



**HAL**  
open science

# Pictorial Composition - Modeling, Perception & Creation

Pierre Lelièvre

► **To cite this version:**

Pierre Lelièvre. Pictorial Composition - Modeling, Perception & Creation. Art and art history. École Normale Supérieure, 2022. English. NNT: . tel-04007348

**HAL Id: tel-04007348**

**<https://hal.science/tel-04007348v1>**

Submitted on 28 Feb 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

**THÈSE DE DOCTORAT**  
**DE L'UNIVERSITÉ PSL**

Préparée à l'École Normale Supérieure

**Pictorial Composition**  
**Modeling, Perception & Creation**

Soutenu par

**Pierre LELIÈVRE**

Le 25 novembre 2022

École doctorale n°540

**Lettres, arts, sciences  
humaines et sociales**

Spécialités

**SACRe, arts visuels  
Sciences cognitives**

Composition du jury :

Norberto GRZYWACZ Pr, Loyola University Chicago	<i>Président &amp; Rapporteur</i>
Sofian AUDRY Pr, Université du Québec à Montréal	<i>Rapporteur</i>
Chien-Chung CHEN Pr, National Taiwan University	<i>Examineur</i>
Alasdair NEWSON MdC, Télécom Paris	<i>Examineur</i>
Katherine STORRS MdC, University of Auckland	<i>Examinatrice</i>
Sylvie TISSOT IdR, ENSAD, Institut Mines-Télécom	<i>Examinatrice</i>
Peter NERI DR, École Normale Supérieure, CNRS	<i>Directeur de thèse</i>



*PhD Thesis*

***Pictorial Composition***  
*Modeling, Perception & Creation*

**Pierre Lelièvre**

February 18, 2023

Supervisor: Peter Neri

Specialities: Research-Creation, Cognitive Sciences, Machine Learning





# Abstract

Pictorial composition, understood as the arrangement of graphical elements on the plane, is typically associated with qualitative rules and heuristics. Although informative for artists and their practice, these norms and guidelines only act as external constraints on the canvas. We believe that artworks are able to fix more fundamental compositional features in their pictorial matter. We therefore develop a paradigm in which every artwork of an artist represents a partial view of a higher-dimensional representation, aggregating intrinsic compositional regularities. We choose to materialize this theoretical hyper-compositional object as a continuous, vectorial and probabilistic space. Our objective is to make regularities explicit for artistic purposes, and to build quantitative metrics for scientific scrutiny. Our research is therefore articulated around a reflexive research-creation agenda: it is grounded in personal artistic material drawing from more than 10 years of practice in abstract composition, it expands along a projective interdisciplinary framework that combines iterative modeling with machine learning, and it engages in perceptual validation using psychophysical techniques.

The sequential non-stationary nature of the compositional process, together with the complex and evolving definitions of its underlying functional units, coalesce into a perceptual phenomenon that cannot be readily modeled through pixel-based deep learning models, such as CNNs. We adopt a different strategy, constructed around a parametric definition of stroke execution and hierarchically nested RNN-VAEs (Recurrent Variational Auto-Encoders), enabling our network to tackle pictorial material by aligning its behavior with the artistic gesture. More specifically, this network architecture extracts compositional regularities by compressing inputs into a reduced number of independent dimensions, ultimately aligned with the representation entertained by artists and observers. These artificial neural networks are trained on >5k personal abstract compositions vectorized as Bézier curves. Although this dataset is large for a single artist, its scale remains relatively small for training large networks. We address this issue by introducing new constraints that support a compact latent space that is both cohesive and expressive.

We then study the resulting compositional space through perceptual judgments of interpolated trajectories spanning targeted locations within this space. In particular, we characterize latent density homogeneity by measuring the perceptual scale adopted by human participants when judging sample similarity. We limit our exploration to circular slices of hyperspheres, along which latent density can be regarded as reasonably stable, and orthogonal linear progressions along the norm,

which imply larger perceptual distortions. We employ a variant of the MLDS method, which we have restricted to local triplets and extended to periodic physical spaces. The empirically measured perceptual scales present regularities that are satisfactorily captured by the notion of Fisher information computed on metrics provided by the model. The resulting algorithms enable artists to explore the dynamical interaction of graphical elements in accordance not only with their own compositional regularities, but also with the perceptual regularities intrinsic to those who view their art. We then come full circle by revealing the hidden compositional dimensions with ink and paper through digitally pen-plotted creations.

**Keywords** Pictorial Art, Composition, Abstraction, Research-Creation, Modeling, Machine Learning, Neural Networks, Deep Learning, RNN, VAE, Perception, Psychophysics, MLDS

## Résumé

La composition picturale, entendue comme la disposition des éléments graphiques sur le plan, est généralement associée à des règles qualitatives et des heuristiques. Bien qu'instructives pour les artistes et leur pratique, ces normes n'agissent que comme des contraintes externes sur le plan. Nous pensons que les œuvres d'art sont capables de fixer des caractéristiques de composition plus fondamentales dans leur matière picturale même. Nous développons donc un paradigme supposant toutes les œuvres d'un-e artiste comme les vues partielles d'une représentation en plus grandes dimensions, agrégeant des régularités compositionnelles intrinsèques. Nous choisissons de matérialiser cet objet hyper-compositionnel théorique par un espace continu, vectoriel et probabiliste. Notre objectif est de rendre ces régularités explicites pour un usage artistique et d'établir des mesures quantitatives pour des études scientifiques. Notre recherche s'inscrit donc pleinement dans un programme réflexif de recherche-création; fondé à la fois sur un matériau artistique personnel, riche d'une pratique de plus de 10 ans de la composition abstraite; et sur une approche interdisciplinaire projective, combinant une modélisation itérative par apprentissage automatique et des vérifications perceptives avec de la psychophysique.

La nature séquentielle et non stationnaire du processus de composition, ainsi que la définition complexe et évolutive de ses unités fonctionnelles sous-jacentes, se combinent en un phénomène perceptif qui ne se modélise pas facilement par les modèles d'apprentissage profond basés sur des pixels, e.g. CNNs. Nous adoptons une stratégie différente, construite autour d'une définition paramétrique de l'exécution des traits, et de RNN-VAEs (Recurrent Variational Auto-Encoders) imbriqués hiérarchiquement, permettant à notre modèle d'aborder la matière picturale en alignant son comportement sur le geste artistique. Plus précisément, cette architecture extrait les régularités compositionnelles en compressant les dessins en un nombre réduit de dimensions indépendantes, alignées dans l'idéal sur la représentation intérieure construite par les artistes et les observateurs. Ces réseaux neuronaux artificiels sont entraînés sur plus de 5000 compositions abstraites personnelles et vectorisées par des courbes de Bézier. Bien que cet ensemble de données soit important pour un seul artiste, son échelle reste relativement réduite pour l'entraînement de réseaux profonds. Nous abordons cette problématique en introduisant de nouvelles contraintes qui encouragent un espace latent à la fois compact, cohésif et expressif.

Nous étudions ensuite l'espace compositionnel résultant à travers des jugements perceptifs de trajectoires interpolées entre des points précis de cet espace. Nous vérifions particulièrement l'homogénéité de la densité latente en mesurant l'échelle perceptive produite par des participants humains jugeant la similarité entre des compositions. Nous limitons notre exploration à des coupes circulaires d'hyper-sphères, dont la densité latente est relativement stable, et des progressions linéaires orthogonales le long de la norme, provoquant des distorsions perceptives plus importantes. Nous utilisons une variante de la méthode MLDS, que nous avons restreinte à des triplets locaux et étendue aux espaces physiques périodiques. Les échelles perceptives mesurées empiriquement présentent des régularités qui sont capturées de manière satisfaisante par la notion d'information de Fisher calculée à partir des métriques fournies par le modèle. Les algorithmes qui en résultent permettent aux artistes d'explorer l'interaction dynamique des éléments graphiques en fonction non seulement de leurs propres régularités de composition, mais aussi des régularités perceptives intrinsèques de ceux qui voient leur art. Nous terminons enfin ce cycle en révélant les dimensions compositionnelles cachées avec de l'encre et du papier, via des créations par traceur numérique.

**Mots-clés** Art Pictural, Composition, Abstraction, Recherche-Création, Modélisation, Apprentissage Machine, Réseaux de Neurones, Apprentissage Profond, RNN, VAE, Perception, Psychophysique, MLDS

## ***Notice to readers***

The work described in this document is funded by an unconventional PhD program called SACRe (Science Arts Création Recherche), which was specifically instituted to support research at the interface between art and science. As a consequence, this thesis is written in an unconventional style that may come across as off-putting to some readers, especially those who are more familiar with scientific reports. The document of the thesis is intended as an artifact in and of itself, in addition to being a document that records certain scientific contributions and results. It is also a story, a reflection, an account of my experience. As such, it is often written in an introspective style that should invite readers to share my experience and see my research from a more intimate perspective, and one that involves multiple dimensions beyond the strictly scientific one. This deliberate effort is in keeping with the remit of the SACRe program. I apologize in advance to readers who may not resonate with the style, and ask for their understanding in accommodating this unusual reading exercise.



# Acknowledgements

A PhD is a long and personal journey, especially when you are intimately bound to your project. Nonetheless, it would not have been possible without the support of, and discussions with, many different people.

First of all, I am deeply grateful to my supervisor Peter Neri. He offered me a unique chance to achieve my objectives with great autonomy. However, this creative freedom only succeeds thanks to his remarkable open-mindedness and implicit trust, making me confident to move forward. Peter has also been a strong scientific support, helping me to clarify my ideas and deepen my reflection. I was delighted by our collaboration on my first paper. Finally, writing in a foreign language can be demanding – Peter generously helped me with the editing process for this manuscript, improving its precision and readability.

I also want to thank the whole LSP team for the many great interactions they offered, and for their material support, such as the opportunity to participate in international conferences and access to computing resources. They have been great colleagues and friends. In particular, I would like to extend my gratitude to Jonathan Vacher, who relentlessly answered my mathematical questions and kindly shared his pertinent insights.

I am also grateful to the singular SACRe program, which provides a rare environment for the fruitful development of research-creation projects. Without their help, projects such as mine would be funded with great difficulty, if at all. I want to thank artists and researchers of this community for their inspiring seminars and discussions.

My gratitude also goes to my *comité de suivi individuel* for their advice. Samuel Bianchini was instrumental in prompting me to explore my artistic vision more deeply, and Cédric Guiard was probably the first person who encouraged me to do a PhD. I would like to thank him again for letting me grow my R&D skills at his company, right after my master's degree.

Finally, I am grateful to the jury for their time and heartening interest in my work.

---



Je voudrais également remercier ma famille et mes amis pour leur accompagnement chaleureux et décisif durant toutes ces années. Je pense particulièrement à Damien et à ma *petite* sœur Chloé, qui ayant déjà traversé ce long parcours de la thèse de doctorat, ont su me donner de précieux conseils. J'adresse aussi une pensée spéciale au soutien indéfectible de mes parents. Je voudrais leur répéter ici les remerciements formulés dans mon mémoire de master ; ceux de m'avoir toujours permis de réaliser les projets qui me tenaient à cœur, comme s'ils avaient été les leurs.

Enfin, et peut-être surtout, je voudrais exprimer ma plus grande reconnaissance à Yunya, ma femme, sans qui je ne serais jamais parvenu au terme de cette aventure. Je n'oublierai jamais son aide exceptionnelle au quotidien, son enthousiasme et son amour dans les moments de doute, ainsi que son regard bienveillant, pertinent et précis dans les phases de création.

Merci.

# Contents

Abstract . . . . .	i
Résumé . . . . .	iii
Notice to readers . . . . .	v
Acknowledgements . . . . .	vii
<b>Introduction</b>	<b>1</b>
<b>I Composition modeling</b>	<b>11</b>
<b>1 Compositional paradigm</b>	<b>13</b>
1.1 Compositional metrics . . . . .	15
1.2 Sources of complexity . . . . .	25
1.3 Axioms . . . . .	39
<b>2 Personal practice</b>	<b>53</b>
2.1 Floating compositional structures . . . . .	55
2.2 Vectorial decomposition . . . . .	61
2.3 Dataset formatting . . . . .	85
<b>3 Model implementation</b>	<b>105</b>
3.1 Probabilistic models . . . . .	105
3.2 Stroke model . . . . .	116
3.3 Composition model . . . . .	120
3.4 Compositional plane model . . . . .	124
3.5 Practical model training . . . . .	127
<b>II Composition exploration</b>	<b>147</b>
<b>4 Model results and tools</b>	<b>149</b>
4.1 Reconstruction and generation . . . . .	149
4.2 Interpolation . . . . .	166
4.3 Measurements . . . . .	175

## **Contents**

<b>5</b>	<b><i>Composition perception</i></b>	<b>185</b>
5.1	Dimensionality issues . . . . .	186
5.2	Perceptual scaling . . . . .	196
5.3	Interpolation — Experimental results . . . . .	213
<b>6</b>	<b><i>Ink and paper</i></b>	<b>233</b>
6.1	Returning to the material space . . . . .	234
6.2	Stroking the line . . . . .	241
6.3	Diversity, continuity, and dynamics . . . . .	253
	<b><i>Conclusion</i></b>	<b>269</b>
	<b><i>Appendices</i></b>	<b>281</b>
A.1	A deep learning framework for human perception of composition . . . . .	281
A.2	ART@VSAC 2019 . . . . .	305
A.3	Defense exhibition . . . . .	315
A.4	Long résumé en français . . . . .	323
	Bibliography . . . . .	340
	Supplementary bibliography . . . . .	348
	List of figures . . . . .	352
	List of algorithms . . . . .	353

# Introduction

## *Animating the painting*

In 2010, at the beginning of my cinematographic studies at the ENS Louis Lumière, one of our first projects was *Animating the painting*. Students typically animate pictorial material from famous masterpieces in the form of a stop-motion sequence of cut-out photographs. Digital tools allow them to stretch and compress the surface of the painting while disregarding the grain, the touch, or the fundamental material nature of the initial work. Their exclusive goal is to tell a story. For me, *Animating the painting* could only make sense in the dynamic revelation of abstract structures. I wanted to make tangible an immaterial displacement over a material surface. If lines could miraculously gain the ability to slide over the surface of the plane, they may not simply do that along their dimensions of width or height, but along a dimension of which we are not aware. If the spectator were able to project himself/herself onto the dimension of plastic depth, perspective would have nothing to do with it. Magic can only operate through a compositional dimension transposed into a temporal dimension.

This short animation, revisiting *Acht Mal* by Wassily Kandinsky, is the story of a morphogenesis. A selection of photograms is reproduced in Fig.0.1. A first phase is virtually operated by the spectator with an input movement towards the canvas. This external dynamic gives the primary momentum to a transversal displacement in the space of the painting, and our trajectory transforms graphical elements. In a second step, a celestial slit appears at the top of the frame, explicitly standing for mother-like generative organs. An eruption of lines completes the metamorphosis of the entire pictorial surface into a panspermia. Once hatching has been completed, we observe a return to the initial state, in negative contrast. In this cycle, the predominance of music acts as a clock. Its rhythmic logic imposes itself as an abstract rule, specific to the metabolism of the forms placed on the canvas. Rhythm confers to the composition its dynamic truth, like an inner logic that resonates outside.

Of course, my view on this project has evolved over time and has now become partially retrospective. However, during the writing of this manuscript, this animation seemed surprisingly significant to me, as if it was the seed of deeper questions, which took me nearly ten years to formulate completely.



## From practice to questioning

The first scribbles of a child are not intended as representation. They are a form of the enjoyable motor activity. [...] It is an exciting experience to bring about something visible that was not there before.<sup>1</sup>

The practice of drawing is originally exploratory. Children draw aimlessly, for the pure pleasure of filling the paper sheet. Representing familiar people, animals, or objects only comes later. For me, there was initially something of the same graphical joy, resulting from an intuitive and unconscious assembly of simple lines. It should also be noted that my drawing activity has often accompanied moments of physical constraints. I would not say *boredom*, but a certain need for concentration without being able to move other than within the free few square centimeters of draft papers, during a class, a meeting, a conference, or a phone call. I surmise that the act of focusing attention on listening made it possible for the visuomotor loop to free itself from a figurative objective. I was then probably freer to explore different abstract building mechanics (Fig.0.2). Subsequently, geometric logic has gradually given way to finer sensitivity for the line and more delicate arrangements (Fig.0.3).

It may have been only in the last five years that these structures have become truly compositional (Fig.0.4). I now feel able to articulate graphical elements of greater diversity and to judge combinations of a higher complexity. The practice has also gradually shifted to dedicated moments of creation. At the same time, the physical size of these propositions has not changed. I continue to draw small figures, approximately contained within a circle of 4cm in diameter. I believe that what fascinates me at this scale – by getting so close to the sheet of paper – is that I can better discern the contact of ink and paper. Around me, the world moves further away, and I have the impression of witnessing the life of a microcosm, to be the eye a little demiurge. It is also a fruitful method for focusing on elementary structural phenomena. The size of the pen or brush, restricted to such a small space, becomes an *ad hoc* limiting factor for the level of detail that can be explored. Indeed, I consider my collection as barely sprouted seeds of hypothetical more complete works. That may be one reason why I started to collect all these fragments, which now count in the excess of 5k.

My artistic practice is therefore initially more introspective than directed towards others. If the practice of composition is an act of knowledge in itself, then, the act of testifying to the evolution of this process, its associated questions and its discoveries is at least as important as the presentation of the corpus itself. Because the interrogations posed by the final drawings are only implicit, I think that viewers are more likely to grasp my point by reading my publications and carrying out their own compositions inspired by the principles exposed in those publications, rather than by merely standing in front of my drawings. Artworks may only provide

---

<sup>1</sup>Arnheim, 1954/2004, p. 171.

# Introduction

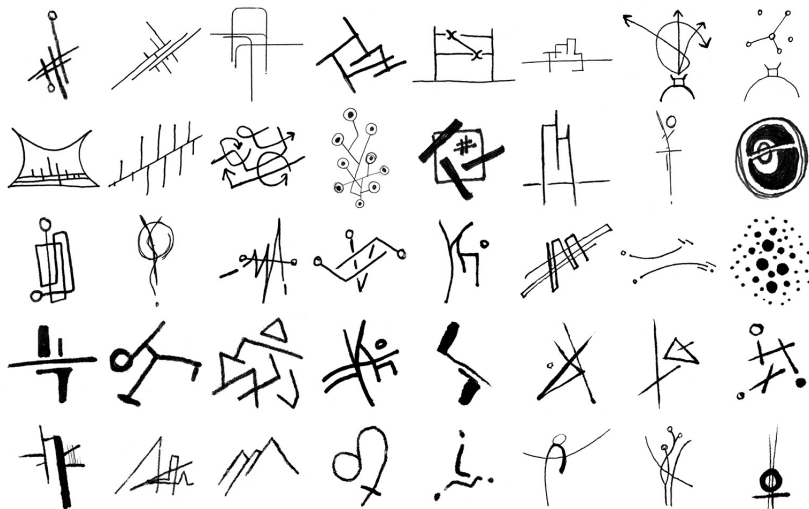


Figure 0.2: Drawings of low complexity.

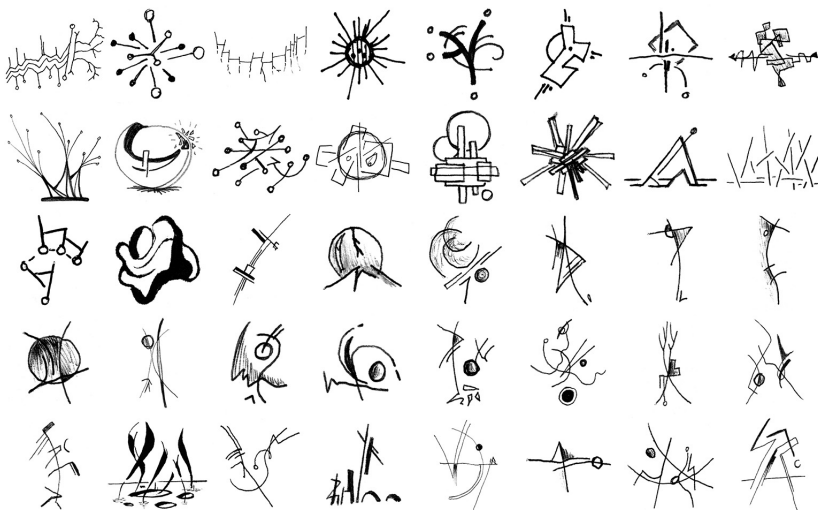


Figure 0.3: Drawings of medium complexity.

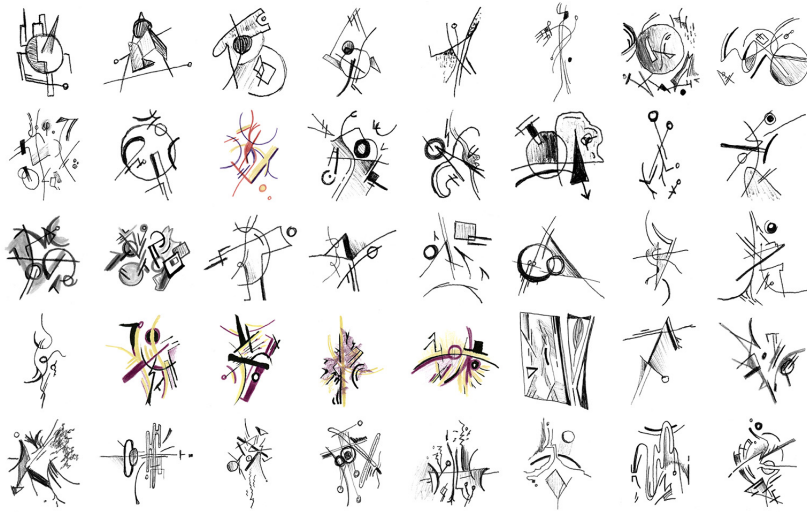


Figure 0.4: Drawings of high complexity.

suggestions. I am always happy to show my work and share it with others, but I believe that, under those circumstances, primal enjoyment supersedes theory. I regard being exhibited as a secondary objective of my artistic trajectory. Above all, I conceive my artistic practice as a motor for personal questioning.

Those that show the thought process of the artist are sometimes more interesting than the final product.<sup>2</sup>

The initial question was rooted in practical necessity: explaining the indescribable need to put one line in one specific place rather than at another location, understanding what factors impose such a graphical element instead of other possibilities from an infinite variety. As the resolution of situations seems both *obvious* and without an objective logic, I began to feel a real tension in my practice and a certain anxiety about doing things randomly. I wanted to understand the laws governing my own practice and make sure I was consistent with myself.

You who speculate on the nature of things, I praise you not for knowing the processes which nature ordinarily effects of herself, but rejoice if so be that you know the issue of such things as your mind conceives.<sup>3</sup>

The real catalyst for a scientific reformulation of these personal questions was my encounter with theoretical writings by artists, in particular Paul Klee's *On Modern Art* and Wassily Kandinsky's *Concerning the Spiritual in Art*. Their work

<sup>2</sup>LeWitt, 1967.

<sup>3</sup>Da Vinci, 1955, p. 70. (G 47 r).



## Introduction

echoed sensations and intuitions which I had perceived during my own practice of composition, and they opened my imagination to a scientific approach to the field of art, beyond a pure historical vision. Their thought will be abundantly represented in the following pages. Despite broader research along my thesis project, this original corpus remains to me most relevant with relation to the compositional paradigm I will develop in this manuscript. Ideally, I would like to be able to join this family of artists whose research has been mainly self-reflective, while at the same time striving to share an experimental method aiming at an objective and inclusive knowledge of art.

For my part, I am now specifically interested in understanding the dynamical dimension of forms. This dimension is implicitly reflected by the constraints between different graphical elements of the plane, and by the regularities that unite all compositions of a given artist. I have the intuition of a pictorial continuity that makes each drawing an approximate artifact of a more complete and coherent whole. Compositional structures belong to a single system, a hyper-compositional object. This view is driven by a fundamental desire to see pictorial forms as living forms, as occurrences of the same complex organism. This highly dimensional object is of course difficult to visualize in our mind, or to represent explicitly as an explorable object. By means of computer simulations I would like to unveil, even slightly, those dimensions that are still hidden from us. Indirectly, this implies being able to measure and quantify compositional regularities, and possibly make them available to artists so that they can make more conscious choices (if they so desire). Concretely, I dream of moving lines, animated by their internal forces, and tools for traveling across the continuous space defined by compositions.

## Method

Traditional analytical methodology is essentially reductionist. The idea is to break down a problem into as many small pieces as necessary. Then, evidence after evidence, scientists go back up the causal chain and synthesize the discoveries. By nature, the composition is an object with a blurred delimitation, the nexus of perceptual and aesthetic considerations. It is also a concept covering both artistic practice and the state of macroscopic spatial organization of sub-elements, where each of them seems essential *a priori*. Therefore, how to segment a composition without alteration? Segmenting a problem in wrong sub-elements can quickly increase its difficulty<sup>4</sup>. We will therefore use a more projective and comprehensive scientific approach: modeling. This method mainly consists in building a tool to simulate complex phenomena, without necessarily partitioning the phenomenon of interest into sub-elements.

---

<sup>4</sup>Le Moigne, 1977/2006, p. 34.

Composition is an essential aspect of pictorial art. However, each epoch and artistic current have developed a set of different methods and appreciation criteria to address their underlying problematics. This makes it inherently difficult to achieve a modeling of the composition that will be universal enough to cover all its manifestations. But is this really necessary? Art, like modeling, are practices requiring to take a position on reality and the studied material. The requirement of universality is then replaced by a commitment to clarity of the adopted approach. This requirement will be concretized by the definition of a compositional paradigm. Modeling is thus perfectly articulated with a research-creation endeavor, which also requires transparency as the only means of legitimizing the general character that is reflected by the particular nature of an individual practice.

In the compositional context, the difficulty for the artist is not so much to free himself/herself of the subjective nature of his/her own vision, but rather to have the necessary hindsight to extract its regularities. This complex problem arises with equal difficulty for the spectator, even if he/she has an external point of view. This issue may be less problematic when artists and spectators live at different times. We therefore require the help of a third party capable of automatic analysis: machine learning and in particular deep learning, which relies on artificial neural networks. On the one hand, deep learning may be viewed as a coldly objective statistical tool, since it is able to extract patterns from data in unsupervised manner. On the other hand, this approach often takes inspiration from the architecture of biological brains, making it particularly relevant to human pictorial composition.

It must be emphasized that *objective* should not be intended in the sense of *universal*. The resulting scope of application is initially and essentially the one pertaining to the dataset itself. In addition, observed regularities may not be related to human perceptual reality. To this end, the cognitive sciences become essential, particularly psychophysics. The latter approach involves quantitative analysis of the connection between a physical stimulus, real or simulated, and its perception. To determine the range of validity for a given model and its potential for generalization, these factors must be assessed by means of an experimental procedure aimed at human participants. While some researchers attempt to explicitly model brain mechanisms, our project developed under looser constraints. Our model is primarily intended as a generative and simulation tool, serving as a basis for perception experiments, and an instrument for statistical measurements on composition.

Our research program will be structured around the following elements: active creation, compositional paradigm, computational implementation and experimental verification. The horizon of application for this work lies at the intersection between the legacy of artists who have made compositional practice the core subject of their work, and the needs arising from a personal practice. This project is then at the crossroads of many fields: research-creation, image processing, machine learning and psychophysics. The interdisciplinary nature of this effort necessarily

## Introduction

involves some difficulties when material from different specialties is incorporated into one document, as each specialty uses its own vocabulary and processes. This, in turn, entails a complex balance between explanation and detail. I hope that the trade-off between these two factors adopted in the rest of this document will be well-matched to the interests of each audience.

Despite a certain proximity of content, this project is not a research effort in aesthetic theory or art history. Even when referring to significant artistic movements or quoting essays written by artists, my goal is not to produce a historical review of compositional evolution. I do not possess the necessary tools and required advanced knowledge for such an endeavor. In addition, I would like to distance the proposed methodology from the field of empirical aesthetics, and more precisely from quantitative approaches using direct aesthetic judgments, such as those relying on *beauty*. Those approaches lead to the discovery of extremely general preferences, e.g. for symmetry rather than asymmetry, for curves rather than angles. These *average* trends sometimes mask high interpersonal variability. I prefer to address aesthetics from artistic artifacts themselves. Human judgment can relate to artistic material, but we must limit ourselves to judgments with a very specific objective, e.g. similarity. *Preference* must be evaluated with a specific purpose. For instance, in our study on the orientation of abstract paintings<sup>5</sup>, participants were asked to determine the optimal orientation of abstract compositions. Thus, it was not an aesthetic preference in general, but a preference constrained to an unambiguous situation.

We conclude this Introduction by outlining the backbone of this manuscript. It is articulated around two main parts. We begin by focusing on the development of the compositional model (Part.I). Chapter.1 discusses the compositional paradigm adopted here, with the purpose of clarifying our position on this concept and delimiting its scope. Chapter.2 describes the processing steps involved in digitizing my personal dataset of compositions, and the chosen representation – structural specifications – that are suitable for algorithmic consumption. Before obtaining a functional tool, the final required step is an effective implementation using artificial neural networks. Chapter.3 details the different architectural choices that were involved in developing the network, and the numerous optimization strategies that were necessary to make model training feasible.

The second part of this thesis is dedicated to exploration (Part.II). Chapter.4 draws an inventory of the raw results and features offered by the model, in particular those available for quantitative measurement. Chapter.5 focuses on verifying continuity of the hyper-compositional object instantiated by the main model. In this chapter, we present the methods and results of psychophysical experiments deployed online

---

<sup>5</sup>Lelièvre and Neri, 2021 describes work conducted during the first year of this PhD. It can be seen as a proof of concept for the method presented above, i.e. deep learning models combined with psychophysical investigations in humans. Nonetheless, it tackles composition through the proxy of painting orientation, and we could not identify a natural place for this research within the main narrative of this thesis. The whole paper is therefore reproduced in the Appendix.A.1.

to quantify the perceptual scale of local similarities along interpolation paths in this novel object. The last chapter (6) is devoted to artistic investigations of the compositional model. Here I explore how to reveal hidden dimensions and how to effectively render their fundamental dynamical character. I also relate the story of a return to the material space, where the unique contact between ink and paper takes place.



*Part I*

*Composition modeling*



# 1 Compositional paradigm

The main goal of this first chapter is to clearly delineate a definition of composition, and to develop this concept into an object that can be modeled. Composition can be understood generically as *the arrangement of forms*. However, this definition does not specify the objectives pursued by artists when they organize graphical elements, nor the means by which they achieve those goals. A statement such as: “follow your instinct to obtain a harmonious whole” is not sufficiently specified to support a compositional model. In addition, modeling is a scientific method that is just as demanding and engaging as artistic practice. These two approaches must position themselves in relation to the world at large in a manner that is inseparable from the person who formulates them. Historical observations, perceptual limitations, intrinsic complexity and axioms are therefore all aspects of the same operational necessity, and are brought together to bear on the world via the formulation of a compositional paradigm.

*Paradigm* refers to a set of fundamental and critical assumptions on the basis of which theories and models are developed. Both theories and models are more completely specified. [...] A number of different models can be generated which have significant differences despite the fact that they all depend upon the same paradigm assumptions.<sup>1</sup>

John Steinbruner defines a paradigm as a set of observations and assumptions from which a whole family of models or theories can follow, without violating the foundational insights from which they originate. This thesis seems better aligned with modeling approaches rather than *theoretical* frameworks, as modeling is an iterative process, “an attempt to fix loose ends, a partial effort that undergoes continuous rearrangements.”<sup>2</sup> A theory is characterized by a more definitive stance, often to the explicit exclusion of alternative positions. In this sense, the research presented here is intended as a proposal under construction, in constant need of empirical verification. In addition, a unique paradigm has the ability to generate several functional models. Only one of them will be detailed in the next chapters, and later deployed through multiple implementations, i.e. operational models that run on a computer. In its transition towards functional implementation, the original paradigm necessarily loses its generality. The modeler is forced to make design choices that are increasingly driven by practical considerations and constrained by experimental limitations.

---

<sup>1</sup>Steinbruner, 1974/2021, p. 11.

<sup>2</sup>Extract from the CNRS strategic project 2002 reproduced in Le Moigne, 1977/2006, p. xii: “un travail de mise en ordre, partiel et continuellement remaniable.”



## 1 Compositional paradigm

In the field of machine learning, a model is almost exclusively understood as an implemented algorithmic process. The underlying *theoretical* framework is often stated only implicitly, because the functional objective is clearly defined (e.g. automated recognition of handwritten postal addresses). Whether the model offers a relevant understanding of the represented phenomenon does not really matter, as long as its practical effectiveness is demonstrated. In our case, pictorial composition is a phenomenon with fuzzy characteristics and often unstated motivations. In this thesis, a complete theoretical specification is therefore required.

Modeling [...] primarily means trying to identify and formulate the problem posed by modelers (a project), by implementing a modeling procedure whose rules are intelligible and accepted. This notion implies a conception of knowledge that is more projective than objective; it solicits the explanation of the postulated axioms - hic et nunc - on the part of the modeler, it calls for a recognition of the creative mind that knows it is formed by the reason it forms.<sup>3</sup>

Thus, modeling means choosing a point of view, and even designing a particular view on a phenomenon, before verifying it. This methodology necessarily engages the responsibility of the modeler in order to acquire scientific legitimacy. By describing his/her motivations and assumptions, as well argued as possible, the modeler develops a kind of ethical contract with the scientific community. Similarly, the artist has freedom only within the limits of the tacit contract of sincerity he/she subscribes to when engaging with spectators. To me, modeling is therefore perfectly aligned with a research-creation project, since it also uses a highly reflective methodological process on its own practice. I personally consider modeling as one of the most humble approaches to art and science.

Modeling initially comes from a reaction to the Cartesian dogma according to which phenomena cannot be fully grasped in their entire complexity. As a result, modeling proposes to simulate, even partially, a phenomenon with an artificial system. Of course, such system is now predominantly computational.

Instead of trying to analyze the mechanisms, we only analyze the functions, which we try to formally describe in the most precise possible manner; it is then a question of realizing - at least on paper, and if possible, concretely - a machine that performs the same functions under identical conditions.<sup>4</sup>

---

<sup>3</sup>Le Moigne, 1977/2006, p. 271: "Modéliser [...] c'est d'abord chercher à formuler - à identifier - le problème que se posent les modélisateurs (un projet), en mettant en œuvre une procédure de modélisation dont les règles sont intelligibles et acceptées. Cette conception de la conception implique, il est vrai, une conception de la connaissance plus projective qu'objective ; elle sollicite davantage l'explicitation des axiomes que postule - hic et nunc - le modélisateur, elle appelle une reconnaissance de l'esprit créateur qui se sait formé par la raison qu'il forme."

<sup>4</sup> Atlan, 1972/2006: "Au lieu d'essayer d'analyser les mécanismes, on analyse seulement les fonctions, qu'on essaie de décrire formellement de la façon la plus précise possible ; il s'agit ensuite de réaliser - au moins sur le papier, et si possible, concrètement - une machine qui accomplisse les mêmes fonctions dans des conditions identiques."

Modeling is therefore not *decomposing*, but *composing*. A model is a representation that accounts for past observations to predict future observations, or to expand its capabilities under more varied, but similar conditions. Despite being inspired by the reality of a phenomenon, this imitation may present discrepancies with the *natural* system. The relationship to a physiological truth is not mandatory<sup>5</sup>. A model of this kind would still be interesting as a simulation tool, and would aid our understanding of an otherwise unimaginably complex reality. It just requires more meticulous behavioral verification.

Importantly, modeling requires a *measurable* analysis of the functions of the *object* under scrutiny. Despite being a core aesthetic dimension of pictorial art, composition remains difficult to evaluate quantitatively. We would like to reiterate our concerns about empirical aesthetics, taking *primary* aesthetic judgments as metrics. We do not believe in straightforward quantitative approaches to beauty. To be relevant, we prefer to develop measuring tools that address aesthetic concepts more indirectly. In this chapter, we will therefore discuss the notion of *measure* in the context of compositional practices from different periods. We will then try to characterize composition in all its complexity, by highlighting compositional aspects preventing any traditional analytical investigation. Finally, we will discuss the specific axiomatic framework that underlies the proposed paradigm. We hope that the sincerity of the approach and the transparency of its objectives will promote composition as a legitimate and operational area for scientific research.

## 1.1 Compositional metrics

Artwork is a measure of space, it is form, and this is what must be considered first.<sup>6</sup>

For art historian Henri Focillon, art is naturally familiar with metrics, since art would itself be a measuring instrument. By virtue of its physicality, a work of art crystallizes its own rules into matter. The invoked type of measure thus appears as a potential means of achieving knowledge about the world. However, in the absence of a more precise definition, this measure remains theoretical, since it has no objective scale, nor a universal frame of reference. If measuring is the basis of many sciences, it is not only because it enables quantitative evaluation of a given aspect of reality: above all, it is because it provides instruments for comparing different aspects of reality, and for verifying their compliance with a specified standard or project. Under these conditions, a metric can become an instrument

---

<sup>5</sup>Nonetheless, let us mention a few examples where analogies have been demonstrated. Computer vision based on convolutional artificial neural networks presents similarities with human vision and its neural architecture (Rajalingham et al., 2018; Yamins & DiCarlo, 2016). Our own work on the orientation of abstract paintings (Lelièvre & Neri, 2021) is also relevant in this context.

<sup>6</sup>Focillon, 1934, p. 6: "L'œuvre d'art est mesure de l'espace, elle est forme, et c'est ce qu'il faut d'abord considérer."

## 1 Compositional paradigm

of creation. But actual implementation requires clear answers to the following questions: what to measure, and how to measure it? This section will describe the use of metrics in different compositional practices, and highlight associated questions. We emphasize that this effort should not be intended as an exhaustive historical analysis. Instead, we focus on a selection of topics that are directly relevant to the paradigm we will introduce in later sections.

### *Building the space*

The first use of measurement in pictorial art is probably visible in the transposition of various rhythms onto the spatial domain<sup>7</sup>. These extremely varied ornaments and patterns span history and culture, but they are largely tangential to compositional issues. We have decided to omit them from the present discussion.

Composition in the Western tradition is intimately connected with numbers, which play both practical and symbolic roles. A number system serves primarily a *practical* role. Like the famous golden number, certain measures, ratios, and other elected values (see Fig.1.1a,b) impart dimensions to the canvas and its constituent parts, relatively to the whole and between them. Numbers are also *symbolic*. The golden number, sometimes called the *divine section*, carries the idea of natural perfection, which cannot be questioned as it is directly handed down by the Creator. Before the Renaissance, proportions also dictated the importance of different human figures in religious art. Large imposing spaces are associated with moral or spiritual greatness. These measures and heuristics contribute to the communicative effectiveness of the religious message, while relieving artists from the need to make decisions or personal judgments.

During the Renaissance, mathematics and geometry (the latter conceived as a variation of the former specifically dedicated to spatial construction) became integral to the pictorial arts. The use of rulers, compasses and learned procedures made it possible to determine the main points of interest on a canvas. As a result, composition is reduced to a rigid mechanic of construction, with perspective as its culmination (see Fig.1.1c). Alongside an obsession for realism, painting enters into a practice that was ultimately more architectural than truly plastic. Perspective also reinforces the idea that the space of the canvas is only an extension of real space: equivalent metrics are applied with comparable relevance. The painting becomes both a window onto the world, and the illusion of a 3-dimensional box. Thus, composition can be understood as the construction of space, the arrangement of forms as an architecture of images, in its literal sense.

At that time, the idea that forms may be dominated by reason appears to take hold, as if all forms should derive from the same mathematical logic. For example, the emblematic *Vitruvian Man* designed by Da Vinci goes beyond a practical

---

<sup>7</sup>Focillon, 1934, p. 22.

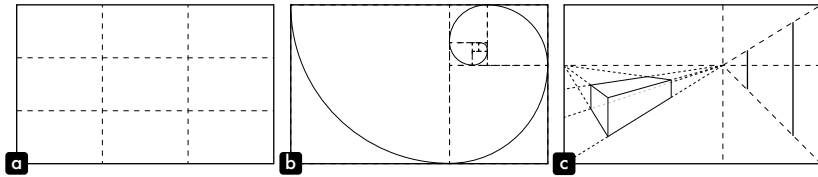


Figure 1.1: Standard ratios and construction lines. In panel **a**, the rule of thirds. In panel **b**, golden ratio and associated spiral. In panel **c**, 2-point perspective.

prescription for the representation of the human figure. Rather, this diagram testifies to the conviction that forms should *conform* to theoretical measures imposed by external sources. While human is placed at the center of the world, this *top-down* standardization of the human body seems to break with traditional spatial measurements, based on customary metrics that were flexible and rich of the diversity of *working hands*. Craftsmen, like artists, could find all relevant measures on themselves, e.g. by looking at their thumbs, feet, elbows. These types of metrics were therefore intrinsically legitimizing forms in a *bottom-up* non-normative fashion: embodied, they were carrying inherent coherence.

This rigid and universalist trend intensified up to the *Age of Enlightenment*. Impressionists finally abandoned this vision to exclusively rely on their raw perception of natural phenomena. Art then seemed to allow some degree of subjugation to reality, rather than serving as an independent instrument of form exploration. Color theory greatly benefited from this artistic current, not so much compositional theory.

Later, the nabi painter Paul Sérusier wrote the following about the *right proportions*:

1 is not a number: it contains and generates all numbers. 2 expresses the struggle between two principles. The fight is sterile, if it does not produce a result, which, together with both principles, constitutes the number 3. From there the idea of a Trinity in several religions. [...] The number 3 means: God or the Creator. The number 4 is no longer a prime number; it is the square of 2. The square means balance in matter.<sup>8</sup>

<sup>8</sup>Sérusier, 1921, pp. 15–16: “1 n’est pas un nombre: il contient et engendre tous les nombres. 2 exprime la lutte entre deux principes. La lutte est stérile, si elle ne produit un résultat, qui, joint aux deux principes, constitue le nombre 3. De là l’idée d’une Trinité dans plusieurs religions. [...] Le nombre 3 signifie : Dieu ou le Créateur. Le nombre 4 n’est plus un nombre premier; il est le carré de 2. Le carré signifie l’équilibre dans la matière.”

On page 29, there is another example of confused mathematical language concerning color: “No matter how clean the powdered colors are, there always remains a little impurity that I will call  $\epsilon$ . So if I mix 2 complementary colors for example: red +  $\epsilon$  and green +  $\epsilon$ , red and green neutralizing each other, there remains  $2\epsilon$ , i.e. impurity.” (Si propres qui soient les couleurs en poudre, en effet, il y reste toujours un peu d’impureté que j’appellerai  $\epsilon$ . Si donc je mélange 2 complémentaires par exemple : rouge +  $\epsilon$  à vert +  $\epsilon$ , le rouge et le vert se neutralisant, il reste  $2\epsilon$ , c’est-à-dire de l’impureté.)

## 1 Compositional paradigm

Despite the apparent mathematical vocabulary and a certain logic of demonstration, this approach to number is only vaguely scientific. It is all the more surprising since these concepts do not seem to have any direct connection with Sérusier's pictorial practice, as if he himself was not convinced by his own argumentation, as if he felt he had to bridge some theoretical gap. It may also be symptomatic of confusion between numbers and metrics. Numbers are values without dimensions or hidden meanings. They have no *value* in themselves. A metric, on the other hand, uses numbers to establish a unit of measurement, using a comparison standard that gives meaning to the values. Modernity approaching, it may have brought forward the ambition to apply the universal language of mathematics to art, in the hope of finding objective laws, similarly universal.

In parallel, the East did not seem to suffer from this rational deference to mathematics. For instance, the Chinese painter Chang Yen-yuan thinks that "a line drawn with a ruler is a dead line."<sup>9</sup> Of course, Chinese painting has not been exempt from certain heuristics, but the validity of these laws seems to have been instantly questioned. The painter Shih Tao wrote in the Ming dynasty:

For a rolled vertical landscape painting, tradition proposes the division into three planes, the bottom one for the ground, the middle one for trees and the top one for the mountain. In front of such an obviously divided painting, how will the spectator enjoy a real perspective? If we mechanically follow this method of the three planes, we only obtain a result close to an engraved plate.<sup>10</sup>

### **Building forms**

Composition finally adopted a new paradigm with the expressionists. Paul Klee wants to "elevate construction to the rank of a means of expression." For him, "some earlier eras had already distinguished themselves [...] by the predominance of construction, but as a scaffolding: a method and not an objective."<sup>11</sup> This concept is perhaps even more radical than the abandonment of figurative representation, which is a mere consequence of matching form against the newly introduced paradigm. Thus, various pioneering currents of abstraction gradually introduced the idea of a mathematical structure, not only organizing space, but creating forms themselves. In *On Modern Art*, Klee clearly summarizes this notion:

---

<sup>9</sup>Cheng, 2006, p. 79: "un trait tracé à la règle est un trait mort."

<sup>10</sup>Cheng, 1989, p. 131: "Pour un tableau de paysage en rouleau vertical, la tradition propose la division en trois plans, celui du bas pour le sol, celui du milieu pour les arbres et celui du haut pour la montagne. Devant un tableau divisé de façon aussi évidente, comment le spectateur pourra-t-il jouir d'une vraie perspective ? Si l'on suit mécaniquement cette méthode des trois plans, on n'obtient guère qu'un résultat proche de celui d'une planche gravée."

<sup>11</sup>Klee, 1924/1998, p. 10: "élever la construction au rang d'un moyen d'expression." and "certaines époques antérieures s'étaient déjà distinguées [...] par la prédominance de la construction, mais comme un échafaudage : moyen et non pas fin."

However, it is not absolutely new to think of the form in precise measures as capable of numerical expression. [...] The only difference is that now the ultimate consequences are drawn from the Number up to the pictorial elements, while former masters were satisfied to metrically determine the main lines of a compositional scheme.<sup>12</sup>

Nonetheless, the void left by the abandonment of the figurative approach may have conferred disproportionate hope in the mechanical logic of geometry. In support of a revolutionary ideology and an intransigent industrial imagination, the *naked* number perfectly suited Russian constructivist movement for its apparent objectivity. Evoking constructivism, Philippe Sers observes that “constructing is to set up in space a coherent whole that is understandable and, of course, reproducible or repeatable.”<sup>13</sup> Therefore, numbers, by virtue of the experimental validation which they enable, represent ideal instruments for a modernity that goes beyond popular beliefs, as if they carried, by nature, structuring legitimacy.

This faith in numbers still appears to me as yet another avoidance of the search for purely pictorial measures. Indeed, it was questioned by Wassily Kandinsky in *Der Blaue Reiter*:

In the search for abstract ratios that manifests itself today, numbers play a crucial role. Any digital formula is cold as a summit covered with ice and, by its absolute regularity, firm as a block of marble. [...] Everything can be translated into a mathematical formula, or simply into a number. But there are many numbers: 1 and 0.3333... are equally legitimate beings, endowed with equal inner resonance. Why would we be satisfied with 1? Why would we exclude 0.3333...?<sup>14</sup>

Geometric shapes and the rudimentary use of numbers are defined with respect to an inconsistent symbolic structure. Because of their contingency, numbers by themselves cannot represent an end. Such a basic vision of the compositional metric would still not be freed of practical constraints. To satisfy the desire for a deeper understanding of the perceptual phenomena involved in relating to pictorial materials, it seems necessary to rethink measure as a *non-spatial* measure.

---

<sup>12</sup>Klee, 1924/1998, p. 11: “Il n'est pourtant pas absolument nouveau de penser la forme en mesures précises susceptibles d'une expression numérique. [...] La seule différence est que maintenant on tire du Nombre les conséquences ultimes jusqu'aux éléments de forme, tandis que les anciens maîtres se contentaient de déterminer métriquement les grandes lignes d'un schéma de composition.”

<sup>13</sup>Kandinsky, 1926/1991, pp. xxiv–xxv: “construire, c'est mettre en place dans l'espace un ensemble cohérent compréhensible et, bien sûr, reproductible ou répétable.”

<sup>14</sup>Kandinsky, 1974/2014, p. 163: “Dans la recherche des rapports abstraits qui se manifeste de nos jours, le nombre joue un rôle capital. Toute formule numérique est froide comme un sommet couvert de glaces et, par sa régularité absolue, ferme comme un bloc de marbre. [...] Tout peut être traduit par une formule mathématique, ou simplement par un nombre. Mais il existe bien des nombres: 1 et 0,3333... sont des êtres pareillement légitimes, doués d'une égale résonance intérieure. Pourquoi se contenterait-on de 1? Pourquoi exclurait-on 0,3333... ?”

## 1 Compositional paradigm

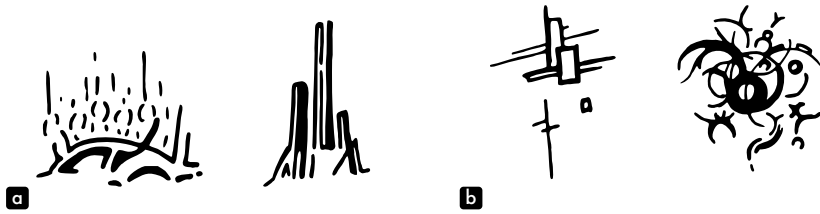


Figure 1.2: Terrestrial (panel a) and celestial (panel b) compositions.

### *Cinematic of forms*

After the neutralization of the figurative logic, the use of abstract forms imposes a new way of evaluating and legitimizing space. Consequently, what is left to the artists that may help them with their task? For Kandinsky, the horizontal line forms the primordial contrast with the vertical line<sup>15</sup>. These lines frame the empty canvas and ideally form mathematical axes. Kandinsky further emphasizes their physical referential aspect with qualifiers such as cold and hot, stable and unstable, whose logic could be associated with thermodynamic or terrestrial gravity. In this context, each graphical element can become a corpuscular entity with some assigned mass.

Thus, a *composition* is nothing other than an *exact law-abiding organization* of the vital forces which, in the form of tensions, are contained within the elements.<sup>16</sup>

Tensions must be understood here as potential movements. Passively immersed in a force field or actively radiating, graphical elements interact with each other and with the canvas. For Kandinsky, the upper part of the canvas evokes flexibility and freedom: *light* elements become lighter, while *heavy* elements become heavier. Conversely, the lower part of the canvas inspires density, gravity and constraint<sup>17</sup>. However, forms are set in motion also by factors beyond terrestrial logic. For Klee, forces housed inside elements can also behave in accordance with “the purest of all mobile forms, the cosmic one, [which] is only created through the suppression of gravity” or within intermediate domains “particularly represented by water and the atmosphere”<sup>18</sup> (see Fig.1.2). These different systems initiate the idea of an interrelationship between forms, but still immutable. Each element is well circumscribed, with a known mass, constituting what could be called a *cinematic of forms*.

<sup>15</sup>Kandinsky, 1926/1991, p. 69.

<sup>16</sup>Kandinsky, 1926/1991, p. 111: “La *composition* n’est donc qu’une *organisation précise et logique des forces vives* contenues dans les éléments sous forme de tensions.”

<sup>17</sup>Kandinsky, 1926/1991, pp. 146–147.

<sup>18</sup>Klee, 1924/1998, pp. 126, 114: “la plus pure des formes en mouvement, la forme cosmique, n’apparaît qu’avec la suppression de la pesanteur” and “représenté notamment par l’eau et l’atmosphère”

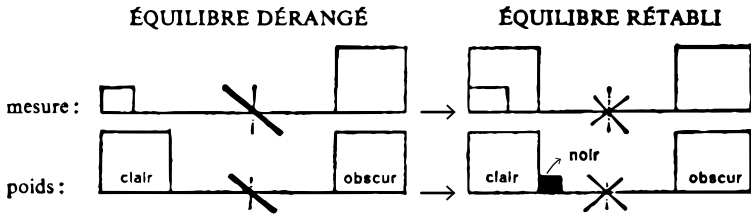


Figure 1.3: Based on drawings by Klee<sup>21</sup>. From *disturbed balance* to *restored balance*. On the first line, *measure*; on the second line, *weight*. *Clair* means light, *obscur* dark and *noir* black.

This new brick is a little too heavy and to my mind puts too much weight on the left; I must add a good-sized counterweight on the right to restore equilibrium.<sup>19</sup>

Although artists commonly use a vocabulary related to the weight of graphical elements, the way in which they measure this quantity often remains logically flawed. What perceptual mechanisms or pictorial aspects do creators rely upon to produce a coherent system, to achieve harmonious dynamics? A popular idea is that surface luminance and color density confer pictorial weight. Variations between light and dark areas seem to determine the perceptual weight of each image fragment<sup>20</sup>. Fig.1.3, taken from Klee's *Pedagogical Sketchbook*, illustrates the theoretical balance adopted by artists. The objective is to keep both sides in alignment by successive adjustments. When qualifying his/her composition, the artist may completely abandon terms like *harmony* in favor of *balance*. In the figure, we also notice that individual lines possess a mass that Klee prefers to simply call *measure*. Such line weights correspond to the space they circumscribe. Thus, we realize that under its apparent and intuitive simplicity, this conception of a compositional metric is actually based on various perceptual mechanisms acting as a whole.

### Inner metrics

For Kandinsky, it is obvious that art “is a domain in itself, governed by its own laws”<sup>22</sup>. In his treatise on composition, *Point and Line to Plan*, he specifies that even if these notions seem “derived from the material world”, pictorially, “they

<sup>19</sup>Klee, 1924/1998, p. 23: “Cette nouvelle pierre, se dit [l'artiste], semble bien un peu lourde et tire mon affaire trop à gauche ; il me faudra un sérieux contre-poids à droite pour rétablir l'équilibre.”

<sup>20</sup>In this chapter we focus on artist's views, but the study of visual weight has a long history in psychophysics too. The psychophysical literature will be further discussed in Subsection.2.3.Spatial standardization.

<sup>21</sup>Klee, 1924/1998, p. 110

<sup>22</sup>Kandinsky, 1974/2014, p. 123: “est un domaine en soi, régi par des lois propres”



## 1 Compositional paradigm

must be understood as tensions living within the elements"<sup>23</sup>. To achieve *true* composition, it seems necessary to overcome ordinary submission to spatial logic.

The means of measurement available to us are, however, exceedingly primitive. It is at present almost impossible for us to imagine how, for example, the *weight* of a scarcely visible point could be expressed by an exact number. The reason for this is that the concept "weight" does not represent a material weight.<sup>24</sup>

In other words, "proportions and balances are not outside the artist, but inside him"<sup>25</sup>. For Focillon:

Is it not the case that these forms, which live in space and matter, first live in the mind? Or rather, is it not really and even exclusively in the mind that they live, their external activity being only the trace of an internal process?<sup>26</sup>

With the same spirit, Apollinaire writes in an article about Matisse:

Ordering chaos here is the creation. And if the artist's goal is to create, we need a type of order that adopts instinct as its measure.<sup>27</sup>

Despite the apparent arbitrariness of possible arrangements offered by abstract graphical elements, they seem to follow intrinsic forces that demand extreme rigor and attention on the part of artists. The principle of an *inner necessity* in composition is therefore more flexible than the cinematic of forms described earlier. At the same time, it is more demanding for the artist in that it requires deep understanding of his/her pictorial material. Even if Kandinsky adds a spiritual dimension to the intuition that supports his judgment, in essence this metric could be consistent with quantifiable perceptual phenomena.

Inner experience is however only partially communicable through language. In the preface of *Concerning the Spiritual in Art*, Philippe Sers reminds us that Kandinsky tries hard to describe his method so that every artist can experience the inner metric for themselves, and then further advance our common knowledge about composition<sup>28</sup>. Knowledge then appears to be potentially acquired at each iteration, by an active practice, by the repetition of measurements and gestures. This conception of art is ultimately connected with oriental painting principles. In

---

<sup>23</sup>Kandinsky, 1926/1991, pp. 146–147: "empruntées au monde matériel" and "elles s'entendent comme tensions intérieures"

<sup>24</sup>Kandinsky, 1926/1991, pp. 148–149: "Les moyens numériques dont nous disposons actuellement sont très primitifs. On peut difficilement imaginer aujourd'hui comment pourrait s'exprimer en chiffres précis le *poids* d'un point à peine visible, d'autant plus que la notion « poids » ne correspond pas à un poids matériel."

<sup>25</sup>Kandinsky, 1912/1989, p. 140.

<sup>26</sup>Focillon, 1934, p. 47: "Ces formes qui vivent dans l'espace et dans la matière ne vivent-elles pas d'abord dans l'esprit ? Ou plutôt n'est-ce pas vraiment et même uniquement dans l'esprit qu'elles vivent, leur activité extérieure n'étant que la trace d'un processus interne ?"

<sup>27</sup>Matisse, 2014, p. 56: "Ordonner un chaos voilà la création. Et si le but de l'artiste est de créer, il faut un ordre dont l'instinct sera la mesure."

<sup>28</sup>Kandinsky, 1912/1989, pp. 16–17.

the *Tao Te King (Dao De Jing)*, Lao Tzu, a pillar of Chinese philosophy, writes that the right action is the non-action, the one that by repeated gesture becomes innate and whose accuracy no longer mobilizes the mind.

A good artist leaves his intuition  
 Taking him where it wants.  
 A good scientist has freed himself from the concepts  
 And keeps an open mind to what is.<sup>29</sup>

At that time, the painter was also a scholar and he/she must represent the point of contact between *good artist* and *good scientist*. François Cheng explains that it “is for [the painter] less about describing the external aspects of the world than about grasping the internal principles [(the *li*)] that structure all things and connect them to each other.”<sup>30</sup> The artist thus seeks to capture the *right* measure of the world, and then to inscribe it in the *right* gesture. The painter Li Jih-hua writes in the Ming dynasty:

More than the *hsing* [external shape], it is important to grasp the *shih* [lines of force]; more than the *shih*, it is important to grasp the *yun* [rhythm or resonance]; more than the *yun*, it is important to grasp the *hsing* [nature or essence].<sup>31</sup>

With Suprematism, part of the Russian avant-garde categorically refuses the influence of the natural order. For its initiator Malevich, art must not serve religious or political purposes either. Thus, Malevich distances himself from the strict constructivist logic. For him, this absolute conception can only be achieved in one way:

The artistic (pictorial) conception, based upon feeling, of linear, two-dimensional and spatial phenomena is not supported by an intellectual understanding and the utilitarian relationships of these phenomena; it is non-objective and subconscious and, viewed from an intellectual standpoint, constitutes, as it were, a *blind, uncontrollable norm*.<sup>32</sup>

Malevich relies on a subconscious, coherent and regular compositional measure, but one which will be forever hidden from him. In contrast to the surrealist movement, painting for him does not represent the subconscious of the artist, but the unconscious use of a pictorial norm, to which abstract artwork would ultimately be the only possible witness. It therefore seems that a relevant compositional measure is not identifiable *a priori*. Whatever the presumed origin of the judgments made by the artist – spiritual, natural, or subconscious – these judgments are *inner*, and only the work of art would be able to fix these metrics in their entirety.

<sup>29</sup>Lao Tseu, 2008, chap. 27: “Un bon artiste laisse son intuition ; le mener là où elle le souhaite. ; Un bon scientifique s’est libéré des concepts ; et garde l’esprit ouvert à ce qui est.”

<sup>30</sup>Cheng, 1989, p. 155: “s’agit pour [le peintre] moins de décrire les aspects extérieurs du monde que de saisir les principes internes [(le *li*)] qui structurent toutes choses et qui les relient les unes aux autres.”

<sup>31</sup>Cheng, 1989, p. 40: “Plus que le *hsing* [forme extérieure], il importe de saisir le *shih* [lignes de force] ; plus que le *shih*, il importe de saisir le *yun* [rythme ou résonance] ; plus que le *yun*, il importe de saisir le *hsing* [nature ou essence].”

<sup>32</sup>Malevich, 1927/2003, p. 20.

## 1 Compositional paradigm

### Decomposing

The science of art advocated by Malevich “must explain the character of the new additional element which has forced its way into the creative organism of the artist and brought about an alteration in our conception of art.”<sup>33</sup> Thus, in order to understand the plastic uniqueness of artworks, Malevich favors an analysis of their environmental and historical context. It is indeed a classical historical approach.

For his part, Kandinsky supports a more autonomous reading of artworks. Research on art must be independent of any moral and aesthetic judgment. “The methods of art analysis have been, until now, far too haphazard and, frequently, too personal in nature.” The observer should ideally occupy a position that is both active (inner analysis) and objective, making “collective work in the science of art possible.”<sup>34</sup> In the background, one can sense the hope of revealing common and universal principles from the pictorial material itself. In this regard, Focillon writes:

The life of forms establishes close relations between masters who have never had the slightest connection between them and who are separated by nature, distance, centuries.<sup>35</sup>

Kandinsky’s approach is therefore fundamentally inclusive. It does not exclude the possibility of finding new metrics consistent with heuristics of the past. Kandinsky summarizes his ambitions as follows:

The research efforts that must become the cornerstone of the new science — the science of art — have two goals and proceed out of two necessities: 1. the need for science in general which grows spontaneously out of a non- or extra-utilitarian urge to know: the *pure* science, and 2. the need for balance in the creative powers that can be grouped under two schematic heads — intuition and calculation: the *practical* science.<sup>36</sup>

Despite the support of important figures such as Kandinsky, the scientific exploration of art always raises fears. For instance, the *decomposition* of artwork would somehow contribute to its desacralization, to the deconstruction of its genius. There is also a certain dread of standardization. However, “dictionaries do not petrify living languages, which are constantly undergoing changes.”<sup>37</sup> In addition:

---

<sup>33</sup>Malevich, 1927/2003, p. 12.

<sup>34</sup>Kandinsky, 1926/1991, p. 91: “Les méthodes de l’analyse de l’art ont toujours été bien trop arbitraires et souvent trop subjectives.” and “possible un travail collectif dans le domaine de l’esthétique expérimentale.”

<sup>35</sup>Focillon, 1934, p. 55: “Entre des maîtres qui n’ont jamais eu entre eux la moindre liaison et que tout sépare, la nature, la distance, les siècles, la vie des formes établit d’étroits rapports.”

<sup>36</sup>Kandinsky, 1926/1991, pp. 19–20: “Les recherches, qui doivent être la base de cette nouvelle science — la science de l’art — ont deux buts et découlent de deux impératifs : 1. du simple désir de savoir, spontanément issu d’un besoin de connaître, sans aucun but pratique, la science « pure » et 2. de la nécessité d’un équilibre des forces créatrices, classées schématiquement en deux composantes — intuition et calcul : la science « appliquée ».”

<sup>37</sup>Kandinsky, 1926/1991, p. 101: “un dictionnaire ne pétrifie pas une langue vivante, qui subit continuellement des changements.”

The general viewpoint of our day, that it would be dangerous to *dissect* art since such dissection would inevitably lead to the abolition of art, originated from an ignorant depreciation of the elements laid bare and of their primary strength.<sup>38</sup>

The fear of unraveling the mystery behind certain aspects of art therefore originates from lack of faith in art, and in the richness of its pictorial material. The search for a metric in art thus cannot be criticized on the basis that it would make the magic of creation disappear. Rather, the critical and legitimate question is whether a scientific approach to art<sup>39</sup> can possibly allow us to understand pictorial phenomena in their full complexity.

### 1.2 Sources of complexity

Art or nature, it is difficult to grasp phenomena of this kind in their entirety as wholes, and even more difficult to make this understanding accessible to others. This is a consequence of the sequential nature that characterizes our methods for studying spatial constructions to obtain a clear and distinct mental representation. It is a consequence of the temporal deficiency of language.<sup>40</sup>

Complexity [is] only the name given to complicated things when the size of the system increases, exceeding the cognitive capacity of the modeler.<sup>41</sup>

For Paul Klee, the temporal dimension of the pictorial work arises as a direct consequence of the limitation of our spatial perception, our inability to grasp spatial concepts all at once simultaneously. For Jean-Louis Le Moigne, this difficulty, introduced as the *complexity of a system*, must be understood more precisely as an exhaustion of our cognitive abilities in comprehending a phenomenon. In both cases, this means that we struggle to extract the underlying meaning of the phenomenon, to consciously analyze its details, to grasp its mechanisms or to memorize its different aspects before representing a coherent whole to our imagination. On the other hand, we have just seen how measurement in composition is essential in practice, and at the same time, how metric evaluation remains beyond formalization: so much so, that the concept of *inner necessity* introduced by Kandinsky appears well formalized despite its rather unspecified nature. This section is concerned with identifying the specific factors that prevent

<sup>38</sup>Kandinsky, 1926/1991, p. 16: "L'opinion, répandue aujourd'hui encore, qu'il serait fatal de « disséquer » l'art, et que cette autopsie mènerait inévitablement à la mort de l'art, résulte de l'ignorante dépréciation des éléments mis à nu et de leurs forces primaires."

<sup>39</sup>Which at that time was mainly Cartesian and positivist.

<sup>40</sup>Klee, 1924/1998, pp. 17–18: "Art ou nature, il est difficile d'embrasser du regard un ensemble de ce genre et encore plus d'en faciliter la vue à autrui. Cela tient aux méthodes échelonnées dans le temps dont nous disposons pour étudier un ensemble spatial afin d'en obtenir une représentation mentale claire et distincte. Cela tient à l'infirmité temporelle du langage."

<sup>41</sup>Le Moigne, 1977/2006, p. 232: "La complexité [n'est] que le nom donné au compliqué lorsque la taille du système augmente, dépassant la capacité cognitive du modélisateur."

## 1 Compositional paradigm

an accurate and objective definition of a compositional metric: what are the sources of complexity inherent to a composition that hinder its metrization?

### *Temporal complexity*

The notion of attributing a temporal dimension to pictorial composition may appear counterintuitive. This idea was not accepted for a long time, and Kandinsky in particular deplored its absence from most artistic theories<sup>42</sup>. Focillon explains that “the work of art is only apparently immobile. [...] In reality, it is born out of a change, and it sets the stage for another change.”<sup>43</sup> He thus suggests that paintings incorporate two distinct temporal moments.

The second of them, which is perhaps the least obvious, is the one already mentioned by Klee in his quote at the beginning of this section. The act of perceiving a composition takes place over time. A cognitive limitation prevents us from accessing the entire work instantly. *A priori*, this may be regarded as a limitation, however the artist can actively exploit this constraint in the pictorial experience. This is what Kandinsky observes in *Rückblicke* with regard to his first contact with time in Rembrandt's painting:

I had the impression that his paintings *were lasting a long time*, and I explained this experience as reflecting the fact that it took me some time to exhaust exploration of a given part of the painting, before I could move on to a different part. Upon further reflection, I realized that this separation [of the chiaroscuro] magically fixes on the canvas an element initially foreign to painting and which seems difficult to grasp: Time.<sup>44</sup>

In the view expressed above, sequential perception acquires a narrative character. In addition, Kandinsky's shift towards abstraction led him to concepts related to the perception of music. For him, “a musical piece lives in our memory in the manner of a painting, i.e. simultaneously via a combination of all its essential parts.”<sup>45</sup> Time in paintings is therefore not a passive experience, but the source of real freedom in movement and exploration. Its illusion of totality gives rise to multiple interpretations for the composition. Being this the case, why does Klee speak of *deficiency* with reference to the necessity for temporal integration in visual perception?

---

<sup>42</sup>Kandinsky, 1926/1991, p. 39.

<sup>43</sup>Focillon, 1934, p. 10: “l'œuvre d'art n'est qu'apparemment immobile. [...] En réalité elle naît d'un changement et elle en prépare un autre.”

<sup>44</sup>Kandinsky, 1974/2014, pp. 102–103: “J'avais l'impression que ses tableaux « duraiient longtemps » et je me l'expliquais par le fait qu'il fallait que je commence par prendre le temps d'en épuiser une partie avant de passer à l'autre. Plus tard je compris que cette séparation [du clair-obscur] fixe comme par enchantement sur la toile un élément initialement étranger à la peinture et qui paraît difficilement saisissable : le Temps.”

<sup>45</sup>Kandinsky, 1912/1989, p. 99: “une œuvre musicale vit dans notre souvenir et comme un tableau, c'est-à-dire simultanément par toutes ses parties essentielles.”

The main handicap of those who contemplate or reproduce [an artwork] is that they find themselves in front of the final result all at once, and that they can only go backwards through the genesis of the work.<sup>46</sup>

The *genesis of the artwork* directly refers to the *iterative* creative process of composition that the artist pursues “until the balance stabilizes.” “A work of art is born of movement.” Creation is the moment when the artist listens to matter set in motion to capture the first breath of forms. “The *form* is the end, the death. *Formation* is Life.”<sup>47</sup> The initial instants of the composition are thus of a completely different nature compared with perceptual time. It is a time that generates diversity. However, spectators are only in contact with this process in very rare cases. Generally, they can only contemplate the resulting artwork where the primordial temporal dimension is flattened.

*Morphogenesis* is a particular consequence of *implemented* pictorial instruments. Tools such as pencils and brushes imply a creative work in *positive*. The material contact reveals the form. In opposition, wood engraving is executed in *negative*, as the artist empties spaces between graphical elements. Therefore, each medium gives an extremely specific meaning to pictorial dynamics and gesture.

The geometric point is, according to our conception, the ultimate and only *union of silence and speech*. [...] A point is the result of the initial collision of the tool with the material surface, with the basic plane. [...] By this first collision, the original plane is impregnated [(fertilized)].<sup>48</sup>

The sudden emergence of a point brings about a change that carries extreme tension. Nothing has happened yet, but everything is potential. There is no possible turning back.

Every point has its own existence; it promises multiple transformations. The act of setting down a point, is equivalent to the act of sowing a seed; it must grow and become...<sup>49</sup>

For the Chinese painter Huang Pin-Hung, a point on a canvas calls for a necessary transformation, which Kandinsky describes in the following way.

There exists still another force which develops not within the point, but outside of it. This force hurls itself upon the point which is digging its way into the surface, tears it out and pushes it about the surface in one direction or

---

<sup>46</sup>Klee, 1924/1998, p. 38: “Le principal handicap de celui qui la contemple ou la reproduit est qu’il est mis d’emblée devant un aboutissement et qu’il ne peut parcourir qu’à rebours la genèse de l’œuvre.”

<sup>47</sup>Klee, 1924/1998, pp. 23, 38, 60: “jusqu’à ce que la balance se stabilise.”, “L’œuvre d’art naît du mouvement.” and “La *forme* est fin, mort. La *formation* est Vie.”

<sup>48</sup>Kandinsky, 1926/1991, pp. 25, 29–30: “Le point géométrique est, selon notre conception, l’ultime et unique *union du silence et de la parole*. [...] Le point est le résultat de la première rencontre de l’outil avec la surface matérielle, le plan originel. [...] Par ce premier choc le plan originel est fécondé.”

<sup>49</sup>Cheng, 2006, p. 79: “Chacun des points a une existence propre; il promet de multiples transformations. Poser un point, c’est semer un grain; celui-ci doit pousser et devenir...”

## 1 Compositional paradigm

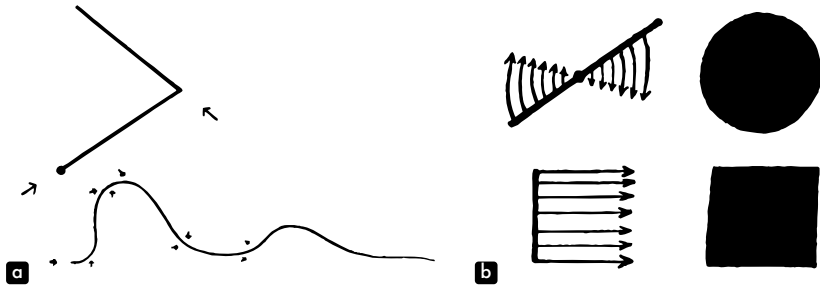


Figure 1.4: In panel **a**, morphogenesis of the line based on drawings by Kandinsky<sup>54</sup>. In panel **b**, morphogenesis of the surface based on drawings by Klee<sup>55</sup>.

another. [The point] perishes and a new being arises out of it which leads a new, independent life in accordance with its own laws. This is the Line.<sup>50</sup>

The point-line couple is then loaded with a unique generative symbolism. From a given point, similar to any other, all lines are born. This is one of the reasons why the brush-ink technique is central to the Chinese tradition. *Hua* generally refers to what we call *painting* in the West, because we put colors on the canvas, while this word originally evokes the drawing of boundaries, i.e. lines. It embodies the philosophical principle of *One and Multiple*, which Francois Cheng summarizes as follows:

The Line, by its internal unity and its ability to vary, is One and Multiple. It embodies the process by which the artist rejoins the gestures of Creation.<sup>51</sup>

More prosaically, graphical elements seem to be generated by a successive elevation of their dimensional order. The point is the primary unit, then “the line can be considered a secondary element”<sup>52</sup>, but the complexity does not stop there. “The time factor intervenes [...] the same way when a line generates a surface while moving.”<sup>53</sup> Kandinsky speaks of a densification of lines. Fig.1.4 illustrates the successive phenomena of emergence of lines and planes.

The theoretical framework for morphogenesis takes its inspiration from an obvious analogy with living beings. Later, it is echoed by scientific discoveries on DNA.

<sup>50</sup>Kandinsky, 1926/1991, pp. 62–63: “Il existe une autre force, prenant naissance non pas dans le point mais à l’extérieur. Cette force se précipite sur le point ancré dans le plan, l’en arrache et le pousse dans une quelconque direction. [...] Le point disparaît et il en résulte un être nouveau, vivant une vie autonome et soumis à d’autres lois. C’est la ligne.”

<sup>51</sup>Cheng, 2006, p. 73: “Le Trait, par son unité interne et sa capacité de variation, est Un et Multiple. Il incarne le processus par lequel l’homme dessinant rejoint les gestes de la Création.”

<sup>52</sup>Kandinsky, 1926/1991, p. 67: “la ligne peut être considérée comme un élément secondaire”

<sup>53</sup>Klee, 1924/1998, p. 37: “Le facteur temps intervient [...] de même lorsqu’une ligne engendre une surface en se déplaçant.”

<sup>54</sup>Kandinsky, 1926/1991, pp. 82, 104

<sup>55</sup>Klee, 1924/1998, p. 76

Together with the development of computer algorithms, some artists<sup>56</sup> consider coding and associated programs as a simulation of genetic determinism. We will see at the end of this section how this vision of the composition cannot embrace all the complexity of compositional practice.

To conclude our discussion of the temporal dimension, I would like to further emphasize how its deployment is articulated in a dual and complex way. In particular, the very term *composition* is ambivalent. This word represents both finished artworks and the practice supporting their elaboration. We can thus conceive composition as diachronic (compositional process) and synchronic (resulting composition). In addition, this transitional process seems to happen twice. In the viewer's eye, the perception of the painting offers a new sequential episode, eventually crystallizing in its memory and imagination as a paradoxically unified whole. It should be noted that, during the two major periods of the composition, the first segment can be considered as absolute, while the second is fundamentally a nonlinear exploration. Therefore, any attempt to model composition must address its temporal complexity by allowing a simultaneous, diachronic and synchronic, dual representation with a non-absolute sequential aspect.

### ***Structural and functional complexity***

Exposed morphogenesis lets us basically imagine *elementary* graphical elements, such as points, lines, circles. Without additional constraints imposed by the artist, this approach even leads to an infinite space of forms. Furthermore, each graphical element can take infinite configurations on canvas by varying its position, orientation, and scale. However, these sources of variation do not represent sources of complexity *per se*. They merely reflect different degrees of freedom.

Let us now tackle *complexity* within an established theoretical framework. First, we should check whether composition can be represented as a *system*. A system is, in its most general conception, an enclosing entity of substructures or modules. These constituent elements do not need to be perfectly defined, but they are supposed to interact with each other. For Kandinsky, even "forms with [objectively] *unspecified* relations will nevertheless engage in fundamental and precise interactions."<sup>57</sup> Without detailing the nature of these relations for the moment, composition seems to fulfill this first condition.

A system also requires precisely defined borders separating it from its environment. For pictorial composition, the canvas seems a trivial, but acceptable delimitation. The environment would then begin from the frame outwards, could include the exhibition scenography, and more importantly would welcome artists and spectators.

---

<sup>56</sup>Verostko, 1990.

<sup>57</sup>Kandinsky, 1912/1989, p. 194: "formes ayant entre elles un rapport [objectif] « quelconque » ont malgré tout, finalement, des relations importantes et précises."



## 1 Compositional paradigm

This limit essentially allows for exchanges at the interface to be formalized, and for potential functions of the system to be identified. Concerning composition, artists and observers are obviously privileged actors for interaction with the system. Indeed, it is on this basis that the cognitive abilities of artists and viewers can take part in the process of compositional complexity.

However, there is no transfer of matter at the composition-observer interface as usually happens with physical systems. The nature of this relation appears purely informational. Thus, from system theory to Shannon's information theory, there is only a small gap which many authors have tried to interpret<sup>58</sup>. More specifically, the notion of *amount of information* associated with a system is not straightforward and is possibly ambiguous. Information was initially defined quantitatively for a message in a communication channel. It is a quantification of the minimum number of signs, in a chosen language, necessary for the transcription of this message. Another way to think about information is in connection with the uncertainty associated with receiving a certain message. The more unexpected a code is, the more information it conveys to the receiver. For example, receiving a yes-no binary response intuitively contains less information than a message about *Which day of the week?*, which spans seven possibilities. Shannon expresses the amount of information via the entropy function  $\mathbb{H} = -\log_2 p$ , with  $p$  the probability of the expected code. For the binary answer, we then have  $\mathbb{H} = -\log_2(1/2) = 1$  and for a day of the week  $\mathbb{H} = -\log_2(1/7) \approx 2.8$ . We therefore understand that the amount of information depends on the realization of one event among possible ones. Assuming that the relation between a system and an observer is a communication channel, and that a system can exist in different states, the amount of information is then reflected by the uncertainty associated with the observation of a particular state for this system. We can thus imagine that a system with more elements will potentially exist in a larger number of different states, and that full characterization of such a system will be more challenging. Therefore, the amount of information carried by a system represents a possible measure of complexity. Depending on the point of view, this quantity can be interpreted as lack of knowledge in the eyes of the observer. This situation is particularly true in physics.

**This lack of information implies the possibility of a wide variety of distinct microscopic structures that are, in practice, impossible to distinguish from each other. Since any of these various microstructures can actually exist at a given time, the lack of information corresponds to a real disorder in the hidden degrees of freedom.**<sup>59</sup>

<sup>58</sup> Atlan, 1972/2006; Brillouin, 1959/1988; Le Moigne, 1977/2006; Moreno, 1998; Schrödinger, 1944/2013; von Foerster, 1960/2003.

<sup>59</sup> Brillouin, 1959/1988, p. 155: "Ce manque d'information implique la possibilité d'une grande variété de structures microscopiques distinctes qui sont, en pratique, impossibles à distinguer les unes des autres. Puisque l'une quelconque de ces diverses microstructures peut exister réellement à un moment donné, le manque d'information correspond à un désordre réel dans les degrés de liberté cachés."

Let us now transfer these concepts to the study of composition. We assume that there is a channel of communication between compositions and artists/viewers. We also assume that a composition is a particular state of all possible compositions constituting a single system. But intuitively, a composition does not objectively present indeterminacy concerning its state. Without time constraints, spectators have access to all graphical elements arranged on the plane. Forms are physically immutable. Perhaps a more relevant way to approach the problem could be to determine to which extent a spectator is able to perceive alterations in a given composition. The observer's perception is then probably not as unambiguous as we sometimes imagine. Fig.1.5 allows us to briefly experience this idea. At first glance, the two altered proposals (Fig.1.5b,c) seem to have a composition similar to the original (Fig.1.5a). In particular, identifying compositional differences or *weaknesses* in Fig.1.5b seems to require significant cognitive efforts. Affine transformations<sup>60</sup> of some elements only make subtle perceptual changes. In comparison, the removal of graphical elements (Fig.1.5c) is much more visible via the strange void that is left in the painting.

This naive examination allows us to assume some uncertainty, nonoptimality, in the perception of artworks. However, this is not enough to encompass the full extent of compositional complexity. Paintings appear to be flexible systems, given that *reasonable* affine transformations of some graphical elements produce hardly any impact on the compositional system. Some resilience happens in visible and spatial dimensions of the painting, across the degrees of freedom mentioned at the beginning of this subsection. The resilience results from a logic of non-absolute positioning of elements, of continuous ambiguity. Deletion of elements seems to further destroy local arrangements. Let us then consider more carefully Brillouin's quote. He writes that information unavailable to the observer can be located at the level of *hidden* degrees of freedom. As a result, the space of alternatives of a graphical element, whose absence would ultimately be the most radical state, could prove to be a transversal and unknown dimension to the spectator.

While alterations in Fig.1.5 have been produced randomly, it would be interesting to study alternatives of the same composition performed by the artist himself. Preparatory drawings and preliminary studies for paintings constitute a large corpus, but it might be objected that they are usually rough versions only, less definite than the final versions, only concerned with distributing main masses in a *schematic* way. Although ideally we should only compare alternatives produced with the same pictorial material, Kandinsky proposed several linear versions (only made of lines) of his own compositions. Fig.1.6 presents two pairs of compositions, where

---

<sup>60</sup>Examples of affine transformations: translation, scaling, symmetry, rotation.

<sup>61</sup>Guggenheim Bilbao: <https://www.guggenheim-bilbao.eus>

<sup>62</sup>Kandinsky, 1926/1991, pp. 235, –

<sup>63</sup>Centre Pompidou: <https://www.centrepompidou.fr>

<sup>64</sup>Miyagi Museum of Art: <https://www.pref.miyagi.jp>

## 1 Compositional paradigm

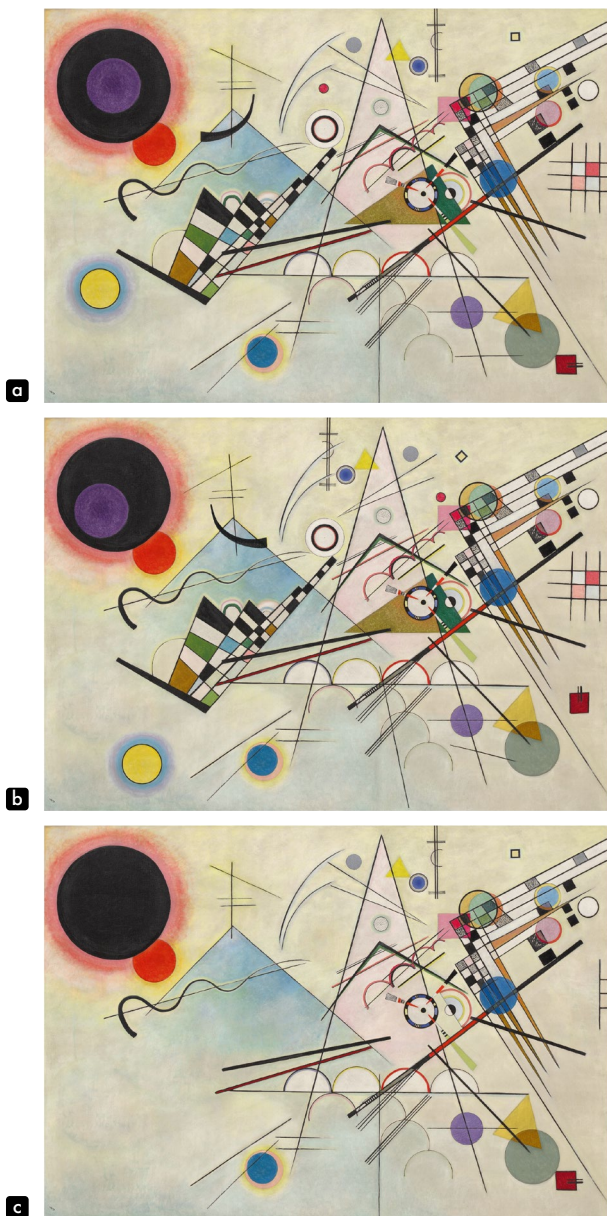


Figure 1.5: Compositional alterations of *Komposition VIII* (1923) by Wassily Kandinsky<sup>61</sup>. Panel **a**, original composition. Panel **b**, with affine transformations. Panel **c**, with deletions.

## 1.2 Sources of complexity

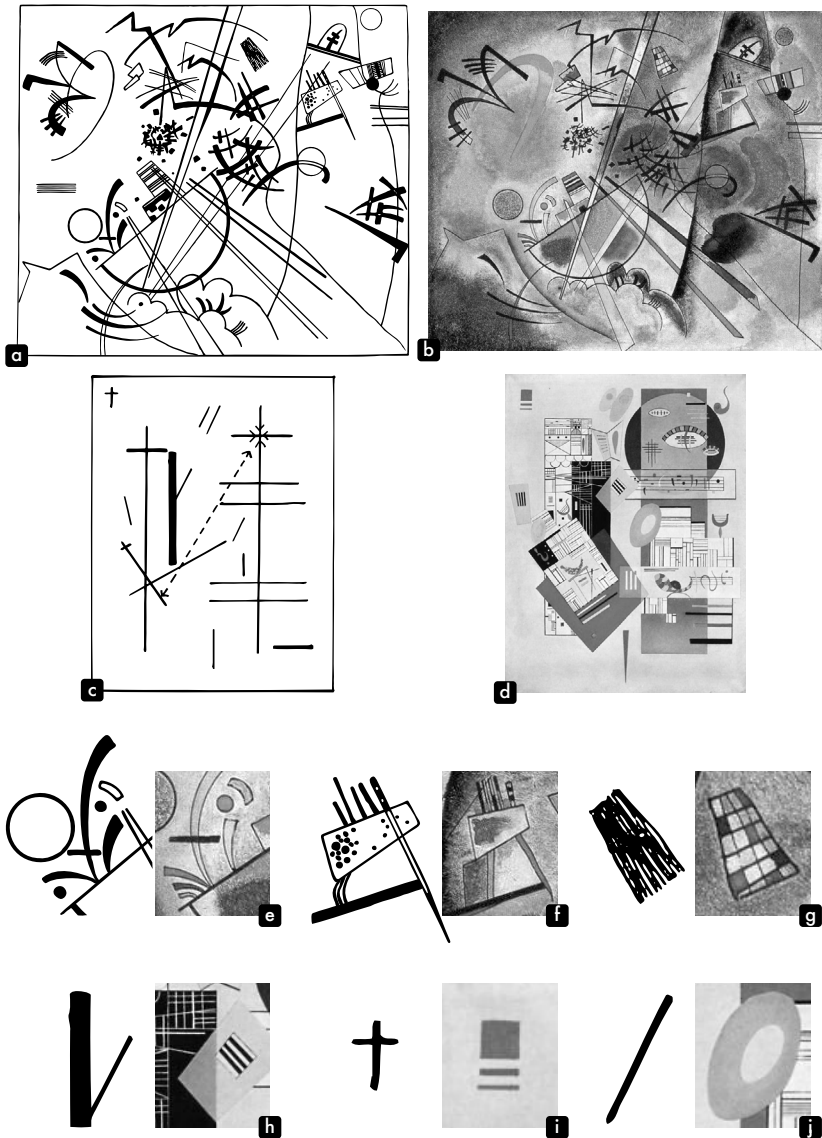


Figure 1.6: Linear and final compositions by Wassily Kandinsky. Panels **a** and **b**, *Small Dream in Red*, linear version (1924), final version (1925)<sup>62</sup>. Panels **c** and **d**, *Animated Stability*, linear version (1938)<sup>63</sup>, final version (1937)<sup>64</sup>. Panels **e-j**, comparison of selected graphical elements.

## 1 Compositional paradigm

Fig.1.6a,c are respectively the linear versions of Fig.1.6b,d<sup>65</sup>. For the first pair, the transposition seems straightforward as it only appears to involve removal of the pictorial matter. But the more detailed comparison in Fig.1.6e,f,g shows clear differences. In Fig.1.6e, the general structure is respected, but the proportion of each element is altered. In the lower left corner, a wide black dot is even substituted by two curved triangles. In Fig.1.6f, the overall composition looks similar, but the structure is not respected, especially for the right *leg*. Little dots filling the main surface are replaced by an indeterminate spot of paint. Finally, Fig.1.6g shows an unrelated texture, despite having similar morphology. Concerning the second pair of images (Fig.1.6c,d), the transformational effect is even more striking. The linear version undoubtedly offers the same compositional presence as the original, but the two versions are visually foreign. There is a flagrant simplification in the linear variant. The number of graphical elements is much smaller, but this fact alone is insufficient to establish compositional equivalence (Fig.1.6h,i,j): a cross corresponds to a square underlined twice, a straight line corresponds to two concentric ovoids.

We are therefore well aware of the volatility of graphical elements and their functions. It seems possible to effortlessly interchange different forms as long as they fulfill the same compositional purpose. In other words, identical graphical elements may have different uses depending on the context. Indeed, it is only on this condition that forms derive their legitimacy. For Kandinsky:

The same form in the same circumstances will always have the same inner appeal. However, circumstances are constantly varying. [...] Nothing is absolute. Form composition rests on a relative basis, depending on 1. alterations in the mutual relations of one form to another, 2. alterations in each individual form, down to the very smallest.<sup>66</sup>

Contextual informational uncertainty is formalized in system theory as follows:

The amount of information [(defined as entropy)] can only measure one kind of complexity: the one related to the large number of components arranged in a certain way in space. [There] is another kind of complexity [...]: the one related to the wide variety of interrelations between components. The first kind is called [...] structural complexity, the second, functional complexity.<sup>67</sup>

<sup>65</sup>These two artworks have been converted to black and white in order to reduce distracting effects of color, and to highlight compositional lines.

<sup>66</sup>Kandinsky, 1912/1989, pp. 127–128: "La même forme a toujours la même résonance sous des conditions inchangées. Cependant les conditions sont toujours différentes. [...] Il n'y a rien d'absolu. C'est pourquoi la composition des formes, qui repose sur cette relativité, dépend – de la variabilité de l'assemblage des formes et – de la variabilité de chaque forme jusqu'au plus petit détail."

<sup>67</sup> Atlan, 1972/2006: "La quantité d'information [(définie par l'entropie)] ne peut mesurer qu'une seule sorte de complexité : celle liée au grand nombre de composants disposés d'une certaine façon dans l'espace. [Il] est une autre sorte de complexité [...] : celle liée à la grande variété des interrelations entre les composants. La première sorte est appelée [...] complexité structurale, la seconde, complexité fonctionnelle."

In practice, it seems that functional complexity is even predominant, as it can already happen in situations of low structural complexity. In *Point and Line to Plan*, Kandinsky proposes to start the analysis of basic elements by studying each “individual phenomenon in isolation” and then the “reciprocal effect of phenomena”. However, he warns the reader about the conclusive third step: “I not only lack the strength to carry out the initial work with sufficient exactitude, but lack space, as well.”<sup>68</sup> Indeed, the catalog of all phenomena produced by few basic elements already implies an insurmountable combinatorial explosion. Unlike gravitational systems, for which the interaction among several elements can be predicted by knowledge of their individual characteristics, it appears that the combined effect of a group of graphical elements exceeds the sum of their fundamental properties. In cognitive sciences, *Gestalt* theory is precisely about the representation of a whole as being different and greater than the sum of its parts. In Kandinsky’s theoretical language, this notion is called *double-resonance* or even *triple-resonance*. Philippe Sers explains these neologisms with the desire to describe a “harmony based on a resonance not unified, but contrasted or at least disparate.”<sup>69</sup> The idea is that there is no hierarchical functional simplification occurring within the composition. A group of interacting elements does not cancel, or at least not completely, the individual capabilities of each unit.

As a result, speaking of *complex system analysis* is somehow improper. A system is not reducible to the enumeration of its elements and interactions. A system remains an experimental object, and that is why we will prefer to use *system exploration* in the second part of this manuscript. The notion of functional complexity also allows us to reinterpret the resilience aspect found in compositions, which we introduced earlier. Resilience may be a form of complicated and silent mechanism (hidden from many spectators) of functional equivalences that go beyond the purely morphological aspects of visual stimuli. More generally, presenting composition as a complex system in structure and functions gives theoretical grounding to many observations by artists. This research field helps to clarify the difficulties encountered when studying the compositional approach, difficulties that were not successfully addressed by analytical enumerations and Cartesian causal synthesis. The systemic view on composition is therefore a cornerstone toward modeling.

### **System complexity and system organization**

The amount of information – or lack of it, in the eyes of the observer – has been so far symbolized by the entropy function. This linguistic choice, referring to the idea of a chaotic disorder, may not be optimal to describe our simple ignorance of

---

<sup>68</sup>Kandinsky, 1926/1991, p. 21: “chaque phénomène isolé”, “effet réciproque des phénomènes” and “ce n’est pas seulement la force qui me manque, mais aussi la place, pour assurer au moins l’exactitude initiale.”

<sup>69</sup>Kandinsky, 1926/1991, p. 241: “harmonie fondée sur une résonance non pas unifiée, mais contrastée ou au moins disparate.”

## 1 Compositional paradigm

the current state of a complex system. Orderliness is not an intrinsic quality of a system, but rather a reflection of our inadequacy<sup>70</sup>. Viewed from this perspective, the qualifiers *order* and *disorder* confer rather unenviable objectives to pictorial composition. Implicitly, this vocabulary depreciates non-ordered systems for their chaotic aspect, and ordered systems for their cold or boring low informational level. In both cases, it is not possible to assign satisfactory meaning to the amount of information with regard to composition. These concepts do not offer a pertinent quantitative goal for compositional practice.

In the previous subsection, we have described the complexity of a system as relying on two components. However, so far we have only taken into account the structural complexity directly related to the system-observer channel of communication. By further taking into account functional information, it is possible to extend the notion of order to that of *system organization*. Henri Atlan introduces this concept in the following way:

The essential point is that each element possesses a set of alternatives associated with it, and that the choices in each set are not independent of those made in the others. [...] Organization is simply the informational amount associated with a set of constraints or correlations.<sup>71</sup>

Behind this formulation, there is a notion of conditional probabilities between graphical elements. In other words, the uncertainty associated with the global compositional structure can be extended by the *virtual* probability of encountering a graphical element at a position *knowing* the properties of the other elements. In the compositional processes engaged by artists, this scheme is easily applicable. Artists make successive choices depending on the elements already present (and/or planned according to their individual practice). Conditional probabilities thus establish an obvious link with the temporal dimension of the composition. It is also possible to imagine all compositional tensions interacting in the eye of the spectator as probabilistic constraints between the expected and present arrangements. This conception of the composition, as the organization of a system, also sets *boundaries* guiding possible objectives of compositional practice.

Total absence of constraints and total constraints between substructures, both correspond to *the absence of organization* of the system: in the first case we have only a juxtaposition of structures completely independent of each other, and in the second, we have only one structure replicated N times. [...] In other words, organization involves a transmission between substructures but with ambiguity or equivocation. Thus, we come to this seemingly paradoxical idea that the organization is better as ambiguity increases, up

---

<sup>70</sup>"Is there any real randomness, apart from our ignorance?" Atlan, 1972/2006 ("Existe-t-il un hasard réel, en dehors de notre ignorance ?")

<sup>71</sup>Atlan, 1972/2006: "Le point essentiel est que chaque élément a un ensemble d'alternatives qui lui est associé, et que les choix dans chaque ensemble ne sont pas indépendants de ceux effectués dans les autres. [...] L'organisation est simplement la valeur informationnelle d'un ensemble de contraintes ou corrélations."

to a certain limit where there is no more transmission at all and the organization disappears.<sup>72</sup>

Beyond statistical vocabulary, we can reformulate relevant notions as follows. If two groups of graphical elements have absolutely nothing in common, if they are neither metonymic nor antithetical, then there is no organization. We could cut the picture in two halves without disrupting anything. Conversely, if one graphical element necessarily involves another one, then the effect produced is as if the same information was reproduced twice. Ambiguity is certainly reduced, but nothing has changed from an organizational point of view for the system. At least, compositional motivations of this repetition should be questioned. In order to better understand the nature of the constraints between elements of a system, Atlan also proposes an analogy with books in a library. In this context, any reference, quotation, or commentary about the content of one book in another, constitutes a form of conditional relation allowing the library, as a whole, to express transversal knowledge. Concerning redundancy on the canvas, it can be naturally executed by an exact or partial repetition of elements and contrasts. However, we must not forget to extend the concept of redundancy to functionally equivalent clusters. One could even imagine a *temporal* redundancy, at the scale of a set of compositions by the same artist, as a form of *expected* regularity of certain groups of forms.

The organization of a system therefore involves a double counter-movement towards redundancy and variety. For the artist, composing entails a fight against both maximum entropy and uniformity. His/her task is delicate because he/she must neither order the elements on the plane in a too efficient and unequivocal manner for the spectator, nor abandon himself/herself to the dull and random placement of graphical elements. The compositional measurements at stake during the creation are therefore complex since they mobilize the evaluation of the impact of new elements *on* and *with* all others. Through moderate choices, which go beyond the most anticipated, automation and logic, he must seek heterogeneity of the parts, while preserving richness of interaction. The notion of system organization thus proposes a clear objective to the composition, that is artistically coherent, and with an associated metric, i.e. the amount of conditional information between graphical elements.

With the words of an artist like Klee, the transition from a purely structural understanding of the compositional practice to an expanded functional vision, is expressed as follows:

---

<sup>72</sup>Atlan, 1972/2006: "Absence totale de contrainte et contrainte totale entre les substructures, correspondent tous les deux à l'absence d'organisation du système : dans le premier cas on n'a qu'une juxtaposition de structures complètement indépendantes les unes des autres, et dans le deuxième, on n'a qu'une structure figurée N fois. [...] Autrement dit, l'organisation implique une transmission entre les substructures mais avec ambiguïté ou équivoque. Nous arrivons donc à cette idée en apparence paradoxale que l'organisation est d'autant meilleure que l'ambiguïté augmente, jusqu'à une certaine limite où il n'y a plus de transmission du tout et où l'organisation disparaît."



## 1 Compositional paradigm

Previously anatomical, the point of view is now becoming more physiological.<sup>73</sup>

We can also illustrate this concept with the words of two Chinese painters of the Tsing dynasty, Chin Tsu-yung and Shen Tsung-chien:

To represent a group of trees, make sure that there is not only balance between the trees, but also contrast; otherwise we fall into platitude and uniformity.<sup>74</sup>

The painting is populated with multiple elements that intersect or respond to each other. The important thing is that, in the middle of the tangled meanders, we can grasp an organic structure and that, within the most compact presence, we still breathe ease.<sup>75</sup>

In their own words, each author clearly evokes the duality of the concept of organization. Despite the figurative nature of their painting and a vocabulary borrowed from living beings, for these authors, forms seem inhabited by a fundamental organizational principle rather than by a purely representative logic.

I would like to mention here pictorial practices that appear to me outside the domain of this proposition, i.e. compositional practice as the organization of a complex system. At the same time, this allows me to justify the under-representation of theoretical thoughts from more contemporary abstract artists. I am particularly thinking to Sol Lewitt who wrote in *Paragraphs on Conceptual Art*:

The form itself is of very limited importance; it becomes the grammar for the total work. In fact, it is best that the basic unit be deliberately uninteresting so that it may more easily become an intrinsic part of the entire work. [...] Using a simple form repeatedly narrows the field of the work and concentrates the intensity to the arrangement of the form.<sup>76</sup>

The compositional aspect of his works is intentionally stifled by a higher principle, a concept, which reduces the functional strength of the graphical elements and lets them dominate the structure. Sol Lewitt is therefore less interested in organizing forms than in constructing them. At the opposite side of the spectrum, we should evoke Jackson Pollock's work, deliberately seeking to escape any organization and structure. His work proposes a graphical magma with its own qualities, but one that goes beyond the present definition of composition.

Finally, we would like to open a reflection on conceptual proximity with the idea of *self-organization*. The main characteristic of this particular type of system is to theorize the phenomenon of emergence<sup>77</sup>. It must be understood as the

<sup>73</sup>Klee, 1924/1998, p. 45: "Anatomique auparavant, le point de vue se fait maintenant plus physiologique."

<sup>74</sup>Cheng, 1989, p. 83: "Pour représenter un groupe d'arbres, veillez à ce qu'il y ait entre les arbres non seulement équilibre, mais aussi contraste ; sinon on tombe dans la platitude et l'uniformité."

<sup>75</sup>Cheng, 1989, p. 139: "Le tableau est peuplé de multiples éléments qui se croisent ou se répondent. L'important est que, au milieu des méandres enchevêtrés, on puisse saisir une structure organique et que, au sein de la présence la plus compacte, on respire cependant l'aisance."

<sup>76</sup>LeWitt, 1967.

<sup>77</sup>See in particular Moreno, 2004.

spontaneous appearance of a global constraint, a pattern, or a macrostructure, that was not predictable from the individual knowledge of sub-elements in a system. Emergence occurs everywhere in our physical world: from the wrinkles on the surface of dunes, to crowd dynamics, to every living organism being partially self-organized. For instance, this concept explains how brain complexity can exceed the amount of genetic information that permits its development. Indeed, DNA cannot materially encode all connections of a human brain, but it provides sufficient organization to brain cells so as to enable effective connections that are almost autonomously produced with the help of the environment (e.g. through learning). Self-organization reflects somehow every epigenetic phenomenon that favorably amends genetic determinism.

In the case of artistic composition, this concept may explain how an artist manages to overcome both structural and functional complexity during creation. For me, organizing graphical elements in a totally precise and conscious way seems cognitively out of reach. But artworks do exist, so artists may only need to create a macro specification of the composition, and then trust themselves, or trust the pictorial matter, to bring out the full composition during realization. In addition, the idea of *compositional self-organization* may be connected with, and expand upon, Klee's beloved concept of *painting organism*. Although exciting, these novel insights will not be explored in this thesis as they are not presently deemed mandatory for modeling composition. Nonetheless, I hope to investigate this idea further in future work.

### 1.3 Axioms

The only thing to ask a painter is to clearly express his intentions. His thinking will gain from this effort.<sup>78</sup>

The only moral constraint that theory therefore imposes on the modeler is to conduct a *a priori* verification: has he explained the few axioms upon which he will gradually support his inferences and create his design?<sup>79</sup>

The essential characteristic of an axiom is to be a supposedly true proposition without a real demonstration. However, propositions must be acceptable to the extent that they support the reasoning that requires them, i.e. our compositional paradigm. This section will therefore seek to introduce two essential concepts needed to develop our modeling framework for composition. We will first look at the idea of an artist's *works*<sup>80</sup> as a *hyper-compositional* object. This implies that

<sup>78</sup>Matisse, 2014, p. 99: "La seule chose qu'on doit demander au peintre c'est d'exprimer clairement ses intentions. Sa pensée y gagnera."

<sup>79</sup>Le Moigne, 1977/2006, p. 21: "La seule contrainte morale que la théorie impose dès lors au modélisateur est celle d'une vérification *a priori* : a-t-il explicité les quelques axiomes sur lesquels il va, progressivement, appuyer ses inférences et graver son dessin ?"

<sup>80</sup>Referring to the complete works of an artist, and corresponding to *l'œuvre d'un artiste* in French.

## 1 Compositional paradigm

all the artifacts produced by an artist weave a continuous *hyper-object*, in which the compositional regularities, specific to that artist, are revealed and accessible. Secondly, the hyperspace covering the works, and the hyperspace hierarchically included of graphical elements within the composition, are considered as vectorial and probabilistic spaces.

### Continuous space

Continuity is generally understood as the permanence of a phenomenon over time. In mathematics, this notion is extended to any  $f$  function whose infinitesimal variation of  $x$  inputs (e.g. time) is accompanied by an infinitesimal variation of outputs  $f(x)$ . There is no location where a phenomenon, an object or  $f(x)$ , can change in nature instantly (see Fig.1.7a, solid line). Continuity is therefore opposed to the notion of discrete structure, where the space of possibilities is filled with distinct states, such as the notes of a piano (see Fig.1.7a, dotted lines).

Let us now consider linear graphical elements only, regardless of their presence in a composition. In previous sections, we have detailed how these elements all arise from the point, from a material contact with the plane and the execution of free movements. Without any other constraint (from the artist or the modeler), this morphogenetic proposal allows the generation of an infinity of forms, and therefore implicitly, of all possible linear forms. This set of lines then constitutes a single and new continuous object, which completely determines the limits of the space that hosts it. By extension, we can talk about a continuous space of lines. At any point in this space, there is a line. At any infinitesimal neighborhood around this line, there is an infinity of other lines, all slightly different. Fig.1.7b illustrates this idea by displaying a portion of what a two-dimensional space of lines could look like<sup>81</sup>.

The generative principle at stake forces us not to reduce a phenomenon to existing or observed occurrences only. It intrinsically guarantees a potential, complete and continuous space. Let us particularly recall the *One and Multiple* Chinese principle mentioned previously (see Subsection.1.2.Temporal complexity). In this sense, the composition also has its own genesis and follows a particular generative process. Then, why not consider all compositions of an artist as belonging to the same entity? An entity ultimately close to what is commonly called his/her *works*. We are indeed able to imagine intermediaries between any two compositions. Of course, we could question the relevance or the interest of such intermediaries, but we cannot invalidate their theoretical existence<sup>82</sup>. Thus, we assume that all compositions and graphical elements can be represented as infinite and continuous spaces.

---

<sup>81</sup>Illustration is sadly discrete for legibility purposes. Representing continuity is actually an artistic challenge, and we will address this question in Chapter.6.

<sup>82</sup>This discussion will happen later in this section.

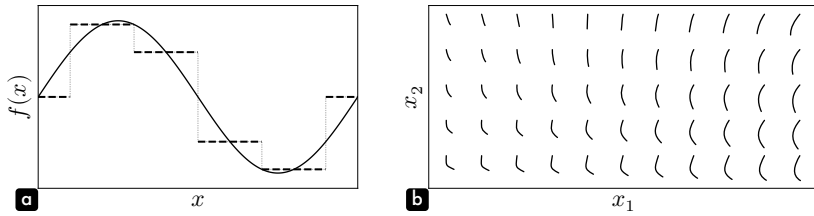


Figure 1.7: Continuous space. In panel **a**, solid line represents a continuous function, while the dotted lines show its discrete version. Panel **b** is an illustration of a portion of what could be a two-dimensional space of lines.

### Hyper-composition

Meaning *beyond*, the prefix *hyper* basically refers to theoretical objects in  $n$  dimensions with  $n$  typically  $>3$ . The hypercube is, for example, the  $n$ -dimensional cousin ( $n > 3$ ) of the cube ( $n = 3$ ) and the square ( $n = 2$ ). It then becomes obvious that *hyper* and *higher* do relate to any aesthetic values about the objects in question. An additional dimension is a mathematical property which, even abstractly manipulated, pragmatically offers new possibilities.

Then, what is the relationship between *hyper-forms*, hyperspaces and the presumed continuity of compositions? It should be noted that in weaving continuity between objects, we implicitly create a higher-order set-object, which is necessarily inscribed into a higher-dimensional space. This encompassing space must be increased by at least one dimension in order to allow that a point on this new dimension corresponds to an initial object of a lower-order. The ability to move from one object to another continuously must be considered as a new degree of freedom for this object, hitherto implicit.

However, our cognitive system is so inscribed and continuously immersed in a 3-dimensional universe, that imagining an object in  $n$  dimensions is a cognitively complicated task. Our perception limits somehow our ability to think in the abstraction of hidden dimensions. Referring to the dimensions of the painting, Klee highlights this problem and gives us some advice.

The instrument is lacking that would make it possible to synthetically discuss a multidimensional simultaneity. [...] It seems that the ever-increasing number of dimensions requires ever more difficult efforts [...], we must exercise great patience.<sup>83</sup>

In his famous book *Flatland: A Romance of Many Dimensions*, Edwin A. Abbott tries to express the vertigo caused by dimensions exceeding us. To do this, he

<sup>83</sup>Klee, 1924/1998, p. 18: "L'instrument manque qui permettrait de discuter synthétiquement une simultanéité à plusieurs dimensions. [...] Il semble que le nombre sans cesse croissant des dimensions demande des efforts toujours plus ardues [...], il s'agira d'avoir beaucoup de patience."

## 1 Compositional paradigm

takes the viewpoint of a square living in a 2-d world, once having the chance to glimpse into the third dimension. The author also likes to describe the perception of lower-dimensional object, such as a point in 0-d.

That Point is a Being like ourselves, but conned to the non-dimensional Gulf.  
He is himself his own World [...]; he has no cognizance even of the number  
Two.<sup>84</sup>

However, Focillon maybe gives a more informative insight on higher hidden dimensions in *The Life of Forms*:

The inhabitant of a two-dimensional world could own the whole series of profiles of a given statue and marvel at the diversity of these figures, without ever understanding that it is one, in relief.<sup>85</sup>

We must object that, in theory, a two-dimensional inhabitant of *Flatland* could not even see the different profiles as solid shapes on a plane (as presented in Fig.1.8a). He could only touch the contours and mentally build up an image. Nevertheless, we do hope that this inhabitant is indeed able to imagine that all these profiles belong to a single volume (e.g. Fig.1.8b), because this is precisely what our research project attempts to achieve for a collection of compositions from the *works* of an artist.

Continuity therefore implies hidden dimensions, which require a certain effort to be convinced of. In reality, it requires even more work to reconstruct them in a relevant way. Rebuilding a whole according to fragments has its share of indeterminacy. Profiles shown in Fig.1.8a comes from the same 3-d object, a Moai from Easter Island, shown in Fig.1.8b. In this object, each section along a plane (e.g.  $a$ ) produces two symmetrical profiles (e.g.  $a$ ,  $a^{-1}$ ) that can appear independent to a viewer not aware of the complete object. Assuming that this ground truth is inaccessible, there are many ways to reconstruct a higher-order object. For instance, assuming a cylindrical 3-d object, it is possible to arrange profiles uniformly every  $\theta = 60^\circ$  (see Fig.1.8c). Missing information is simply interpolated. The 3-d result is different from the ground truth, but similar in nature. From this *hyper-object*, we can then make new slices. For example, the section of Fig.1.8d produces the profile of Fig.1.8e, which is consistent with original profiles. Nonetheless, another possible reconstruction is to consider the different profiles as photographs over time (see Fig.1.8f). The true morphology of the 3-d object is completely lost, but a new cut at another moment  $t$  gives a rather convincing profile. The temporal aspect is actually artificial, and no dimension is *a priori* preferred to operate a new section. Produced crosswise (see Fig.1.8h), the resulting profile shown in Fig.1.8i has then nothing in common with initial fragments. Thus, we realize that this  $n$ -dimensional reconstruction process does

---

<sup>84</sup>Abbott, 1884, p. 81.

<sup>85</sup>Focillon, 1934, p. 28: "L'habitant d'un monde à deux dimensions pourrait posséder toute la série des profils d'une statue donnée et s'émerveiller de la diversité de ces figures, sans se représenter jamais que c'est une seule, en relief."

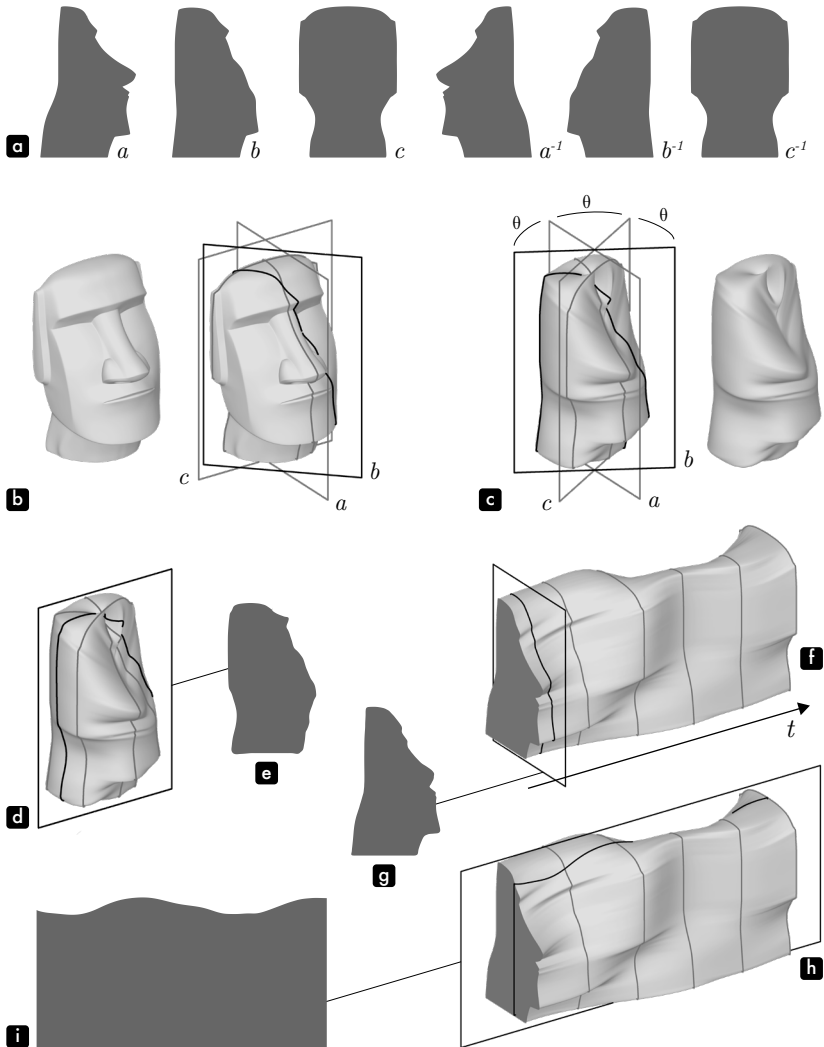


Figure 1.8: Hyper-dimensional reconstruction. Panel **a** shows 2-d profiles, which are slices along planes ( $a$ ,  $b$ , and  $c$ ) of a ground truth object, a Moai, displayed in panel **b**. Supposing this true object inaccessible, there are several ways to reconstruct a 3-d object from panel **a** profiles. With the assumption of a cylindrical object (panel **c**), profiles are arranged every  $\theta = 60^\circ$ . Missing information is then interpolated. From this 3-d object, we can now make a new slice (panel **d**), resulting in a new profile (panel **e**) coherent with the original ones. Another possible reconstruction is a sequential arrangement of profiles along a temporal dimension  $t$  (panel **f**). The 3-d object have nothing in common with the true object, however slices along  $t$  produces good new profiles (e.g. panel **g**). On the contrary, a transversal slice like in panel **h**, provides a spurious result (panel **i**).

## 1 Compositional paradigm

not have a single solution. When a ground truth is not accessible, this approach requires a conscientious experimental control to be relevant.

Let us return to *hyper-composition*. In our proposition, a composition is considered as an incomplete view from a richer space. However, unlike perspective in figurative paintings, where an image is a 2-d flattening of a 3-d space, a composition is not a projection in fewer dimensions of a hyper-composition. A composition must rather be understood as a planar section of a space in  $n$  dimensions, whose additional dimensions are basically not spatial, nor temporal, but specifically compositional. Nevertheless, to imagine these hidden dimensions, to visualize them for ourselves or to show them to spectators, it is possible to *render* them as spatial and/or temporal. This representational question is obviously artistic. It is both a source of wonder and a visual challenge for the artist. We will further explore this aspect in Chapter.6.

### Vectorial space

Although it is conceivable that a composition may be viewed as a planar section of a hyper-object, this conception is not practical from a mathematical and computational point of view. Indeed, a theoretical proposition is all the more relevant as tools for manipulating this new object exist. This is why we assume that compositions can be represented as a point, and more precisely as a vector, in a vectorial space.

To grasp associated advantages, we will set compositions aside for now and focus instead on colors. Individually, light waves can be considered as colored according to their wavelength. Evidently restricted to the visible light spectrum, the resulting range of possible colors would be rainbow colors (see Fig.1.9a). But it does not reflect all the perceptual color domain. For instance, magenta and white do not exist as single light rays. Colors are therefore the result of complex combinations of wavelengths. Colors necessarily involve *energy* weighting of light rays that are present in the observed beams. Despite being accurate, this physical representation is difficult to manipulate. Communicating a color to another person would be extremely challenging. We should transmit measurements of hundreds (actually an infinity) of numerical values (see the 3 complicated spectra plotted in Fig.1.9a).

We should then prefer a color definition that makes more sense to human perception. In our eye, light beams are integrated (filtered) by three types of photoreceptors (i.e. cone cells). The brain is therefore able to interpret colors from a tri-stimulus only. Moving a little away from biological truth, we could say that it is possible to represent any color by three coordinates in a 3-dimensional vectorial space with e.g. red, green, and blue primary colors/dimensions. In other words, a color spectrum, which is a complex set of light rays, can be *encoded* by an artificial system (e.g. a digital camera) into a 3-dimensional vector (see Fig.1.9b, where the three spectra

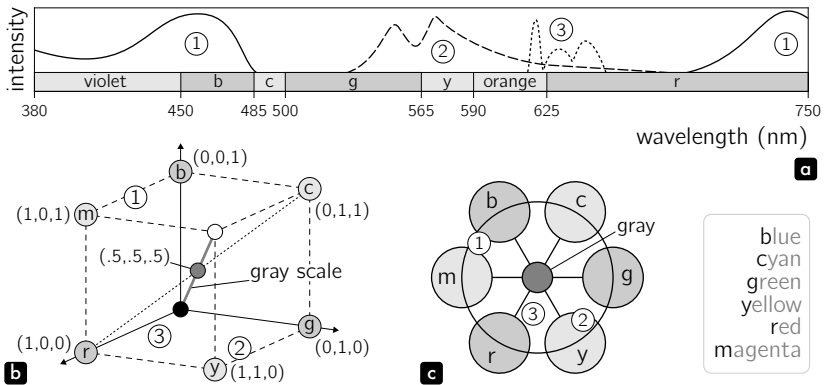


Figure 1.9: Color spaces. Three colors, denoted (1), (2) and (3), are represented in different ways. Panel **a** shows their light spectra, limited to the visible wavelengths in nanometers. Panel **b** is a vectorial 3-d (RGB) representation of color space. Panel **c** is the traditional color wheel, which can be considered as a 2-d color space, restricted to chromaticity (luminance excluded). In other words, panel **c** is a flattened version of panel **b** along the gray scale axis.

from Fig.1.9a correspond to three individual points). Thus, this representation is a compact abstraction of the initial object, while guaranteeing a certain correlation between its physical reality and its digital transposition.

A vector space is also *functional*, because it presupposes that some mathematical operations are allowed in this space. For instance, it is possible to obtain the mix of two colors by adding their vectors, i.e. by adding the numerical values in each dimension (e.g.  $r[1, 0, 0] + b[0, 0, 1] = m[1, 0, 1]$ ). Fig.1.9c presents the traditional color wheel that can be considered as a two-dimensional version of an RGB space, flattened along the luminance axis. In this representation, it becomes apparent how the addition of complementary colors converges towards gray. However, a general caveat of vectorial representations is that there is no guarantee that a vectorial summation is consistent with perceptual reality. Dimensions may not be independent, and may not be relevant. These issues can only be addressed adequately through experimental verification.

Unlike the color domain, the form domain lacks a well-defined inventory of reference points. Thus, the circle of colors, well-defined at the beginning, is radically different from the unfathomable blur of the infinity of forms.<sup>86</sup>

Philippe Sers' remarks should be qualified to emphasize that color space is just as infinite as form space. However, this author implicitly (and correctly) assumes that the infinity of shapes is greater than the infinity of colors because of the greater

<sup>86</sup>Kandinsky, 1926/1991, pp. xiii–xiv: "Contrairement au domaine des couleurs, dans le domaine des formes, l'inventaire n'est pas donné. Ainsi le cercle des couleurs, bien défini au départ, se distingue-t-il radicalement de l'insondable flou de l'infini des formes."



## 1 Compositional paradigm

number of dimensions, and is thus more difficult to conceive rationally. In addition, no physical reality (like the three types of diurnal receptors of the eye) provides an objective vectorial representation of forms and compositions. There is for instance a great indeterminacy about the minimal number of dimensions to cover all possible phenomena. Nevertheless, we will assume that it is possible to represent a composition, as well as its graphical sub-elements, with  $n$ -dimensional vectors. The following chapters of this manuscript will basically describe how to build tools operating this conversion between the different domains (physical/vectorial) and show the legitimacy of this axiom through practical experiments.

### *Probabilistic space*

From the beginning of this chapter, we have hypothesized an infinity of possible compositions and their continuity in a vectorial space. However, without additional constraints, the resulting hyper-compositional object does not really enrich our knowledge about composition. As already mentioned, there exists an infinity of possible hyperspaces, all describing the same reality, so the reconstructed spaces would still be unspecified.

Travelling across such spaces randomly, many locations would appear *uninteresting*. We would be surprised by the qualitative heterogeneity of this space. Indeed, even if all arrangements of graphical elements are theoretically possible, composition arises from precise choices made by the artist. Some of these choices are taken to the detriment of others. This principle, at best described as the *inner necessity*, mechanically reduces *true* compositions to a subset of the complete space. However, we have forbidden ourselves from making any direct aesthetic judgment on compositions, so we must seek for a more satisfactory formulation.

Perhaps, by the term *uninteresting* we essentially designate arrangements of forms that do not *globally* resemble other compositions – their characteristics are not identifiable, or they do not belong to a certain standard. So, these *intermediate* compositions may not present a minimal amount of references to some implicit compositional regularities. For Kandinsky, there is no in-between, a painting is alive or it is not. A less polarizing viewpoint would refer back to the variability of human perception. A quantitative metric of the deviation from a compositional norm cannot therefore be binary, nor uni-dimensional. The compositional norm must be a rich norm, proposing many ways to comply with it.

This way, we finally draw the contours of a probabilistic constraint of the space, where each dimension would reflect different expressive regularities, and follow their own law of probability. Concretely, it means that each value of a given dimension is assigned a probability for this characteristic to be expressed in a composition. As a result, hyper-composition could take shape, and meaning, by an objective densification interplay, freed from any personal judgment. It would

involve transposing the *inner* metric to a measure of density in hyperspace. The relevance of a graphical proposition would be executed against an ensemble, as the evaluation of the sincerity of a composition in relation to the implicit rules set by the artist himself/herself through his/her complete works.

Nevertheless, let us immediately clarify that the probabilistic nature of compositional space must not be confused with, or interpreted as, a creative objective. There is no interest, nor optimality in producing the most likely composition<sup>87</sup>. In this sense, my vision does not align with that advanced by scientists who have already realized this type of connection between art, aesthetics, probabilities, and information theory.

To efficiently encode previously viewed human faces, [...] it is useful to generate the internal representation of a prototype face. To encode a new face, it must only encode the deviations from the prototype. Thus, a new face that does not deviate much from the prototype will be subjectively more beautiful than others.<sup>88</sup>

For Juergen Schmidhuber, *beauty* therefore lies in the average object, in the most expected one. The problem is that beauty is then reduced to the most immediate and easiest cognitive action, with extremely poor informational content. Schmidhuber also thinks that the creation of this mental prototype is what supports our interest. Once known, beauty becomes boring. Despite being an accurate description of visual pattern learning as a cognitive phenomenon, the aesthetic implications of this conception are too narrow and limited to be artistically compelling. In addition, this theory only applies to uni-dimensional representations of the object under scrutiny. Such representations may be applicable to faces under some conditions, but they are certainly inadequate for pictorial compositions. The more an object relies on numerous independent dimensions of potential regularities, the more the very existence of a single global optimum is questionable.<sup>89</sup>

With their own vocabulary, each successive artist expresses the same subtle and intermediate position regarding compositional laws and norms:

Dissonances, sometimes even false notes, are possible, but they must be used with great caution; otherwise, harmony risks to be destroyed.<sup>90</sup>

Precious is the knowledge of laws, provided you beware of schema that confuse simple laws with living reality.<sup>91</sup>

<sup>87</sup>This is a common pitfall of some computational approaches to art, as emphasized by Audry, 2021 in his book *Art in the Age of Machine Learning*.

<sup>88</sup>Schmidhuber, 2009.

<sup>89</sup>We will detail this affirmation in Subsection 4.2. Hyperspace density.

<sup>90</sup>Sérusier, 1921, p. 34: "Les dissonances, parfois même les fausses notes, sont possibles, mais il faut s'en servir avec beaucoup de prudence; sans cela, l'harmonie risque d'être détruite."

<sup>91</sup>Klee, 1924/1998, p. 51: "Précieuse est la connaissance des lois, à condition de se garder d'un schématisme confondant loi nue et réalité vivante."

## 1 Compositional paradigm

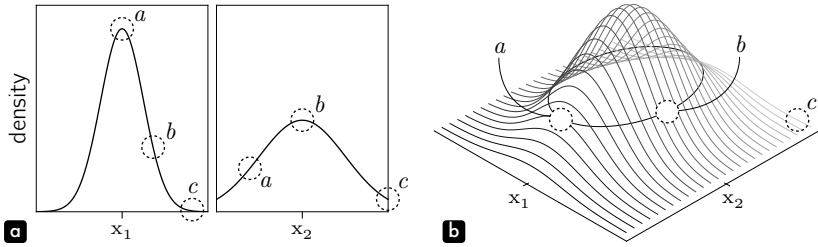


Figure 1.10: Two-dimensional probabilistic space of independent random variables  $x_1$  and  $x_2$ . The individual (marginal) probability density functions are plotted in panel **a**. Panel **b** shows the joint distribution of  $x_1$  and  $x_2$ . Points *a*, *b* and *c* are plotted in both representations.

**Blind following of scientific precept is less blameworthy than its blind and purposeless rejection.<sup>92</sup>**

These observations are indeed similar to the one associated with the organization of a system. Artistic attitudes towards regularities typically compromise between the very probable and the improbable, between the total constraint and absence of constraint. The hyper-compositional object is in fact itself a system. Compositions within the artist's works need to maintain a certain level of connections, i.e. to share regularities, while avoiding repetition, which would impoverish the overall amount of information, i.e. the richness of the resulting works.

Let us try to illustrate this idea concretely. For simplicity, we assume a two-dimensional compositional space only, with  $x_1$  and  $x_2$  as its independent dimensions. Fig.1.10a shows density functions for each dimension. We also assume these distributions to be *normal*. This assumption is reasonable (based on the central limit theorem<sup>93</sup>) and can be found in many everyday life phenomena. Normal distributions are symmetrical with greater central density. In Fig.1.10a,  $x_1$  is more tightly distributed around zero than  $x_2$ . Fig.1.10b displays the joint probability of the two dimensions. On this surface, a black circle defines the location of compositions with the same probability at the scale of the hyper-composition. We are particularly interested in two compositions *a* and *b* located on this circle of equiprobability. Detailed analysis of marginal probabilities (Fig.1.10a) shows that *a* and *b* are averages for one dimension, but not for the other. In particular, *a* frees itself from the norm on  $x_2$ . The *risk* taken in this dimension by the artist can be considered as particularly informative since, in information theory, the unexpected produces meaning. However, as *a* is *excessively* banal on  $x_1$ , at the hyper-compositional level this composition is still relevant. *b* is also just as relevant as *a*, but for different reasons, in another dimension. It therefore

<sup>92</sup>Kandinsky, 1912/1989, p. 199: "L'observance sans but des règles scientifiques n'est jamais aussi nuisible qu'un inutile renversement de celles-ci."

<sup>93</sup>The central limit theorem proves that the sum of various noises with unspecified distributions converges to a normal distribution.

seems incorrect to believe that a probabilistic vision of art would necessarily be normative or limiting. Our proposition in fact gives meaning to certain regularities, while allowing many degrees of freedom. It ultimately allows artists to make choices with a clear *conscience*. The global low density of composition  $c$  finally echoes Kandinsky's warning reported above. When an artist rejects laws across all dimensions simultaneously, his/her actions seem purposeless and *erroneous*, as they reject their own higher-level coherence.

Finally, we must not forget that this illustration is two-dimensional only. The concept of hyper-composition must be imagined in  $n$  dimensions. A complete demonstration will be produced in the Subsection.4.2.Hyperspace density, but in short we can say that the more dimensions there are, the more statistically unlikely it is to have a composition in the average for all dimensions. Therefore, the fear of normalizing compositional objects by revealing their regularities does not find theoretical justification.

### **Probabilistic plane**

The temporal dimension of the composition, particularly in relation to graphical elements considered as individual entities, has not been addressed yet. At the level of the hyper-compositional object, we only explicitly addressed *completed* artworks. However, inner compositional properties on the plane are necessarily related to higher level regularities. We must discuss *how*.

We have seen that directly studying interactions between pictorial elements is not analytically feasible. The combinatorial complexity of the task prevents us from an exhaustive approach. In addition, the functional complexity exposed earlier makes these interactions highly contextual. In practice, the main difficulty with capturing probabilistic constraints at the scale of one compositional plane derives from the fact that only a reduced number of graphical elements are presented for a given composition, and only in a limited number of configurations. Under these conditions, we are not able to capture local regularities. As Klee rightly says, "the typical will come automatically from series of examples."<sup>94</sup> In other words, a pattern can only be studied if it is repeated in many different contexts. Reciprocally, we only build up a distribution corresponding to a set of elements, if this context gives rise to enough alternatives at the hyper-compositional level. Indeed, only series of completed compositions can carry objective information about local interactions. In isolation, no subset of graphical arrangements can claim to be legitimate, but in comparison to the norm of a coherent whole, a specific choice can be measured and make sense. Captured hidden dimensions of the hyper-dimensional object then reflect *inter* and *intra* compositional regularities.

---

<sup>94</sup>Klee, 1961, p. 22.

## 1 Compositional paradigm

Let us rephrase this idea from the artist's perspective. When he composes, he is confronted with a series of choices among many possibilities, and the measurement of the *pertinence* of a particular graphical element results from an *inner* assessment. This judgment is not only based on current context, but also constrained by some global knowledge: all compositional laws are supported by his/her own unconscious regularities. It is as though local probability fields of possible next elements were redistributed according to a final mental objective, incorporating this implicit knowledge in the process. Through an initial choice of a global location in the hyper-compositional space, the artist actively reconfigures the possibilities offered on the plane of the painting. Therefore, this plane must be understood as a complex field of conditional probabilities, knowing a global position in the artist's works and a local context on the plane.<sup>95</sup>

### Active emptiness

We have so far only developed a vision *in positive* of artistic actions. Compositional activity is usually understood as *filling* space, so discussing the concept of *emptiness* may seem paradoxical. For instance, it appears fundamentally impossible to represent the absence of pictorial matter within a modeling framework: how can emptiness fit the probabilistic view described above?

We model clay to make a vase, but it is the emptiness within; which retains what we pour into it. [...] We work with the being; but it is the non-being that we use.<sup>96</sup>

Lao Tzu observes that actions, necessarily defined as *positive* operations, may lead to functions operating *in negative*. In the pictorial domain, Fan Chi of the Tsing dynasty, specifies:

It is generally believed that it is enough to save a lot of unpainted space to create emptiness. What is the point of this void if it is an inert space? In a way, the true Void must be more fully inhabited than the Full.<sup>97</sup>

To become useful, emptiness must be active. But how? Fan Chi subtly indicates that an empty space is not empty because of being unpainted, but rather because it should/could have been painted. It therefore seems relevant to imagine emptiness as a place of high probability for the presence of one or even multiple graphical elements. However, this potential must remain unfulfilled, so that plane emptiness

<sup>95</sup>We are aware that the presented idea is difficult to grasp. This concept will be clarified via direct implementation (Section.3.4) and associated practical measurements (Section.4.3, in particular Fig.4.32 and Fig.4.33).

<sup>96</sup>Lao Tseu, 2008, chap. 11: "Nous modelons de l'argile pour en faire un vase,; mais c'est le vide au-dedans; qui retient ce que nous y versons. [...] Nous travaillons avec l'être.; mais c'est du non-être dont nous avons l'usage."

<sup>97</sup>Cheng, 1989, pp. 45–46: "On croit en général qu'il suffit de ménager beaucoup d'espace non peint pour créer du vide. Quel intérêt présente ce vide s'il s'agit d'un espace inerte ? Il faut en quelque sorte que le vrai Vide soit plus pleinement habité que le Plein."

is not an inert place. It is the location of a probable pictorial emergence, fully in tension by the lack it produces, or by a deliberate deviation from the norm chosen by the artist. For Klee, some over-neutral pictorial elements, such as the gray point, can possess an equally paradoxical nature.

This point is gray because it is neither white nor black or because it is white and black at the same time. [...] Gray because it is a non-dimensional point, a point *between* the dimensions and at their intersection, at the junction of paths. [...] It is the Center of Everything, virtually containing any color, any value, any line.<sup>98</sup>

Gray is for Klee the master of all averages. It embodies the most expected, but at the same time a potential generative magma of all possibilities. We could therefore consider the gray point as some kind of informational emptiness.

In conclusion, emptiness and its complementary inherent active potential, is actually another peripheral insight in favor of a probabilistic approach to art materials and a demonstration of the relevance of the associated concept of hyper-composition.

This pure canvas [...] is itself as beautiful as a painting.<sup>99</sup>

Nonetheless, beyond the poetry of Kandinsky's words, for potential to exist in emptiness, the hyper-compositional object and its complex projection on the plane must be filled and nourished with real artistic matter, such as a personal practice.

---

<sup>98</sup>Klee, 1924/1998, pp. 56, 51: "Ce point est gris, parce qu'il n'est ni blanc ni noir ou parce qu'il est blanc tout autant que noir. [...] Gris parce que point non-dimensionnel, point *entre* les dimensions et à leur intersection, au croisement des chemins. [...] Il est le Centre de Tout, contenant virtuellement toute couleur, toute valeur, toute ligne."

<sup>99</sup>Kandinsky, 1974/2014, p. 115: "Cette toile pure [...] est elle-même aussi belle qu'un tableau."



## 2 Personal practice

In the previous chapter, I have described the overarching paradigm supporting my research. In principle, I wish for those propositions to be as general as theoretically possible. In practice, they are necessarily tailored to my personal practice and experience of composition. For instance, the feasibility of any modeling strategy must rely on the existence of compositional regularities among artworks that are sufficiently consistent to cohere into a unified object, however this requirement is not a universal artistic intention: artists may not create new works with the intention of complying with regularities set out by previous work. In order for my research program to retain any degree of feasibility, it is therefore necessary to restrict its applicability to a specific kind and range of composition. This chapter will detail the specific choices that have gone into the process of delineating and contextualizing my approach through the description of my personal corpus, alongside the processing steps that were necessary in order to make it available to quantitative analysis.

Data requirements and my personal work are in theory two different issues, but in practice they are intimately intertwined. My artistic approach has directly shaped the proposed method of research, and current limitations of existing computational tools have constrained the scope of inclusion for external artistic materials. If one were to challenge the scientific legitimacy of a dataset arising from my own personal experience on the basis that it is not representative of composition in general, the main argument in its defense would be that this is a necessary point of departure: without it, the entire research program becomes a practical impossibility, despite its theoretical existence beyond my own personal experience. *A minima*, we must regard my own personal dataset as a starting point, a proof of concept, a stepping stone towards future generalization of this research approach to a greater range of compositional efforts.

The first limitation imposed by state-of-the-art computational tools is their data hunger. Machine learning, and especially deep learning, requires a large number of samples from the same data family in order to make pertinent discoveries. Trivially, this means that we must acquire thousands of inputs from the same artist. To start with, it is unlikely that any artist would dedicate such a huge amount of time and effort to produce material of this kind. Even if such a devoted artist were to be found, the end result would be an artificial representation of the artist's creative work, not rooted in his/her own practice. Some deep



## 2 Personal practice

learning models aggregate paintings from several centuries<sup>1</sup> to generate novel images, but what does this tell us about composition? All those artists may only have in common basic ratios and simple symmetries. Furthermore, trends arising from these approaches hide inherent bias associated with the process of selecting datasets, which mainly consist of occidental paintings and largely ignore non-Western traditions.<sup>2</sup>

Besides the necessity to restrict inquiry to a consistent and sizeable compositional corpus, there is an additional characteristic that makes it necessary to focus on my own personal corpus. The practical feasibility of the novel research program presented in this thesis calls for a *minimal artistic grammar*, where *minimal* should not be interpreted to mean *limited*. Our intention is to address composition in all its complexity. For Le Moigne:

It is now necessary to consider the modeling of any *phenomenon perceived and conceived as complex* by the refusal of its simplification, of its mutilation.<sup>3</sup>

In the same vein, modern artists have turned to abstraction to avoid representational issues or, like Kandinsky, they have adopted simple geometrical forms. This was not an imposed simplification, but rather the appropriate vocabulary for addressing their pictorial questions at a deeper level. We wish to adopt a similar vocabulary for our own research program, a symbolic structure within which to discover and express the compositional process. Where is such a material to be found?

We cannot – at least for now – take an existing masterpiece and automatically extract its compositional structure to any acceptable degree. To express this naively, we cannot reduce the complexity of art material by applying an edge detection algorithm to its luminance pattern<sup>4</sup>. A more sensible start would involve the analysis of preparatory sketches and studies from a given artist, however current availability of such material is hopelessly insufficient for the purposes of our research program: even a relatively large corpus of this kind, such as Picasso's *Meninas* series involving nearly 60 exemplars, is infinitely smaller than the size required to successfully support machine learning.

In a different approach, we may take the reductionist view to its extreme, until we encounter psychophysics. In this discipline, elementary visual stimuli are regarded as adequate stimuli for the investigation of low-level visual phenomena in the brain<sup>5</sup>. It is, for example, legitimate to work with very simple polygons<sup>6</sup> to study the early mechanisms of shape perception. When this approach is translated to art, however,

---

<sup>1</sup>A. Elgammal et al., 2017.

<sup>2</sup>Most of the time, datasets are chosen to be subsets of WikiArt, n.d.

<sup>3</sup>Le Moigne, 1977/2006, p. 16: "[Il faut] entendre désormais la modélisation de tout *phénomène perçu et conçu complexe* par le refus de sa simplification, de sa mutilation."

<sup>4</sup>Y. J. Lee et al., 2011; Redies et al., 2017.

<sup>5</sup>Neri, 2014.

<sup>6</sup>Behrman and Brown, 1968.

## 2.1 Floating compositional structures

it appears lacking. It is true that some art works only involve a few straight lines and even sometimes an *empty* canvas, but in general this characterization does not seem expressive enough to study composition on a wider scale, in its more commonly accepted meaning. With a too simple *structural* complexity, any arrangement is likely to be pictorially *functional*.

With the above in mind, it appears once again that the only feasible option is to restrict inquiry to a corpus that was intentionally and deliberately embedded within a common minimal grammar, such as my own personal work. This dataset is therefore attractive not only because of its size, but also thanks to its cohesive symbolic structure that makes it amenable to quantitative inquiry.

Finally, I understand that using one's own artwork for scientific purposes may be controversial. For an existing example that appears accepted by the wider community, I may refer readers to the work of Christoph Redies, who employs his own creations in the field of empirical aesthetics<sup>7</sup>. In my mind, and in light of the considerations made throughout this chapter, a more important question is whether one's own artworks and artistic practice are coherent with one's modeling intentions. When it comes to my own personal trajectory, I believe a strong case can be made in the affirmative. Beyond that, I can only hope that my sincere, yet personal artistic effort will serve as a good scientific tool for prompting and guiding future research. If there is meaning to the point of contact between creative endeavors and the methods of scientific research, it is especially on how the particular can bring insights to some more universal knowledge.

## 2.1 Floating compositional structures

Before detailing the processing tools developed to transpose my artworks into a formatted dataset ready for modeling, it is important to precisely describe the nature of the drawings.

### *Years of drawing small*

When we think about small drawings erratically spread among notes and draft papers, as it is my case (Fig.2.1), we may naturally regard this practice as doodling. Mostly associated with childish scribbles or a lazy-bored attitude, doodling remained until the surrealism movement an undervalued activity. In the early 20<sup>th</sup> century, with the development of psychoanalysis, people began to appreciate the value of these uncontrolled free-forms for their supposed hidden meaning. Raised to the status of a legitimate artistic practice, doodling is now a popular means of

---

<sup>7</sup>Redies et al., 2015; Schwabe et al., 2018. He claims to address composition with his work, but it appears to me more related to texture and homogeneity perception.

## 2 Personal practice

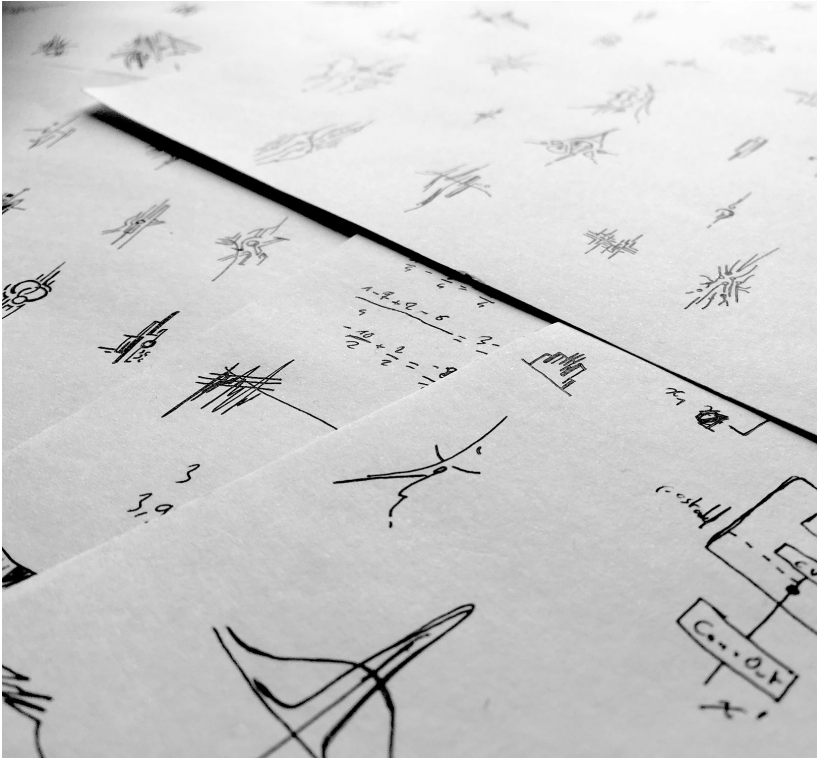


Figure 2.1: Raw drawings among notes.

expression with diverse approaches. The following comments will therefore present my view on this interesting practice.

I would like to emphasize that, even if this activity is not targeted to a predefined result or objective, it should be distinguished from randomness. In my case, only the first stroke may be random or totally free. Then, some kind of detached concentration drives the logic of the next strokes. My main focus of attention may be captured by some other auditory stimuli, but my hand and vision are dedicated to the task. To proceed, this activity requires active and careful evaluation of each line. Concurrently, there is an increasing fear of ruining the generated structure. Any new stroke involves more hesitation than the former. My drawings are small ( $\sim 4\text{cm}$ ) and can happen everywhere; but at the same time, creation moments are now, after 10 years of practice, more relaxing and meditative moments. Then, is it still doodling?

### *Increasing complexity*

As already stated in the introduction, the complexity of the compositional structures increased over the years (see Fig.0.2, Fig.0.3, Fig.0.4). This observation is not supported by any quantitative metric proposed in previous studies<sup>8</sup>, possibly because complexity is not merely a function of the number of strokes or graphical elements present in the structure. Rather, complexity is driven by the quality and the diversity of the plastic interactions between elements. It is connected with the difference, expressed in the previous chapter, between the structural and functional properties of a system. Building a tool for quantifying complexity is central to the present research program, so how to define beforehand if my artistic work fulfills the right amount of complexity? On which basis, for example, may one exclude compositions with an excessive level of element interactions?

Without an objective metric, we can only rely on our own estimate. Subjective judgments are highly variable, but they can be constrained via, for instance, a requirement on viewing duration. Then, to make our analysis feasible, it is advisable to remove compositional structures that exceed our cognitive abilities. For example, we can exclude drawings for which any individual element is not easily comparable to other elements. This occurs in accordance with two main configurations:

1. In preparatory sketches of final art pieces (e.g. Fig.2.2a), compositions are designed on bigger surfaces, letting empty areas and voids play a more important role in grouping autonomous sets of elements and adding a second layer of interactions. Together with scaling differences and distances between individual elements, maintaining a clear mental image of the compositional structure quickly becomes onerous.
2. In Fig.2.2b, strokes are arranged so that the finer grain elements of the composition prove difficult to memorize. The exploration of such texture and nature-inspired patterns induces a feeling of pleasure more related to the *sublime* than to composition. Summarizing Kant:

In front of the mathematical sublime (the celestial vault, the ocean), which impresses by the immensity of its grandeur, as in the face of the dynamic sublime (the storm on the raging ocean), which impresses by the immensity of its power or of its strength, one is not fascinated by a form, but by the formless, by the absence of form, and the spectacle first appears as doing violence to the subject, as infinitely exceeding what the imagination can grasp in a unified way.<sup>9</sup>

---

<sup>8</sup>Redies et al., 2015.

<sup>9</sup>Lories and Lenain, 2002, p. 112: "Devant le sublime mathématique (la voûte céleste, l'océan), qui impressionne par l'immensité de sa grandeur, comme devant le sublime dynamique (la tempête sur l'océan déchaîné), qui impressionne par l'immensité de sa puissance ou de sa force, l'on n'est pas fasciné par une forme, mais par l'informe, par l'absence de forme, et le spectacle apparaît d'abord comme faisant violence au sujet, comme dépassant infiniment ce que l'imagination peut saisir de manière unifiée."

## 2 Personal practice

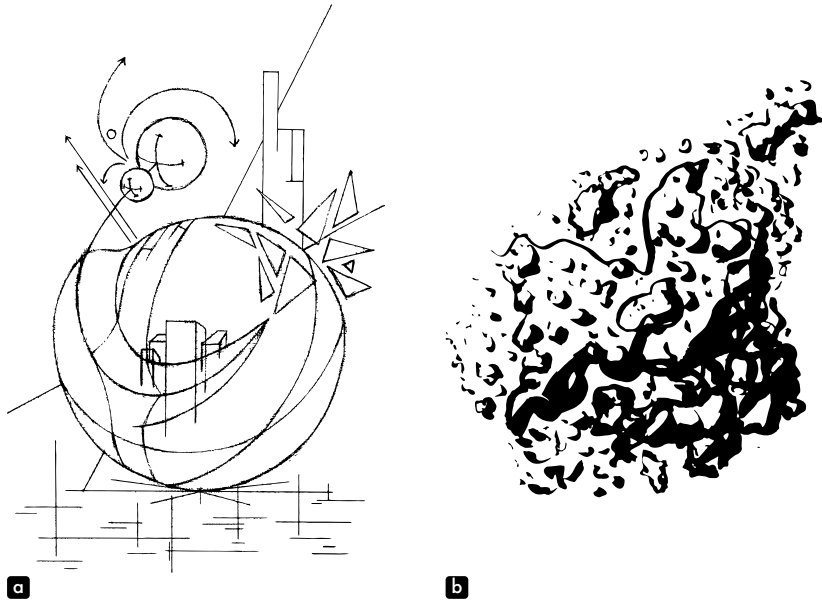


Figure 2.2: Two samples of excluded drawings.

I do not totally agree with this formless nature of the sublime, because sublime typically represents for me the emergence of a form, a new form from globally smaller indistinguishable ones. Sublime occurs exactly in the conflict between known unorganized elements and an unidentified higher-level pattern. At any rate, this perceptual confusion is far beyond the objective of this study.

In practice, the aforementioned rules keep some ambiguous boundaries, but they guarantee a minimal level of complexity for the compositional phenomena we are trying to model.

### *Figurative drawings*

Another question that emerges in the selection of drawings for the dataset relates to the inclusion of figurative drawings (Fig.2.3). We can obviously project upon them the appearance of some animal or vegetation. Nonetheless, if we could prevent our brain from desperately matching familiar patterns, we would notice that their figurative attributes are only roughly suggested. If we pay more attention to these sets of lines, their compositional arrangements are not very different from abstract drawings. I have therefore decided to retain them as part of the dataset.

## 2.1 Floating compositional structures



Figure 2.3: Figurative drawings of vaguely defined creature, human, and tree entities.

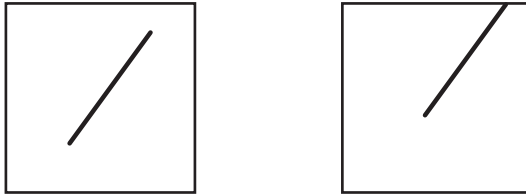


Figure 2.4: Interaction of a line with the B.P. Based on drawings by Kandinsky<sup>11</sup>.

For Klee, such emergence of the figurative is actually a non-event, not mandatory, neither an issue for the composition.

An association of ideas can sooner or later occur in him and nothing will prevent the artist from accepting it anymore if it presents itself under a really appropriate manner. This acquiescence to the object can then inspire one or another addition that a given object, once its nature has been specified, is necessarily calling for. [...] The debate therefore bears less on the presence of the object as such than on its particular mode of existence, on its presentation.<sup>10</sup>

### *Floating structures*

The Basic Plane, B.P., as denominated by Kandinsky, is the material surface hosting the art piece. Even if it is usually blank, the canvas does not passively support strokes. The B.P. introduces forces in the composition by virtue of its boundaries (Fig.2.4).

On approaching the boundary of the B.P., a form increases in tension until, at the moment of contact with the boundary, the tension suddenly ceases.

<sup>10</sup>Klee, 1924/1998, p. 24: "Une association d'idées peut tôt ou tard se produire en lui et rien n'empêchera plus l'artiste de l'accepter si elle se présente sous un nom vraiment approprié. Cet acquiescement à l'objet peut alors inspirer telle ou telle adjonction qu'un objet donné, une fois précisée sa nature, se trouve nécessairement appeler. [...] Le débat porte dès lors moins sur la présence de l'objet comme tel que sur son mode d'existence particulier, sur sa présentation."

<sup>11</sup>Kandinsky, 1926/1991, p. 172

## 2 Personal practice

Furthermore: the farther a form lies from the edge of the B.P., the weaker becomes the attraction of the form to the edge.<sup>12</sup>

The frame and the shape of the frame is therefore a key element of the composition. This point is mainly admitted and shared among artists, nonetheless it is never precisely addressed in models. Only implicit boundaries, like a closed interval  $[0, 1]$ , is usually chosen for every stroke, but this choice is arbitrary. The upper limit is for instance of 109 in a dataset of vectorial kanji<sup>13</sup>. With pixel-based classification models trained on ImageNET<sup>14</sup> like VGG<sup>15</sup>, the B.P. remains an implicit discrete square area of 224px. None of these models actually materializes the boundaries, and how this may be achieved in practice remains an open question. Luckily, my artistic practice bypasses this issue entirely: all compositional structures are drawn without regard for the borders of the canvas, or for surrounding drawings. In this case:

They [the elements] are so loosely knit with the B.P. that the latter's accompaniment is scarcely audible; it disappears, so to speak, and the elements *hover* in space which, however, knows no precise limits (especially in depth).<sup>16</sup>

Then, the central point of the composition becomes the main locus of tension created with the B.P. Even if *top* and *bottom* concepts are still occasionally present, the centroid of all the strokes is where the composition becomes fixed. Lines evolve around the centroid with a concentric quality.

A simple complex of lines can finally be treated in two ways – either it has become one with the B.P. or it lies free in space. The point clawing its way into the plane is also able to free itself from the plane and to *float* in space.<sup>17</sup>

The cosmic form is only created through the suppression of gravity (through elimination of material ties).<sup>18</sup>

In the compositional process, this central point is therefore not firmly imposed from the beginning as for compositions in a frame. The center is slightly rebalanced along the composition, depending on its constituent strokes (see Fig.2.5).

<sup>12</sup>Kandinsky, 1926/1991, pp. 171–172: "Une forme gagne en tension autant qu'elle s'approche des limites du P.O. [Plan Originel], jusqu'au moment où la tension cesse subitement quand cette forme atteint cette même limite. Et autant que cette forme s'éloigne des limites du P.O., autant la tension entre la forme et les limites diminue."

<sup>13</sup>Apel, 2009.

<sup>14</sup>Russakovsky et al., 2015.

<sup>15</sup>Simonyan and Zisserman, 2014.

<sup>16</sup>Kandinsky, 1926/1991, p. 156: "Leur rapport [des éléments] avec le P.O. est si relâché que celui-ci ne résonne pour ainsi dire plus, disparaît presque, et que les éléments « planent » dans un espace sans limites précises (surtout en profondeur)."

<sup>17</sup>Kandinsky, 1926/1991, p. 182: "Une simple composition linéaire peut être traitée de deux façons – ou bien elle est intégrée au plan originel, ou elle flotte librement dans l'espace. Le point, qui s'incruste dans le plan, peut lui aussi, se libérer de la surface et « planer » dans l'espace."

<sup>18</sup>Klee, 1924/1998, p. 126: "La forme cosmique, n'apparaît qu'avec la suppression de la pesanteur. (Avec la disparition des amarres terrestres.)"

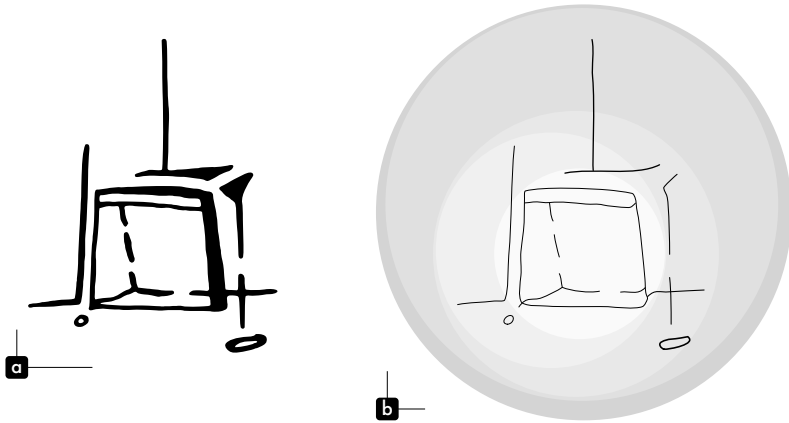


Figure 2.5: Central composition. My graphical structures evolve around a center slightly rebalanced along the compositional process. Shading in panel **b** is illustrating this idea for the original drawing shown in panel **a**.

In this setup, an empty artwork cannot theoretically exist; no tension comes from strokes, nor from the B.P. On the other hand, this initial state is easy to model because it is closer to the mathematical definition of  $\mathbf{0}$ .

## 2.2 Vectorial decomposition

This research program relies on computational tools that only accept numbers as input. In this section, I will therefore detail the process of digitizing and vectorizing my drawings. This transposition is critical because it defines what information will be accessible to the model and how. There is no standard procedure, and each step is tailored to specific modeling objectives and artistic expectations. Because of the large number of elements ( $>5k$ ) and the associated cost of manual work, it is imperative to look beyond the present project and be minimally destructive, so that the database may be used for other purposes in the future. If information must be trimmed, this should happen as late as possible in the data-processing pipeline.

### Digitizing

Some drawing datasets<sup>19</sup> like Google's *Quick, Draw!* have been collected directly on computers with simple drawing interfaces that record sequences of points at

<sup>19</sup>Eitz et al., 2012; Jongejan et al., 2016.



## 2 Personal practice

regular time steps. Users could use their mouse or a stylus. The SUSIG signature dataset<sup>20</sup> has been recorded with pressure-sensitive tablets at high frame-rate. Such tools would be extremely pertinent to study composition in the creative process, as they could provide an absolute (non-ambiguous) and accurate stroke definition. However, this type of interface represents a huge artistic constraint. One may not like this medium as it does not provide nice stroke effects or the right surface feeling. This setup may also be incompatible with the inspirational atmosphere that is required for creative efforts to succeed. At least for me and this dataset, the very small amount of compositional structures realized on a computer or a tablet reflects these uncomfortable constraints.

Moreover, because this collection started 10 years ago, before any specification of the present research project, most of my drawings had already been produced on paper. The first processing action therefore involved a digitizing step. Each drawing was individually digitized with an Epson Perfection V550 Photo scanner at 1200 dpi and with a low sharpening correction. Files were saved in sRGB, 8bit per channel (gray or RGB) tiff with the LZW (Lempel-Ziv-Welch) lossless compression. Fig.2.6a-k show raw scanned drawings at their actual size. This selection is representative of the diversity of ink and paper used throughout the dataset. Paper is mainly solid white, but there are also several ruled-paper examples (Fig.2.6b,c), recycled paper examples printed on the other side (Fig.2.6d), and darker yellowish draft paper (Fig.2.6f). Black ink with roller pens is the norm, but there are also examples of colored ink (Fig.2.6d; if a single colored ink was used throughout the drawing, it was scanned in gray), felt-tip pen (Fig.2.6f), brush pen (Fig.2.6g), pencil (Fig.2.6h,i), and colored pencils (Fig.2.6a). Finally, the few drawings produced digitally (Fig.2.6l) involved various formats (jpg, png, psd) and qualities (compressed, uncompressed).

For the remaining processing steps, we will consider digitized images as linear (converted from sRGB) float (double) arrays of values in the range [0, 1].

### *Inverted Scale drawings*

The first processing step involves standardizing the definition of canvas versus strokes for digitized images. For instance, in opposition with other drawings, Fig.2.6l presents a dark background and white lines. This is the negative of the expected scale of intensities. So, a *contrast inversion* or *color inversion* is required (this step involved the simple operation  $img_{out} = 1 - img_{in}$ ). Fig.2.7b shows the outcome of applying this transformation to Fig.2.7a. Supposing there are more background pixels than stroke pixels, the *inverted* status of an image can be initialized to true if its median value is below 0.5.

---

<sup>20</sup>Kholmatov and Yanikoglu, 2009.

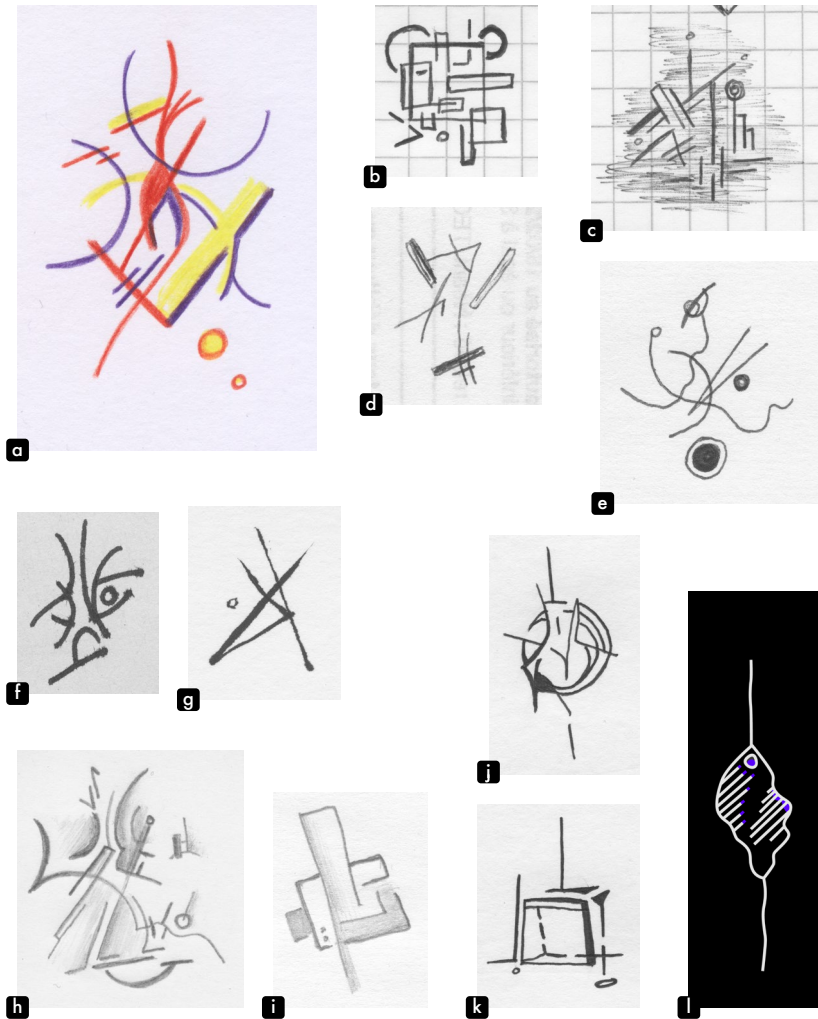


Figure 2.6: Diversity of pen, paper, and support found in the dataset. Scanned drawings are at their actual size, except for panel l which is natively digital.

## 2 Personal practice

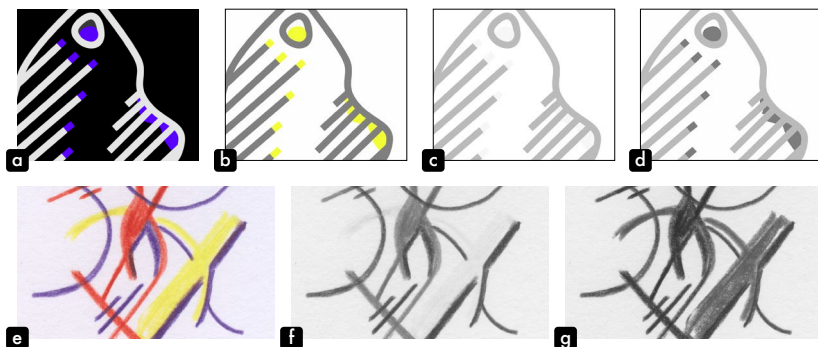


Figure 2.7: Conversion of colored drawings to grayscale images. Panels **a** and **e** show two examples of colored drawings. Panel **b** is the *color-inverted* version of **a**. Panels **c** and **f** are grayscale versions using the standard luminance definition from the CIE 1931. Panels **d** and **g** use  $gray = \min(r, g, b)$  for grayscale conversion.

### Colors

Because the large majority of drawings is monochromatic, removing colors represents a major means of simplification. Color is important to the composition but, as Kandinsky stated:

Form can stand alone [...]. Color cannot [...]; it cannot dispense with boundaries of some kind.<sup>21</sup>

Similarly, Klee<sup>22</sup> gave pictorial elements three nested dimensions, (color-quality (grayscale-weight (line-measure))). It means that a given color is first and foremost a quality, but at the same time it is weight and measure. A line, on the other hand, only involves measure. Thus, simplification of the compositional vocabulary can only be directed towards the line, a binary line. There is no color without an outline and strokes represent minimal compositional words. In addition, if we think about traditional Chinese painting<sup>23</sup>, lines represent ideal tools for creation.

If Chinese painting [...] has privileged ink to the detriment of the colors, it is because ink, on the one hand, by virtue of its internal contrasts, seems sufficiently rich to express the infinite shades of nature and, on the other hand, combining with the art of the line, it offers this unity which [...] solves the contradiction between drawing and color [...]. By virtue of its double quality of being both one and multiple, ink, like the brush, is considered a direct manifestation of the original universe.<sup>24</sup>

<sup>21</sup>Kandinsky, 1912/1989, p. 115: "La forme seule, [...] peut exister indépendamment. La couleur non. La couleur ne se laisse pas étendre sans limite."

<sup>22</sup>Klee, 1924/1998, pp. 19–20.

<sup>23</sup>Cheng, 1989, 2006.

<sup>24</sup>Cheng, 2006, p. 88: "Si la peinture chinoise [...] a privilégié l'Encre au détriment des couleurs c'est parce que l'Encre, d'une part, par ses contrastes internes, semble suffisamment riche pour exprimer

Without invoking such mystical concepts, we may accept that excluding colors from our modeling represents a reasonable choice. Images are usually converted to grayscale by extracting luminance as defined by the  $Y$  component of the CIE 1931. For linear images with sRGB color components, luminance is given by  $Y = 0.2126729 r + 0.7151522 g + 0.0721750 b$ .<sup>25</sup> The result of this operation is shown in Fig.2.7c,f. In both cases, yellow strokes almost disappear in the background, which is problematic. The choice of black and white is purely conventional. The presence or absence of ink, more than the colors of the couple ink/paper, is what matters. Therefore, we are not merely turning images to grayscale, we need to apply a more elementary transformation. We want to convert drawings to binary maps of *being* and *not being*, of the mark of a tool or the absence of a gesture. Remembering our childhood, we want to record the regressive pleasure of seeing something where there was nothing<sup>26</sup>. With this premise, we should extract any information that is far from white, i.e. far from 1 along each component. To do so, we can simply define  $gray = \min(r, g, b)$  (see Fig.2.7d,g).

### Binary maps

In Fig.2.8, close inspection of example drawings from Fig.2.6 highlights the difficulties associated with binarization of the dataset: Fig.2.8b,c present light gray rulers in the background; Fig.2.8c contains shadowing sketches; the lines in Fig.2.8d have some inner white scratches; the paper in Fig.2.8f is darker than other examples and presents ink soaking into its fibers; Fig.2.8g presents involuntary ink dots on the right; Fig.2.8h,i contain very light pencil strokes with grainy shadowing; Fig.2.8j,k use better quality pen and paper, but still display unwanted smudges.

The basic approach to binarization is to set a threshold: values below this threshold are set to 0, those above to 1. Because of the changing intensities of ink and paper, the threshold is initialized as the midpoint between the 5-percentile and the 95-percentile of image values. Fig.2.9a,d,g,j show results from this automatic thresholding procedure. This heuristic generally produces good results, however some manual adjustment is usually required.

To limit ink soaking (Fig.2.9g) and grainy shadowing (Fig.2.9j), a Gaussian filter is applied on the image which is then re-thresholded using a value of 0.5. The default Gaussian kernel standard deviation is 2.0, and is manually adjusted afterwards on a per-drawing basis. Fig.2.9b,e,h,k show example results.

---

les infinies nuances de la nature et, d'autre part, se combinant avec l'art du Trait, elle offre cette unité qui [...] résout la contradiction entre dessin et couleur [...]. Par sa double qualité à la fois une et multiple, l'Encre, comme le Pinceau, est considérée comme une manifestation directe de l'Univers originel."

<sup>25</sup>Lindbloom, 2017.

<sup>26</sup>Arnheim, 1954/2004, p. 171.

## 2 Personal practice

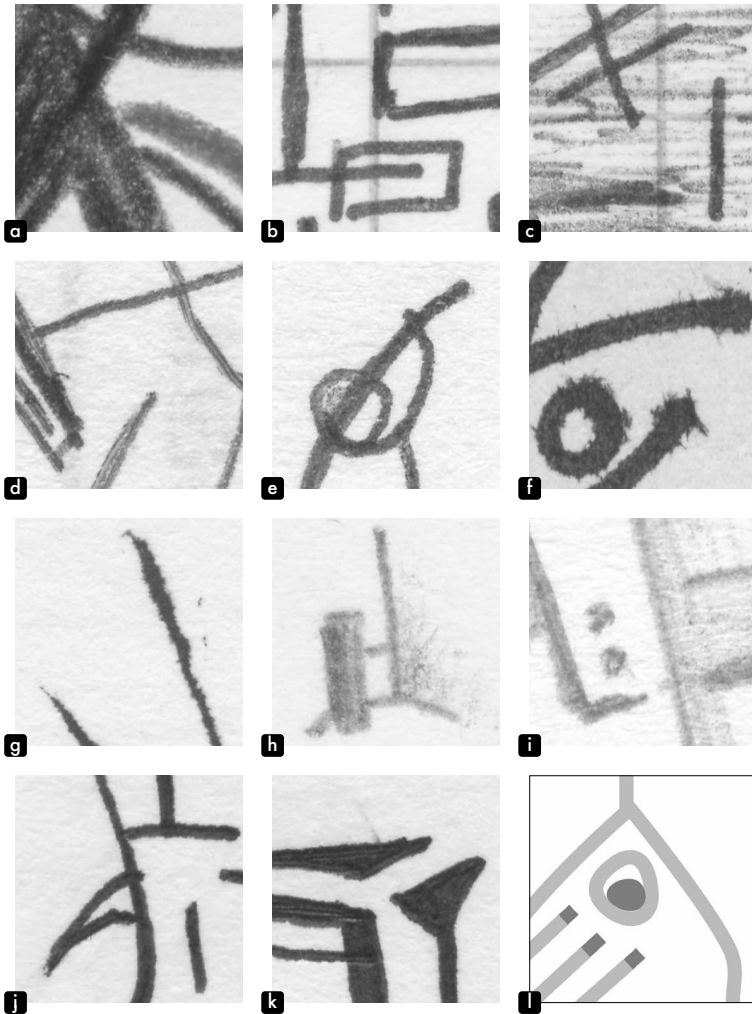


Figure 2.8: Close-up (x5) of scanned drawings. Panel letters correspond to the same compositional structures of Fig.2.6.

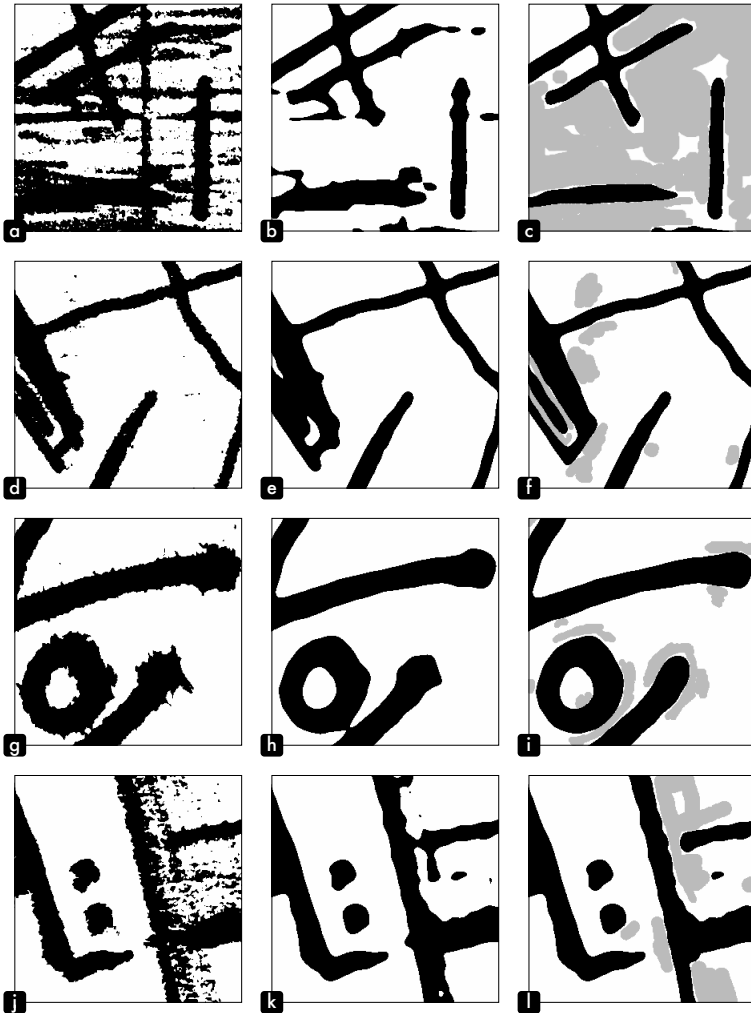


Figure 2.9: Binary map processing. Panels **a**, **d**, **g**, **j** are binarized with automatically determined thresholds. Panels **b**, **e**, **h**, **k** show binary maps smoothed with a Gaussian filter. Panels **c**, **f**, **i**, **l** are corrected with hand-drawn masks (highlighted in gray).

## 2 Personal practice

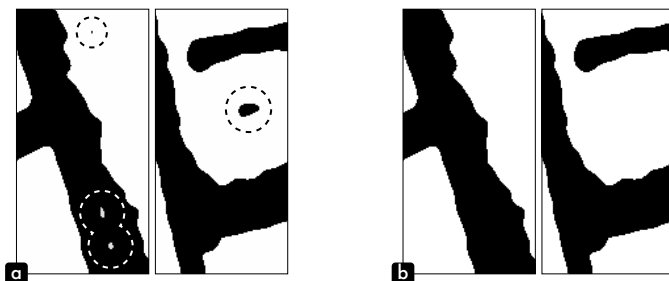


Figure 2.10: Binary maps are automatically cleaned of small black and white unwanted dots/stains (indicated by dotted circles in **a**). Panel **b** shows the final result.

Lines become more distinguishable, but many stains remain (Fig.2.9b,k) and some tangential strokes are merged (Fig.2.9e,h,k). A manual cleaning operation is therefore required. A mask is hand-drawn on unwanted regions. This mask is multiplied with the binary image before the smoothing step. This way the masking operation does not introduce sharper edges. Fig.2.9c,f,i,l present example results (masks are highlighted in gray).

In some cases (Fig.2.10a), black or white dots are still scattered randomly. They usually occupy only a few pixels and are difficult to identify. To get rid of them automatically (Fig.2.10b), we compute all contours on the binary map, i.e. the boundaries between black and white pixels, and then extract the associated hierarchical tree<sup>27</sup>. For instance, contours of white dots at the bottom of Fig.2.10a can be considered as children of the surrounding black area. As a result, it is possible to filter out all contours that do not correspond to leaves in the hierarchical tree and that are bigger than a manually defined area in pixels (default is 4). Furthermore, if we add 1 background pixel around the image before contour extraction, we can assign a value (0 or 1) to the interior of a contour based on its depth within the tree structure (even or odd). We then fill selected spurious surfaces with the computed value to obtain a clean binary map from a digitized drawing. The whole process is summarized in Algorithm.2.1 (Algorithms are grouped at the end of this chapter).

### Surfaces to lines

As we target a model of the composition considering artworks as a sequence of strokes, we need to recover from binary maps the generating trajectories of the forms. We want to extract the skeleton inside surfaces, the centerline of strokes. The application of this idea to Fig.2.11a produces sensible results (Fig.2.11d). If

<sup>27</sup>We used the function *findContours* from OpenCV (<https://opencv.org>). Please check their documentation for further details.

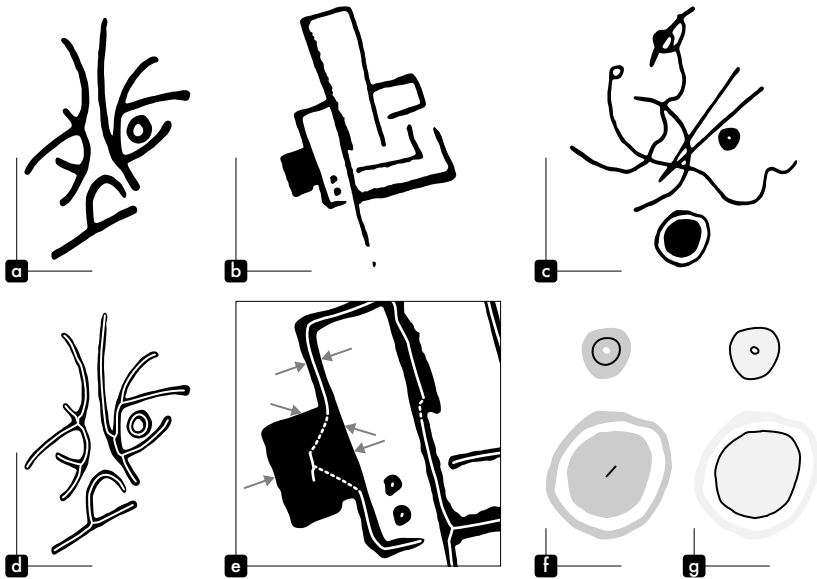


Figure 2.11: Practical issues arise when transposing surfaces to lines. Panels **a**, **b**, **c** are binary maps. Panel **d** shows the nominal case. In panel **e**, extracted skeleton does not reflect original stroke widths. Panels **f** and **g** explore possible line definitions of large dots and circles.

we thicken the skeleton to match the original pen size, we perfectly recover the drawing. But this procedure only works because the different lines have similar width. If we consider Fig.2.11b,e, the extracted skeleton seems lost over wider surfaces. In order to make this approach practicable, we must extract centerline and stroke width at the same time. Strokes then become 3-dimensional objects. They may result from 3-d gestures by the artist. The third axis is virtually like the vertical varying pressure of a calligraphic pen brush<sup>28</sup>. To retain the original number of strokes and trajectories as closely as possible, centerlines should be broken in correspondence with sudden width changes, e.g. by discarding dotted portions of the skeleton in Fig.2.11e.

The extraction, storage and visual representation of stroke width are difficult tasks. Available vectorization algorithms are not designed to record width. Additionally, the main open source vectorial graphics format, SVG (Scalable Vector Graphics), does not propose an implementation of varying stroke width or stroke profile<sup>29</sup>. The most popular commercial solution is Adobe's Illustrator<sup>30</sup>. Its *.ai* files carry a width feature, but their format is proprietary and is not easily modified computationally

<sup>28</sup> *Introduction à la calligraphie chinoise*. 1997; Yee, 1974.

<sup>29</sup> Official documentation at <https://www.w3.org/TR/SVG2/>

<sup>30</sup> More details at <https://www.adobe.com/products/illustrator.html>



## 2 Personal practice

as with SVG (which is XML<sup>31</sup> based). Width data could be stored outside the SVG file, but the complexities and costs associated with implementing this strategy and subsequently visualizing the results pose numerous difficulties. The development of a versatile SVG library to precisely manipulate 2-d graphics programmatically already required a large investment of time<sup>32</sup>. For all these reasons, I have decided to bypass the width issues temporarily and postpone 3-d representation of strokes to future efforts.

The above simplification, although necessary to allow smooth progress of my research program, does introduce compositional losses that must be addressed. A very noticeable loss concerns dots, disks, and rings. In Fig.2.11b, we have for instance two dots that are skeletonized as a short line in their center (Fig.2.11e). This result appears coherent because the size of these elements is obviously on the scale of a dot when compared with the rest of the drawing. But what about the bottom round area of Fig.2.11c? In this case, an outline (Fig.2.11g) would be a better representation of the form and its compositional impact/expressiveness, rather than the short centerline in Fig.2.11f.

The point can grow and cover the entire ground plane unnoticed – then,  
where would the boundary between point and plane be?<sup>33</sup>

The limit between a dot and a disk or, similarly, between a line and a filled square<sup>34</sup>, is already difficult to establish perceptually. Then, how can we define a simple heuristic to discriminate those instances? For example, the thick annular form in the center-right of Fig.2.11c is not better defined by its double outlines as opposed to its centerline (Fig.2.11f,g). We considered manually fixing such configurations, but the ratio cost/modeling improvement seemed to low to further investigate the issue.

Concerning the implementation of the algorithm, skeletonization is a procedure that attempts to iteratively thin surfaces to centerlines by using morphological operations (erosion) together with local connectivity heuristics.<sup>35</sup> Two classical algorithms<sup>36</sup> are already implemented in libraries like scikit-image<sup>37</sup> and are easily accessible. Fig.2.12a,b show the differences between these algorithms. The most recent alternative in Fig.2.12b has been preferred for its clearer outputs. The end of squarish strokes is clean, without small branch-like line segments. The big round area is also skeletonized as a proper dot rather than a messy angular line.

---

<sup>31</sup>XML (Extensible Markup Language): <https://www.w3.org/TR/xml/>

<sup>32</sup>It will be detailed in Chapter.6 when discussing art production from the model.

<sup>33</sup>Kandinsky, 1926/1991, p. 30: "Le point peut grandir, devenir surface et remplir imperceptiblement toute la surface de base – où serait alors la limite entre point et surface ?"

<sup>34</sup>Kandinsky, 1926/1991, p. 108.

<sup>35</sup>There are some exceptions to this scheme. Noris et al., 2013 proposes a gradient-based clustering algorithm on images that do not require binarization, and Feldman and Singh, 2006 addresses the problem with a Bayesian probabilistic approach.

<sup>36</sup>T. Lee et al., 1994; T. Y. Zhang and Suen, 1984.

<sup>37</sup>van der Walt et al., 2014. Find more information at: <https://scikit-image.org>

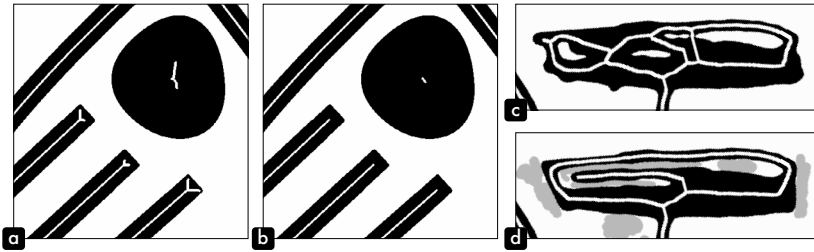


Figure 2.12: Skeletonization. Panels **a** and **b** show the differences between two skeletonization algorithms (**a**: T. Y. Zhang and Suen, 1984, **b**: T. Lee et al., 1994). Panels **c** and **d** depict the importance of manual masking to disentangle messy skeleton structures.

However, the selected algorithm still produces complex structures in ambiguous situations (Fig.2.12c). Fig.2.12d demonstrates the critical role of manual masking (in gray) to unravel intricate graphical elements.

Besides the centerline utilized for modeling, an outline vectorization is useful for illustration purposes and some reproductions on paper. Specific algorithms exist<sup>38</sup>, but we do not require precise control of the level of simplification applied to curves, so we relied on the open source software Potrace<sup>39</sup>, which produces SVG files of sufficient quality.

### ***Skeleton disentanglement***

It is necessary to handle individualization of lines. The skeleton structures introduced so far do not incorporate functional features. In other words, we need to decompose each skeleton into appropriate sub-elements. This processing step is typically difficult at intersections, where joining lines may be converging or overlapping. We first cut all lines between intersections, and then find smart ways to group the right line segments in order to preserve stroke continuity. This issue is central to vectorization of line drawings, and has a long history in computer vision<sup>40</sup>. Each paper proposes its own heuristics and claims the best general results. Depending on the targeted application (font or kanji/hànzì vectorization, design sketches, artistic drawings), there are different trade-offs between fidelity to the original and simplicity of the result (e.g. geometrical correctness in the case of architectural drawings). Furthermore, when tools are publicly available (for instance, it is not the case for the excellent paper from Disney<sup>41</sup> dedicated to animation drawings), they provide an integrated solution, from file to file. Individual processing steps

<sup>38</sup>Hurtut et al., 2011.

<sup>39</sup>Selinger, 2019.

<sup>40</sup>Among others: Favreau et al., 2016; Hilaire and Tombre, 2006; Janssen and Vossepel, 1997; Liao and Huang, 1990; Noris et al., 2013.

<sup>41</sup>Noris et al., 2013.

## 2 Personal practice

are not accessible and adaptation to a particular purpose is difficult. Results from recent approaches<sup>42</sup> are smooth and clean, but I was unable to integrate these algorithms into my processing pipeline with sufficient control. On the other hand, open-source projects like Autotrace<sup>43</sup> are easy to use, but produce unsatisfactory results. This issue has recently been tackled using machine learning. One study<sup>44</sup> proposed an approach to isolate strokes in a drawing based on a learning procedure. The results are promising, but each trained network is dedicated to a stereotyped class of drawings, e.g. *stroked* or *constant-width* kanji, which is problematic for application to a diverse dataset such as mine.

With the goal of precisely controlling each aspect of vectorization, I ended up implementing a simple, yet qualitative algorithm of skeleton disentanglement. Fig.2.13b demonstrates good connection of strokes under different configurations of line style and width variation. I will describe the method using the three example intersections highlighted by gray disks in Fig.2.13a. Labels for these intersections correspond to the close-ups shown in Fig.2.13c,g,k. The results (Fig.2.13f,j,n) are compared with Autotrace (Fig.2.13e,i,m).

We define the neighbor of a pixel as its 8 directly adjacent surrounding units. In Fig.2.13d, we can then disambiguate line-pixels (light gray, 2 neighbors) from one intersection-pixel (dark gray, 3 neighbors). To know which lines to connect, we need to focus on the intersection point and compute orientation vectors for each line-end. We know from the general context (Fig.2.13a) that the two *horizontal* lines should be connected. Autotrace fails here (Fig.2.13e) because it computes these vectors from the local neighbor only (dotted vectors). The angle between *horizontal* lines is smaller than the angle formed by each of them with the *vertical* line. The algorithm then picks one of two equally satisfactory possibilities. Our method works because we compute the orientation vector from a larger distance along the line. This distance is set to an estimate of the mean stroke width used in the drawing. This approach incorporates contextual information and helps overcome the skeletonization artifacts typically found at intersections. Under these conditions, a connection between the two *horizontal* lines is favored (see Fig.2.13f).

Intersection *g*, with two crossing lines (four ends), presents a different challenge to Autotrace. In Fig.2.13h, we notice that there are four intersection-pixels. In Autotrace, this situation is handled by iteratively applying its policy four times. First, it addresses the top intersection pixel and connects the top and right lines because of the local configuration. It is the same inverted Y, and the same incorrect choice as in the previous example. Then, other ends are excluded from possible connections (Fig.2.13i). Moreover, they are locally orthogonal and remain

---

<sup>42</sup>Favreau et al., 2016.

<sup>43</sup>Weber, 2016.

<sup>44</sup>B. Kim et al., 2018.

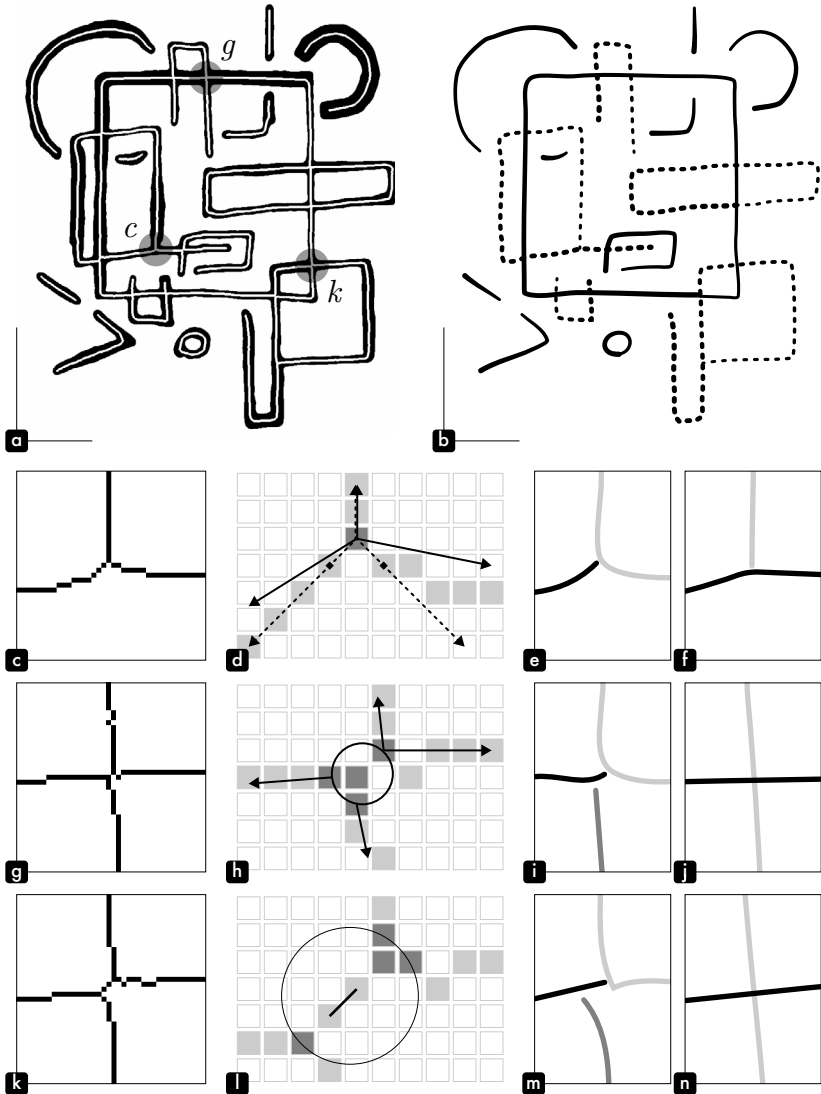


Figure 2.13: Skeleton disentanglement. Panel **a** shows the skeleton extracted from a drawing. Three intersections are highlighted in gray and named after panels **c**, **g** and **k**, being close-ups of these intersections. Panel **b** is the result of our skeleton disentanglement algorithm. Continuous strokes are represented by different line styles and width variations. Close-ups of this result are given by panels **f**, **j**, **n** and compared with Autotrace in **e**, **i**, **m**. Panels **d**, **h**, **l** show explanatory diagrams of our method.

## 2 Personal practice

split. Inspired by previous work<sup>45</sup>, our method clusterizes intersections so that all line-ends belong to the same problem. Fig.2.13j shows how opposing segments are successfully joined.

Concerning case  $k$ , the issue here is that a very short line separates the intersection into two groups (see Fig.2.13l). This artifact often happens with larger stroke width. Four line-ends do not belong to the same problem and cannot be addressed all at once. Unsurprisingly, Autotrace fails (Fig.2.13m), but our method overcomes this issue (Fig.2.13n). We consider all lines between two intersections and smaller than some length (set by default to 75% of the estimated stroke width) as *connection* lines. Intersections are then clustered and resolved in the same manner as described above. If all lines in contact are connected, the small *connection* line is discarded. Otherwise, it can be re-evaluated as a normal line during a second run. Our complete procedure for converting a *binary map to individual lines* is detailed in Algorithm.2.2.

### Parametric curves

Strokes are now defined as sequences of 2-d points separated by  $\sim 1$  pixel. Furthermore, a stroke is by definition a continuous event, i.e. the pen is not lifted from the paper. It implies that these points should be connected to build the curve. So, we need to *repair* the discretization introduced by the digitization of the drawing as an array of pixels. The most simple parametric formulation to join two points is the linear interpolation i.e. a line segment. Let us call these two points  $\mathbf{p}_0$  and  $\mathbf{p}_1$ . The newly created line  $\mathbf{l}$  will be parameterized by  $u$ , so that:

$$\mathbf{l}(u) = (1 - u)\mathbf{p}_0 + u\mathbf{p}_1 = \mathbf{p}_0 + u(\mathbf{p}_1 - \mathbf{p}_0), \quad u \in [0, 1] \quad (2.1)$$

Joining linearly all consecutive pairs of points, the whole curve is called a polygonal line or a polyline. Although accurate, this definition is far from optimal (Fig.2.14a). We are looking for a parametric definition that makes sense for artists. With *pixel-polylines*, the granularity of information is very unlikely to match the rate of conscious actions displayed by artists. Simplification is therefore required. For instance, aligned points could be easily simplified using their extremities without any loss. The idea is then to remove points that lie *almost* on the same line produced by their neighbors. The Ramer-Douglas-Peucker (RDP) algorithm<sup>46</sup> applies this idea recursively. It first draws a new line between the first and the last points. If no middle points exceed a certain user-defined distance to the new line, then all these points can be discarded. Otherwise, if at least one of them is too far from the new line, the furthest point is kept and the procedure is repeated on both sides around this point. Results of the RDP procedure are shown for increasing tolerance distances in Fig.2.14b,c,d. This algorithm has been used in previous

<sup>45</sup>Favreau et al., 2016; Noris et al., 2013.

<sup>46</sup>Douglas and Peucker, 1973; Ramer, 1972.

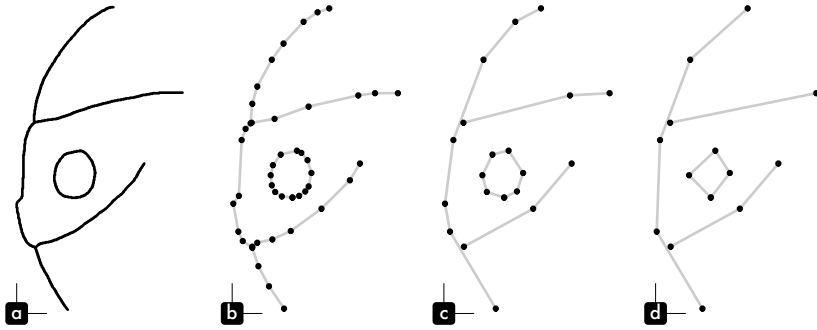


Figure 2.14: Polylines from RDP algorithm. Panel **a** shows initial *pixel-polylines*. Panels **b**, **c**, and **d** show simplified versions generated by the RDP algorithm for increasing tolerance distances (2, 10, 20).

work<sup>47</sup>. It is interesting because it is easy to implement, and because model inputs remain sequences of 2-d points. However, it introduces over-sampling near curved regions (see circle in Fig.2.14b). In addition, it is still quite dissimilar from the definition of stroke that would be provided by an artist.

Among physiologically plausible models of drawing movements<sup>48</sup>, the *Sigma Lognormal* model has proved effective in multiple contexts. It has successfully described signature, handwriting, and simple sketching phenomena<sup>49</sup>. The model provides a parametric definition of curves produced by human rapid movements<sup>50</sup>. It decomposes a stroke as the sum of its neuromuscular actions. A curve  $s$  is then represented by the combination of  $n$  independent components suitable for describing wrist rotations via circular arcs, and muscular accelerations/decelerations via lognormal velocity profiles. Time  $t$  becomes the driving parameter and evolves over an open interval  $[0, t_{end}]$ . With  $s_0$  the initial point of the curve, and  $p_i$  the *shape* parameters of each component:

$$\mathbf{s}(t) = \mathbf{s}_0 + \int_{\tau=0}^t \mathbf{v}(\tau) d\tau = \mathbf{s}_0 + \int_{\tau=0}^t \sum_{i=1}^n \mathbf{v}_i(\tau, p_i) d\tau \quad (2.2)$$

We will now consider only one component and omit subscript  $i$  for legibility.  $\mathbf{v}(t, p)$  can be decomposed into polar coordinates via its norm  $v(t, p) = \|\mathbf{v}(t, p)\|$  and angle  $\phi(t, p)$ , where  $p = [D, t_0, \mu, \sigma, \theta_0, \theta_1]$ :  $D$  (intensity of the neuromuscular action),  $t_0$  (time occurrence of this action),  $\mu, \sigma$  (lognormal distribution parameters),  $\theta_0, \theta_1$  (starting and ending angles of the action arc).

<sup>47</sup>Clanuwat et al., 2018; Ha and Eck, 2017.

<sup>48</sup>Flash and Hogan, 1985.

<sup>49</sup>Berio et al., 2017; Fischer and Plamondon, 2015; Leiva et al., 2015; O'Reilly and Plamondon, 2008.

<sup>50</sup>Plamondon, 1995a, 1995b.

## 2 Personal practice

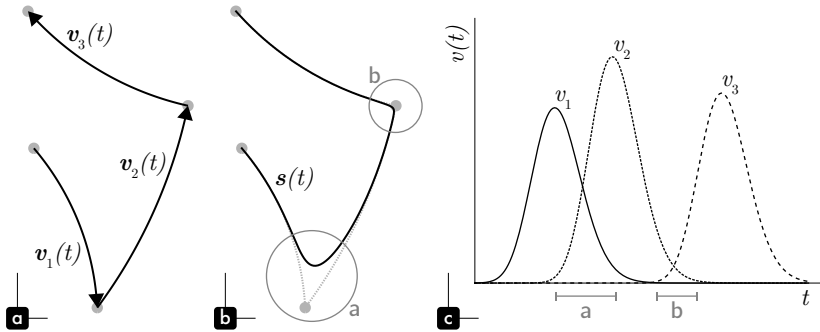


Figure 2.15: *Sigma Lognormal* parameterization (figure based on Berio et al., 2017). A *Sigma Lognormal* curve  $s(t)$  (black line, panel **b**) is a sequence of circular arcs ( $v_1(t)$ ,  $v_2(t)$ ,  $v_3(t)$  in panel **a**). As the velocity profiles of these components usually overlap over time (panel **c**, especially at intervals  $a$  and  $b$ ), the real path of the curve blends smoothly between arcs (highlighted by gray circles in panel **b**).

$$v(t, p) = \frac{D}{\sigma(t - t_0)\sqrt{2\pi}} \exp\left(\frac{-(\ln(t - t_0) - \mu)^2}{2\sigma^2}\right), \quad t > t_0$$

$$\phi(t, p) = \theta_0 + \frac{\theta_1 - \theta_0}{D} \int_{\tau=t_0}^t v(\tau, p) d\tau$$
(2.3)

In short, to rephrase the whole *Sigma Lognormal* parametric definition, consider a stroke as a sequence of circular arcs (e.g.  $v_1$ ,  $v_2$ ,  $v_3$  in Fig.2.15a). These different components do not happen exactly one after the other: their velocity profiles usually overlap over time (see Fig.2.15c, especially at intervals  $a$  and  $b$ ). As a result, the real path of the curve blends smoothly between arcs (Fig.2.15b). Endpoints of components actually only constitute visual targets (represented by gray points in Fig.2.15a,b), where the amount of smoothing is directly driven by the velocity norm ratio of each component along the path (see gray circles in Fig.2.15b corresponding to aforementioned intervals  $a$  and  $b$ ).

At this stage, we must consider whether a physiologically plausible definition of curve decomposition is actually necessary, or even desirable, for the specific application of our research program. For instance, when the true velocity of a stroke is recorded, this model is legitimate enough to enable biometric signature validation by the analysis of neuromuscular constants and other shape parameters<sup>51</sup>. Even for cases in which tracing dynamics is not available (as in our case), an approximate procedure has been proposed<sup>52</sup> that produces plausible data augmentation of handwriting datasets, conserving individual stylistic characteristics. Notwithstanding its benefits, this parametric definition relies on stroke executions faster than those associated

<sup>51</sup>Fischer and Plamondon, 2015.

<sup>52</sup>Berio et al., 2017.

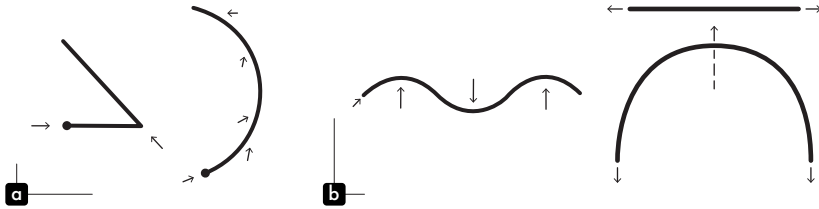


Figure 2.16: Tensions of the line and the curve. Based on drawings by Kandinsky<sup>55</sup>.

with lines in my dataset. The structure of curves in my drawings is primarily dictated by compositional factors and less so by neuromuscular events. In addition, my modeling intentions are more focused on shapes than on their generating dynamics. Stroke widths are discarded and cannot be used as proxy for drawing velocity. Given the above, extraction of relevant physiological parameters from the final shape becomes too expensive. Having said that, the idea of smooth curve interpolation between visual targets is definitely interesting, provided we find simpler motivations/interpretations for these targets and their associated parameters. We can probably get some insights from modern artists who have evoked a vectorial definition of curves.

The external forces which transform the point into a line can be diverse. The variation in lines depends upon the number of these forces and upon their combinations.<sup>53</sup>

For Kandinsky, as a line results from the movement of a point, a natural way to materialize its active forces is by the use of tangential arrows (Fig.2.16a). However, his pedagogical drawings in Fig.2.16b show a conflict with arrows normal to curves. They seem related to a different phenomenon, as forces specifically implied by the curve, like perceptual tensions.

The inner difference [of the curved line] from the straight line consists in the number and kind of tensions: the straight line has two distinct primitive tensions which play an unimportant role in the case of the curved line, whose chief tension resides in the arc.<sup>54</sup>

This notion of tensions housed in arcs could be the right way to decompose complex curves: one tension per component. Mathematically, it could be related to locations of curvature inversion.

<sup>53</sup>Kandinsky, 1926/1991, p. 67: "Les forces extérieures qui transforment le point en ligne peuvent être de nature très différente. La diversité des lignes dépend du nombre de ces forces et de leurs combinaisons."

<sup>54</sup>Kandinsky, 1926/1991, p. 96: "La différence intérieure entre les lignes courbes et droites consiste dans le nombre et la nature des tensions : la ligne droite subit deux tensions primitives définies qui ne jouent qu'un rôle insignifiant pour la ligne courbe – dont la tension essentielle se situe dans l'arc."

<sup>55</sup>Kandinsky, 1926/1991, pp. 96, 97, 103



## 2 Personal practice

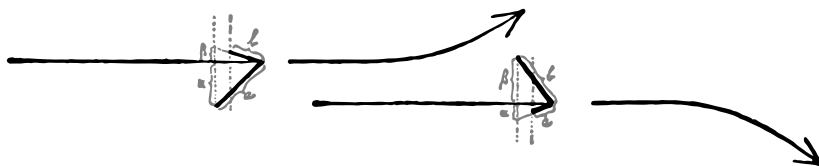


Figure 2.17: Imaginary trajectories associated with arrows for Klee. Uneven lengths  $[a, b]$  and uneven angles  $[\alpha, \beta]$  result in a deviating course<sup>58</sup>.

Klee also considers lines as points in movement, but his vectorial thinking takes even more materiality in arrows, present in many of his final works<sup>56</sup>. This graphical element enters the composition like any others, but possesses a unique expression: “the father of the arrow is the thought.”<sup>57</sup> Arrows embody the physically unachievable will. An arrow head even confers to a line a hypothetical target. Even the subtle arrangement of its head can inflect an imaginary trajectory (Fig.2.17)

What seems important to artists is therefore physical properties of the line reflecting its genesis, as well as hidden intentions, or more pragmatically, the suggestion of inflecting targets. In other words, a curve seems representable by specific landmarks lying on its path, together with tangential inputs guiding the trajectory. This description is not so far from popular Bézier curves, which are precisely popular for the legibility of their parameters. An artist without any mathematical background can easily handle multiple control points to shape a specific design. Their low computational cost and optimized rendering algorithm are also key elements of their success. Bézier curves are implemented in computer typography technologies, in web browser graphics and particularly in the SVG format. For these reasons, Bézier curves represent a reasonable choice for parameterization. Nonetheless, this choice must be appropriately justified. I will try to detail its benefit and suitability for artistic interpretations and modeling purposes.

### Bézier curves

Bézier curves have been independently invented by two french automobile engineers in the late 50s. De Casteljaun has been the first to use it at Citroën, but he did not publish his work due to patent restrictions. At roughly the same time Bézier, who worked at the competing company Renault, shared his work widely and the curves were therefore named after him. Let us begin with their mathematical definition. Given  $n + 1$  control points  $[p_0, p_1, \dots, p_n]$ , a Bézier curve  $c$  of degree  $n$ , parameterized by  $u$ , is defined by:

<sup>56</sup>To cite a few: *Affected Place* (1922), *Wavering Balance* (1922), *Mural from the Temple of Longing* ↙ *Thither* ↘ (1922), *Eros* (1923), *Arrow in the garden* (1929)

<sup>57</sup>Klee, 1924/1998, p. 128: “Père de la flèche est la pensée.”

<sup>58</sup>Klee, 1924/1998, p. 131

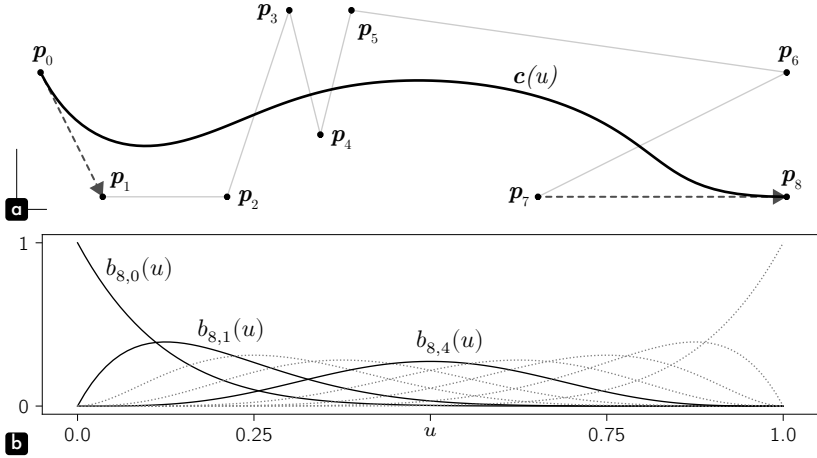


Figure 2.18: Panel **a** shows a Bézier curve  $c(u)$  of degree  $n = 8$  and its 9  $p_i$  control points. The thin gray line is the control polyline. Dotted arrows are tangents at  $c'(0)$  and  $c'(1)$ . Panel **b** shows associated Bernstein polynomials.

$$c(u) = \sum_{i=0}^n b_{n,i}(u) p_i, \quad u \in [0, 1] \quad (2.4)$$

A Bézier curve (e.g. Fig.2.18a) can therefore be considered as a weighted average of its control points by  $b_{n,i}(u)$ , called the basis functions. These coefficients correspond to Bernstein polynomials (see examples in Fig.2.18b), defined as:

$$b_{n,i}(u) = \binom{n}{i} u^i (1-u)^{n-i} = \frac{n!}{i!(n-i)!} u^i (1-u)^{n-i} \quad (2.5)$$

This simple definition carries several interesting properties, such as the behavior of  $c$  at its extremity, i.e. when  $u = 0$  and  $u = 1$ . With the convention of  $0^0 = 1$ :

$$\begin{aligned} b_{n,i}(0) &= \binom{n}{i} 0^i = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise} \end{cases} \implies c(0) = p_0 \\ b_{n,i}(1) &= \binom{n}{i} 0^{n-i} = \begin{cases} 1 & \text{if } i = n \\ 0 & \text{otherwise} \end{cases} \implies c(1) = p_n \end{aligned} \quad (2.6)$$

This result shows that any Bézier curve passes through its first and last control points (see Fig.2.18a). Another important aspect of the desired parameterization relates to tangents and curvature along the path. Let us compute derivatives for  $b_{n,i}$  and  $c$  w.r.t  $u$ .

## 2 Personal practice

$$\begin{aligned}
 b'_{n,i}(u) &= \frac{n!}{i!(n-i)!} i u^{i-1} (1-u)^{n-i} - \frac{n!}{i!(n-i)!} u^i (n-i) (1-u)^{n-i-1} \\
 &= n \frac{(n-1)!}{(i-1)!(n-i)!} u^{i-1} (1-u)^{n-i} - n \frac{(n-1)!}{i!(n-1-i)!} u^i (1-u)^{n-1-i} \\
 &= n(b_{n-1,i-1}(u) - b_{n-1,i}(u))
 \end{aligned} \tag{2.7}$$

and,

$$\begin{aligned}
 \mathbf{c}'(u) &= \sum_{i=0}^n b'_{n,i}(u) \mathbf{p}_i \\
 &= \sum_{i=1}^n b_{n-1,i-1}(u) (n\mathbf{p}_i) - \sum_{i=0}^{n-1} b_{n-1,i}(u) (n\mathbf{p}_i) \\
 &= \sum_{i=0}^{n-1} b_{n-1,i}(u) (n(\mathbf{p}_{i+1} - \mathbf{p}_i))
 \end{aligned} \tag{2.8}$$

By recursion, we can obtain the second derivative of  $\mathbf{c}$ :

$$\begin{aligned}
 b''_{n,i}(u) &= n(n-1)(b_{n-2,i-2}(u) - 2b_{n-2,i-1}(u) + b_{n-2,i}(u)) \\
 \mathbf{c}''(u) &= \sum_{i=0}^{n-2} b_{n-2,i}(u) (n(n-1)(\mathbf{p}_{i+2} - 2\mathbf{p}_{i+1} + \mathbf{p}_i))
 \end{aligned} \tag{2.9}$$

Let us observe once again the behavior of  $\mathbf{c}'$  and  $\mathbf{c}''$  at extremities  $u = 0$  and  $u = 1$ . Reusing results from Eq.2.6:

$$\begin{aligned}
 \mathbf{c}'(0) &= \sum_{i=0}^{n-1} b_{n-1,i}(0) (n(\mathbf{p}_{i+1} - \mathbf{p}_i)) = n(\mathbf{p}_1 - \mathbf{p}_0) \\
 \mathbf{c}'(1) &= \sum_{i=0}^{n-1} b_{n-1,i}(1) (n(\mathbf{p}_{i+1} - \mathbf{p}_i)) = n(\mathbf{p}_n - \mathbf{p}_{n-1}) \\
 \mathbf{c}''(0) &= n(n-1)(\mathbf{p}_2 - 2\mathbf{p}_1 + \mathbf{p}_0) \\
 \mathbf{c}''(1) &= n(n-1)(\mathbf{p}_n - 2\mathbf{p}_{n-1} + \mathbf{p}_{n-2})
 \end{aligned} \tag{2.10}$$

As a result, the vector produced by the first two control points is tangential to the starting point of the curve. The last two points operate similarly with respect to the end point of the curve (see dotted arrows in Fig.2.18a). In addition, it can be shown that other intermediary control points would also have a global impact on the curve, but without a direct interpretation, as a point lying on the curve or a tangent. If we want to maintain an *artistic* meaning for every control point, the upper limit on curve degree is 3. These curves are called cubic Bézier and correspond to 4 control points. Higher degrees present more powerful characteristics and abilities to model complex strokes, but the impact of each control point reduces proportionally to the number of points (see for instance  $\mathbf{p}_4$  in Fig.2.18a which does not inflect the whole curve as we may expect).

Another approach to the approximation of real strokes of arbitrary complexity is to join multiple Bézier curves. The resulting curve is then called composite Bézier curve or polybezier (see Fig.2.19). A condition of  $C^0$  continuity is naturally required, i.e. successive endpoints must be joined, but we can add other constraints to ensure higher degrees of continuity and to produce smooth curves, i.e. without sharp angles<sup>59</sup>. For a minimal example, let us consider two Bézier curves  $c$  and  $d$  (with degree  $m$  and  $q_j$  control points). Enforcing  $C^0$ ,  $C^1$  and  $C^2$  implies:

$$\begin{aligned}
 C^0 &\implies c(1) = d(0) \\
 &\quad \mathbf{q}_0 = \mathbf{p}_n \\
 C^1 &\implies c'(1) = d'(0) \\
 &\quad n(\mathbf{p}_n - \mathbf{p}_{n-1}) = m(\mathbf{q}_1 - \mathbf{q}_0) \\
 &\quad \mathbf{q}_1 = \mathbf{p}_n + \frac{n}{m}(\mathbf{p}_n - \mathbf{p}_{n-1}) \\
 C^2 &\implies c''(1) = d''(0) \\
 &\quad n(n-1)(\mathbf{p}_n - 2\mathbf{p}_{n-1} + \mathbf{p}_{n-2}) = m(m-1)(\mathbf{q}_2 - 2\mathbf{q}_1 + \mathbf{q}_0) \\
 &\quad \mathbf{q}_2 = \mathbf{p}_n + \frac{2n}{m}(\mathbf{p}_n - \mathbf{p}_{n-1}) + \frac{n(n-1)}{m(m-1)}(\mathbf{p}_n - 2\mathbf{p}_{n-1} + \mathbf{p}_{n-2})
 \end{aligned} \tag{2.11}$$

Therefore, if  $c$  is fixed,  $\mathbf{q}_0$ ,  $\mathbf{q}_1$  and  $\mathbf{q}_2$  do not have any degree of freedom. Their positions are completely constrained by  $\mathbf{p}_{n-2}$ ,  $\mathbf{p}_{n-1}$ ,  $\mathbf{p}_n$  and curve degrees  $n$ ,  $m$ . Intuitively, tangent orientation and direction should be sufficient to ensure first-order continuity between  $c$  and  $d$ , but here  $\mathbf{q}_1$  has to be placed at a specific distance. So, to better fit our intuitive idea of continuity, the concept of geometric continuity  $G$  has been introduced. The mathematical ground is that the “traditional measure of continuity [ $C$ ] is affected by reparameterization.”<sup>60</sup> For this reason, the introduction of free parameters  $\beta$  partially release the constraints of  $C$  continuity.

$$\begin{aligned}
 G^0 &\implies \mathbf{q}_0 = \mathbf{p}_n \\
 G^1 &\implies \mathbf{q}_1 = (1 + \beta_1)\mathbf{p}_n - \beta_1\mathbf{p}_{n-1} \\
 &\quad = \mathbf{p}_n + \beta_1(\mathbf{p}_n - \mathbf{p}_{n-1}) \\
 G^2 &\implies \mathbf{q}_2 = (\beta_1^2 + 2\beta_1 + 1 + 0.5\beta_2)\mathbf{p}_n - (2\beta_1^2 + 2\beta_1 + 0.5\beta_2)\mathbf{p}_{n-1} + \beta_1^2\mathbf{p}_{n-2}
 \end{aligned} \tag{2.12}$$

To retain  $G^1$ ,  $\mathbf{q}_1$  can be at any positive distance  $\beta_1 \|\mathbf{p}_n - \mathbf{p}_{n-1}\|$  from  $\mathbf{p}_n$  on the tangent line defined by  $\mathbf{p}_n - \mathbf{p}_{n-1}$ . So, while fitting the curve  $d$  to a *pixel-line*,  $\beta_1$  can be adjusted to any positive real. Optimization of  $\beta_2$  would behave similarly. Fig.2.19 illustrates  $G^0$ ,  $G^1$  and  $G^2$  continuity with two cubic Bézier curves.

Even if we do not consider strokes of the dataset as rapid movements (in the *Sigma Lognormal* definition), a stroke is, in principle, the result of an uninterrupted gesture.

<sup>59</sup> $C$  stands for mathematical parametric continuity, sometimes called the *smoothness* of a function.

<sup>60</sup>Fournier and Barsky, 1985a, 1985b.

## 2 Personal practice

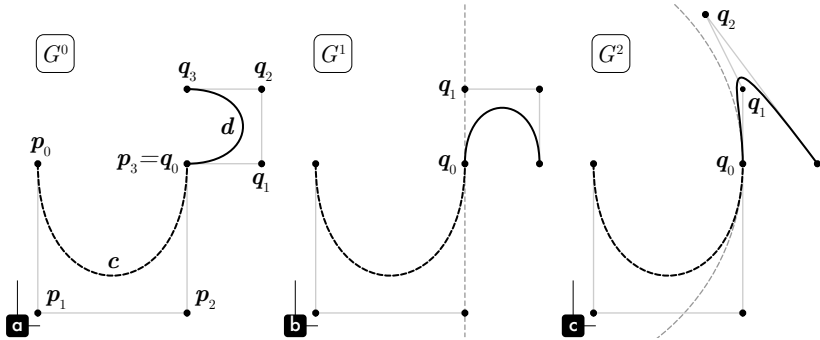


Figure 2.19: Illustration of geometric continuity with two cubic Bézier curves  $c$  ( $p_i$  control points, dotted black line) and  $d$  ( $q_i$  control points, black line). In panel **a**,  $G^0$  is not violated because  $p_3 = q_0$ . Panel **b** shows  $G^1$  continuity by enforcing  $q_1$  to lie on the tangent line passing by  $q_0$  (gray dotted line,  $\beta_1 = 0.5$ ).  $G^2$  is depicted in panel **c**, where  $q_1$  and  $q_2$  are computed from Eq.2.12 ( $\beta_1 = 0.5$ ,  $\beta_2 = 1.0$ ). As a result, an arc can be drawn to follow the curvature of the composite Bézier curve on each side of  $q_0$  (gray dotted line).

Therefore, the presence of a sharp angle along a curve (breaking  $G^1$ ) would imply two different strokes (Fig.2.19a). Nonetheless, a quick change of direction can still satisfy the  $G^1$  constraint. Tangents at a point can take different magnitudes (e.g. a very small  $\beta_1$ ) and remain aligned, allowing local, yet smooth, inflection. On the other hand, breaking  $G^2$  would only provoke unbalanced curvatures at junctions. In other words, the center of the circle that would fit inside the curve would be different on each side. Conforming to  $G^2$  could enforce the natural aspect of true dynamic strokes (see Fig.2.19c), but this requirement is too restrictive for our dataset.  $G^1$  seems intuitively sufficient for good visual continuity (Fig.2.19b).

However, is this choice practical? For modeling purposes, the degree of Bézier curves must remain constant for every component of a composite curve and for every stroke of the dataset. Therefore, what degree corresponds to a feasible fitting scenario? Let us consider three consecutive Bézier curves  $c$  ( $p_i$  control points),  $d$  ( $q_i$  control points) and  $e$  ( $r_i$  control points) of degree  $n$ . The fitting procedure will focus on  $d$  only. Adjacent  $c$  and  $e$  will be considered as already fitted and fixed. If  $n = 1$ , a Bézier curve is equivalent to a linear interpolation (Eq.2.1) and possesses only two control points. In this condition:

$$G^0 \implies q_0 = p_1 \quad q_1 = r_0 \quad (2.13)$$

Linear Bézier curves are already problematic for a fitting procedure with  $G^0$  continuity, because there is no degree of freedom. If the line segment  $q$  is not a good approximation of its curve portion, there is no alternative to subdividing the problem. This configuration is actually similar to the RDP algorithm described earlier. Then, let us look at quadratic Bézier curves with  $n = 2$  (3 control points):

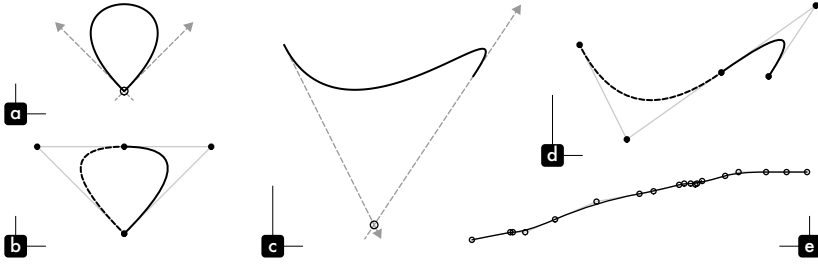


Figure 2.20: Fitting difficulties associated with quadratic Bézier curves. Panels **a** and **c** show two scenarios for which extremity tangents (gray dotted vectors) join at an impossible location for an approximation by one quadratic Bézier curve (black empty circles). In **a**,  $\beta_1 = \beta'_1 = 0$  and in **c**,  $\beta'_1 < 0$ . At least one subdivision is required, and panels **b**, **d** present the resulting fit by two quadratic Bézier curves (dotted and solid dark lines). Panel **e** demonstrates this issue in a real-world case. Even if the line to fit is simple, its alternating concave and convex parts produce a large number of components.

$$\begin{aligned}
 G^0 &\Rightarrow \mathbf{q}_0 = \mathbf{p}_2 & \mathbf{q}_2 &= \mathbf{r}_0 \\
 G^1 &\Rightarrow \mathbf{q}_1 = \mathbf{p}_2 + \beta_1(\mathbf{p}_2 - \mathbf{p}_1) & \mathbf{q}_1 &= \mathbf{r}_0 + \beta'_1(\mathbf{r}_0 - \mathbf{r}_1) \\
 \beta_1 &= \frac{\det(\mathbf{r}_0 - \mathbf{p}_2, \mathbf{r}_0 - \mathbf{r}_1)}{\det(\mathbf{p}_2 - \mathbf{p}_1, \mathbf{r}_0 - \mathbf{r}_1)} & \beta'_1 &= \frac{\det(\mathbf{p}_2 - \mathbf{p}_1, \mathbf{p}_2 - \mathbf{r}_0)}{\det(\mathbf{p}_2 - \mathbf{p}_1, \mathbf{r}_0 - \mathbf{r}_1)}
 \end{aligned} \tag{2.14}$$

To satisfy the  $G^1$  requirement, we notice that  $\mathbf{q}_1$  becomes the intersection of the tangents from both sides. As a result,  $\beta_1$  and  $\beta'_1$  are deterministically computed. When specified this way, **d** may represent its curve portion poorly, and we face the same problem described earlier: we require a subdivision. However, this is a manageable problem. Tangents at extremities can converge onto impossible locations (i.e.  $\beta_1 \leq 0$  or  $\beta'_1 \leq 0$ ), also requiring at least one subdivision (see examples of Fig.2.20a,b,c,d). Depending on the configuration (e.g. a line with many alternating concave and convex parts), this issue can occur recursively and may artificially increase the number of components for a given curve to fit (Fig.2.20e). Ultimately, when **d** is required to fit a region between two adjacent pixels, we have no choice but to violate the  $G^1$  constraint and fit linearly.

To ease interpretation of control points, we have fixed the upper limit to cubic ( $n = 3$ ) Bézier curves and their 4 control points.

$$\begin{aligned}
 G^0 &\Rightarrow \mathbf{q}_0 = \mathbf{p}_3 & \mathbf{q}_3 &= \mathbf{r}_0 \\
 G^1 &\Rightarrow \mathbf{q}_1 = \mathbf{p}_3 + \beta_1(\mathbf{p}_3 - \mathbf{p}_2) & \mathbf{q}_2 &= \mathbf{r}_0 + \beta'_1(\mathbf{r}_0 - \mathbf{r}_1) \\
 G^2 &\Rightarrow \mathbf{q}_1 = \mathbf{r}_0 + 2\beta'_1(\mathbf{r}_0 - \mathbf{r}_1) + \beta_1^2(\mathbf{r}_0 - 2\mathbf{r}_1 + \mathbf{r}_2) + 0.5\beta_2'(\mathbf{r}_0 - \mathbf{r}_1) \\
 & \mathbf{q}_2 = \mathbf{p}_3 + 2\beta_1(\mathbf{p}_3 - \mathbf{p}_2) + \beta_1^2(\mathbf{p}_3 - 2\mathbf{p}_2 + \mathbf{p}_1) + 0.5\beta_2(\mathbf{p}_3 - \mathbf{p}_2)
 \end{aligned} \tag{2.15}$$

As expected, cubic Bézier curves easily conform to  $G^1$  continuity. Extremity tangents do not have to meet, so the fitting procedure can adjust  $\beta_1$  and  $\beta'_1$

## 2 Personal practice

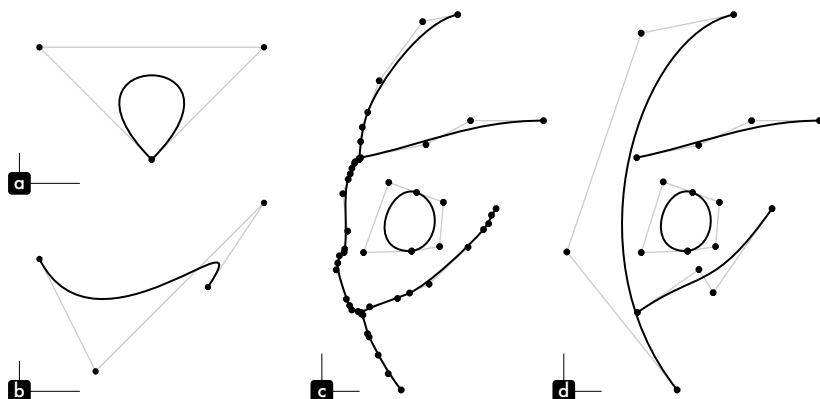


Figure 2.21: Panels **a** and **b** show how cubic Bézier curves can successfully handle the problematic scenario depicted in Fig.2.20a-d i.e. only one component is required. Panels **c** and **d** show fit and simplification of the polyline from Fig.2.14a by cubic Bézier curves with tolerance values of 1.0 and 4.0. Notice that the result of panel **d** remains faithful to the original despite using only a few control points (5 cubic Bézier curves components only).

independently to best match the stroke portion of interest. Problematic configurations for quadratic curves are well handled using cubic Bézier (see Fig.2.21a,b). Even if a solution is not provided here, notice that  $\beta$  parameters would have to be deterministically computed if  $G^2$  were required, leading to the same issue depicted with quadratic curves for  $G^1$ . Composite cubic Bézier curves therefore represent the most reasonable choice for a parametric definition of strokes. This decision is also supported theoretically by the fact that higher degree curves require fewer data than lower ones. Even if each component has more parameters (control points), it should be compensated by fewer components for the same visual result. The only drawback of using Bézier curves is their inability to perfectly reproduce circles and arcs: these can only be approximated. However, perfect circles are unlikely to be present within my hand-drawn dataset.

### Fitting cubic Bézier curves

The last step in the vectorial decomposition of my drawings involves actual fitting of strokes with cubic Bézier curves,  $G^1$  continuity, and some chosen accuracy level. Two different levels of fitting accuracy are required. A first fitting procedure should maintain maximum fidelity to the original curves, while ironing out processing artifacts (mostly at intersections). These vectorial compositions are suitable for art production (e.g. Fig.2.21c with a 1.0 tolerance), but the modeling dataset needs to target fewer components per stroke. This can be achieved by increasing fitting tolerance (e.g. Fig.2.21d with a 4.0 tolerance).

Our vectorization procedure is mostly a customized version of an algorithm published in Graphics Gems<sup>61</sup> in the 90s. Implementation details can be found in Algorithm.2.3. The basic idea is very close to the RDP algorithm. A fit of the whole curve is first attempted and, if the approximation error exceeds some tolerance value, the fitting procedure is recursively applied on subdivisions. The main difference is that adjacent tangents of the curve portion of interest are required to satisfy  $G^1$  continuity. In Eq.2.15, these tangents are specified by the control points of neighboring curves, but they are unavailable during the first fitting attempt on the whole curve. This is also the case after subdivision, because adjacent Bézier curves may not be already fitted. Nevertheless, these tangents can be approximated from the sequence of 2-d points we are trying to fit. We can compare the position of an endpoint with another point at some distance along the path, and estimate the tangent. Once these tangents are specified, the fitting procedure must only search for optimal  $\beta_1$  and  $\beta'_1$  values. Our modifications of the original algorithm provide better qualitative control. First, different parameters define the way tangents are approximated. Then, for closed curves (loops), we have added a procedure to improve  $G^1$  at endpoints. Finally, the distance error between original points and fitted curve is not computed as maximum error along the path, but as average of all errors. In this manner, the longer the curve, the more subtle vibrations are smoothed out by a fixed tolerance level. The underlying idea is to retain the dynamic aspects of a gesture, the essential shape of a stroke.

## 2.3 Dataset formatting

The processing steps described in the previous section must be applied individually to each one of the >5k drawings with some manual intervention via custom interactive software (which we wrote for this specific purpose). Furthermore, to be ready for modeling, a dataset needs to be adequately standardized e.g. different compositions must be uniformly centered and scaled. We also limited quantitative discrepancies between individual elements i.e. by constraining the number of strokes per composition, as well as the number of cubic components per stroke. All these different aspects are described in this section.

### Manual work

A custom app (see Fig.2.22) has been designed and developed in Python (interface with PyQt<sup>62</sup>). It gives the ability to adjust the variables described in Algorithm.2.1 (*inverted*, if a drawing has an inverted scale; *threshold*, the maximum pixel

---

<sup>61</sup>Schneider, 1990.

<sup>62</sup>PyQt is a Python binding of Qt application framework. See details at: <https://riverbankcomputing.com/software/pyqt/intro>



## 2 Personal practice

intensity of stroke-pixel; *kernelSize*, the standard deviation of the Gaussian kernel that smooths contours; *minArea*, the minimum area of black or white surfaces to keep). The user can also label each drawing as *figurative*, *element* (first class composition, *non-element* are still included in the dataset, but are subjectively judged as less *accomplished*) or *excluded*<sup>63</sup>. Finally, the main purpose of this app is to interactively draw cleaning masks. They are overlaid on original images (left side), and another panel (right side) helps the user to check the resulting cleaned binary map, as well as its skeleton and three framing guides (2 circles and 1 rectangle, which will be discussed in the next subsection).

All parameters and masks are stored separately from original files, so that each action is non-destructive. To be functional, each drawing needs to be uniquely identified with a naming convention. This identifier will also be useful for artistic reproduction purposes, as a specific set of drawings can be selected. The chosen convention is *yy-mm-ppp-ddd* with *yy*:year, *mm*:month, *ppp*:page index, *ddd*:drawing index. When date information is not available, 00 is used. For example, the drawing shown in Fig.2.22 is assigned the identifier 1600-006-10: this means that it was drawn in 2016, but the precise month is unknown.

In the masking operation, some inputs require not only local cleaning of strokes, but also simplification actions. For instance, I could decide to simplify the global composition or the structural skeleton. When intricate lines must be disambiguated, these choices carry an artistic value. Despite the aid of the app, manual editing of the entire dataset required  $\sim$  one month. The processing algorithms also required refinement as new model requirements came along, thus involving iterative manual labor. Every minor change on  $>5k$  elements took days to implement.

### *Spatial standardization*

As shown in Fig.2.6, drawings have been produced at very different sizes. In addition, during the scanning operation, they were framed unevenly to capture the whole drawing. No centering was intended. For modeling purposes, it is necessary to center and scale each drawing to standardize the dataset. Then, different policy can be imagined for very different purposes. For instance, if we search for the minimal enclosing rectangle (dotted rectangle in Fig.2.23c), computed extents are optimal to fit the drawing in rectangular frame of any ratio. On the other hand, if we want to fit multiple compositions within a grid (regular or honeycomb, see Fig.6.12), the computation of the minimum enclosing circle is preferable (dotted circle in Fig.2.23c). Although enforcing the filling of a surface is interesting for artwork reproduction, this is not optimal for composition modeling. Imagine the drawing in Fig.2.23c without its top and bottom vertical lines. The smallest

---

<sup>63</sup>Concerning drawing labels, in order to guarantee coherent judgments over the entire dataset (especially for *element/non-element*), a second app has been developed. It offers an overview of multiple drawings of a specific category, and enables individual/batch exclusion of misclassified compositions.

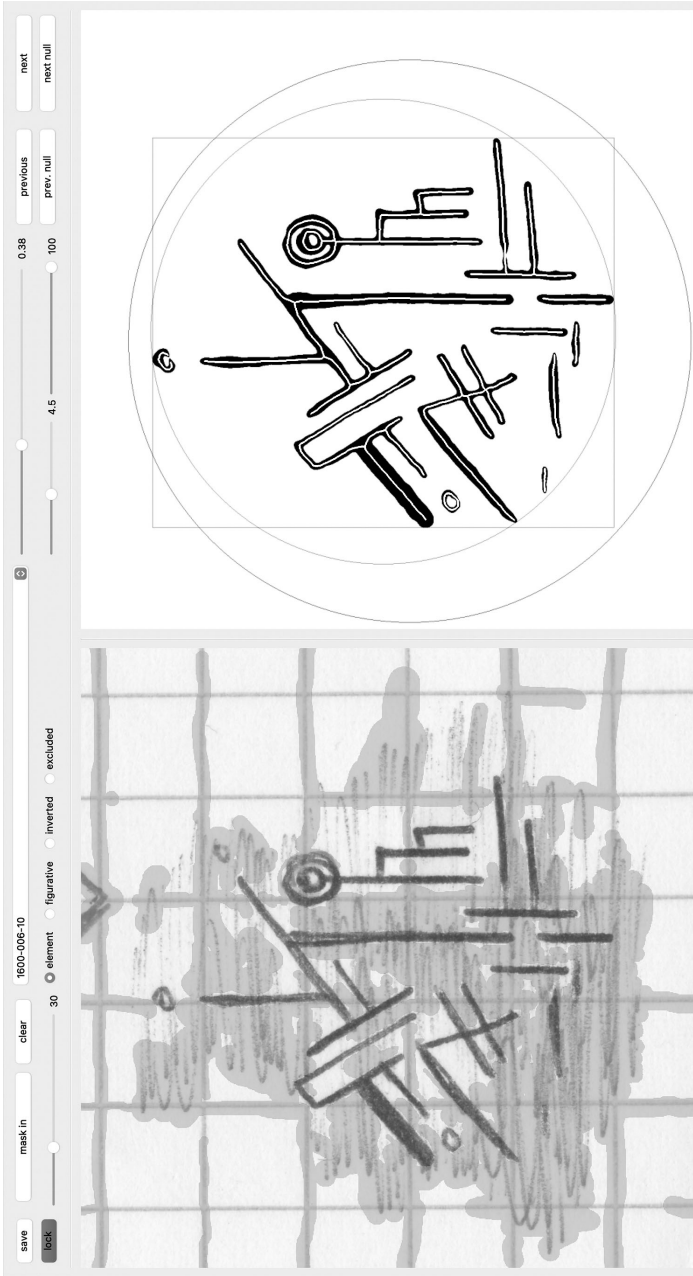


Figure 2.22: Dataset application for manual editing. The top panel lets the user adjust processing and database variables. On the left, a cleaning mask can be edited. It is overlaid on the original drawing. On the right side, resulting cleaned binary map, its skeleton and three framing guides can be checked interactively.

## 2 Personal practice

enclosing circle would nearly double the size of the remaining central part. So, small changes in compositions would produce dramatic encoding effects at the dataset level. Furthermore, as we have already explained in the first section of this chapter, my compositional structures evolve around a center in an open field (under cosmic gravitational rules, in Klee's words). We are therefore looking for some statistical distribution of pictorial masses, rather than absolute boundaries.

The concept of *balance*, through a physical interpretation of visual element weights, has long been discussed<sup>64</sup> and the *Center of Mass* theory has not been proved in the general case. However, this approach is at least well correlated with human judgments of balance for simple binary images. Without better insights into the perceptual importance of real strokes and their mutual interactions (which we are actually trying to model), we will therefore consider stroke-pixels of a drawing to have equal mass, and then attempt to fit this two-dimensional density to a 2-D standard normal distribution.

To do so, we estimate first- and second-order moments, i.e. the mean and covariance of the pixel coordinates  $\mathbf{p}_i = [p_{x,i}, p_{y,i}]$  associated with strokes<sup>65</sup>. With  $n$  being the number of stroke pixels, the mean/center is simply computed as  $\boldsymbol{\mu} = \mathbb{E}_n[\mathbf{p}_i]$ . Concerning covariance, we initially assume the  $x$  and  $y$  components to be independent. In addition, we want to keep the scaling of these dimensions homogeneous. The covariance is therefore of the form  $\text{Cov}_n[\mathbf{p}_i] = \sigma^2 \mathbf{I}$ , implying:

$$\sigma^2 = \frac{1}{2} \sum_{k \in [x,y]} \mathbb{E}_n[(p_{k,i} - \mu_{k,i})^2] = \frac{1}{2} \mathbb{E}_n \left[ \sum_{k \in [x,y]} (p_{k,i} - \mu_{k,i})^2 \right] = \frac{1}{2} \mathbb{E}_n[d_i^2] \quad (2.16)$$

with  $d_i$  being radial distance from the center<sup>66</sup>. Once stroke pixel density is standardized as described, the radial distances  $d_i$  are supposed to follow a  $\chi_2$  distribution. The *rescaled* density of the drawing in Fig.2.23c is presented in Fig.2.23a. We notice that, individually, a composition can be quite far from the theoretical  $\chi_2$  distribution. However, at the dataset level (Fig.2.23b), our procedure produces a distribution of radial distances coherent with the theoretical distribution.

Finally, we would like all drawings to be inscribed within a circle of unit radius. The  $\chi_2$  distribution is defined for any positive real, so in order to set an absolute bound we have chosen that the unit radius must contain 99% of the distribution. This unit circle is materialized by a black line in Fig.2.23c. Some parts of the vertical stroke exceed the unit circle, but it is expected. They represent marginal pixels

<sup>64</sup>Among others: Arnheim, 1954/2004; Hübner and Fillinger, 2016; McManus et al., 2011; Ross, 1907; Wilson and Chatterjee, 2005.

<sup>65</sup>In the MNIST Database (LeCun et al., 1998), the *Center of Mass* has also been used to position digits in the center of their 28px square frame. However, the scaling was operated with a minimum enclosure strategy.

<sup>66</sup>This is actually equivalent to fitting the mean of the squared radial distances  $d_i^2$  to the mean of a  $\chi_2^2$  distribution equal to 2.

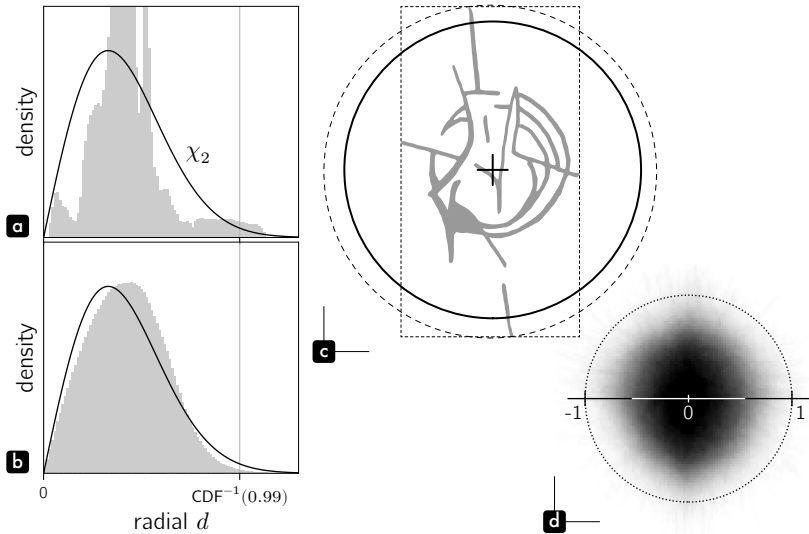


Figure 2.23: Panels **a** and **b** show the standardized density of the radial distance of stroke-pixels (**a**, of the drawing from panel **c**; **b**, of the whole dataset). In panel **c**, a composition is displayed with its smallest enclosing rectangle, its smallest enclosing circle (dotted line) and its statistical unit circle (solid line). Panel **d** is the *rescaled* 2-d density of the whole dataset.

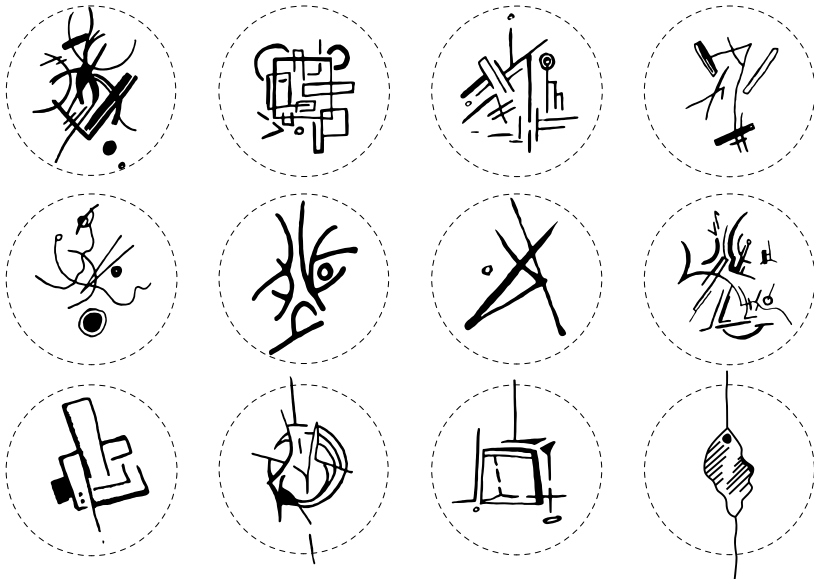


Figure 2.24: Spatial standardization. Compositions are fitted to a statistical unit circle.

## 2 Personal practice

compared to the compact center. Fig.2.23d shows the *rescaled* 2-d density of stroke pixels for the whole dataset, which appears homogeneous in all directions. Sample results from this procedure are shown in Fig.2.24. To summarize, the scaling radius  $r$  of a drawing is computed as:

$$r = \text{CDF}_{\chi_2}^{-1}(0.99) \sqrt{\frac{1}{2} \mathbb{E}_n [d_i^2]} \quad (2.17)$$

### Two distinct datasets

Previous approaches to simple sketches and kanji<sup>67</sup> have considered their inputs as a continuous sequence of positions. The pen could possibly be raised between strokes to give an impression of discontinuity, but the overall sequence of lines was fixed. In Subsection.1.2.Temporal complexity, we have expressed the requirement to model the natural ability to travel within a composition in different orders. Please refer to this subsection for details on the temporal duality of the composition. While the ability to shuffle efficiently the order of strokes within a composition is crucial, it is very unlikely that an artist would draw a line randomly by part. Strokes and compositions are therefore intrinsically of a different temporal nature, respectively continuous and discontinuous. With a unique sequence of points, it is still possible to shuffle this sequence by part, keeping stroke integrity, but it would not respect the essential temporal difference. It would produce semantic discrepancies and uneven information granularities in the model inputs. Strokes represent our unit of artistic intention, and at the same time an acceptable minimal definition of graphical element. Our idea is to then decompose our modeling approach into a stroke model and a composition model.

As a result, we will now consider a composition as an unordered arrangement of strokes, and a stroke as an absolute sequence of ordered points. From a more practical standpoint, it means that each stroke will start from the origin (0, 0) and its initial point  $p_{0,0}$  will be stored at the composition dataset level. This way, we will be able to address each temporal specificity independently for each model. Finally, as strokes can be initiated from both extremities, we must introduce the additional convention that strokes are centrifugal, i.e. the initial point will always be the one closer to the center.

### Stroke simplification

In this subsection, we examine the dataset dedicated to strokes and study its statistical characteristics. Our main concern relates length, i.e. the number of cubic Bézier components per complete curve. If we collect strokes from all compositions,

---

<sup>67</sup>Clanuwat et al., 2018; Ha and Eck, 2017.

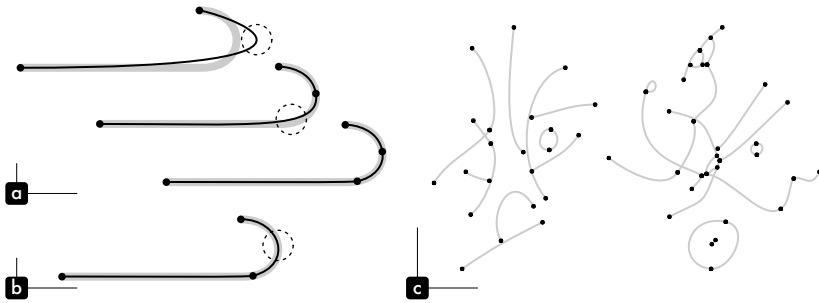


Figure 2.25: Stroke simplification. Panel **a** illustrates how the curve fitting algorithm adds components to achieve some target error. It places intermediary control points on the curve where the error was maximal at the previous optimization step (highlighted by dotted circles). Panel **b** depicts how our simplification algorithm can merge adjacent components, while keeping the same visual shape. Results of similar procedure on real compositions is shown in panel **c**. Note that only the extremity control points of components (2 on 4) are materialized by small black dots.

this gives a mean length of 4.2 and a maximum length of 138. These numbers refer to the fitting procedure with a fixed tolerance of 1.0, corresponding to a mean error of 1 pixel per component. Even if this seems reasonable, because all compositions are now centered and scaled to fit the unit circle, the actual tolerance expressed in pixels becomes uninterpretable. The fitting tolerance needs to be, somehow, adaptive. Originally larger compositions should call for greater tolerance. We therefore defined  $tolerance = r * unitTolerance$  with  $r$  the scaling radius, and  $unitTolerance = 0.01$  (1% of the unit circle). We also found it necessary to set a lower bound to this adaptive  $tolerance$  ( $minTolerance = 4.0$ , i.e. for any drawing with  $r \leq 400px$ ).

In addition, we wanted to address an important issue associated with the curve fitting algorithm. During the recursive procedure, the algorithm can only add new components to meet the expected error threshold (see Fig.2.25a). However, in some situations, two adjacent components could possibly be merged and simplify the overall curve afterwards. In Fig.2.25b, two components have been joined at the dotted region without any shape loss. For this purpose, on top the original fitting algorithm, a simplification step has been introduced with the adaptive tolerance described above. All details can be found in Algorithm.2.4. Example results of this procedure are displayed in Fig.2.25c. There are only a few components per cubic Bézier curve, yet the main compositional expressiveness is retained.

### Splitting of long strokes

After simplification, the raw distribution of stroke lengths takes an exponential shape (see Fig.2.26c, light gray). We notice that, beyond 4 components, the

## 2 Personal practice

density is almost negligible. However, looking at Fig.2.26a, longer strokes (in terms of components) are also the longest (in distance) and the most complicated ones. In this plot, we measure complexity by the cumulative curvature along the path. The two stroke samples of length 24 and 26 have respectively a cumulative curvature of  $17\pi$  and  $13\pi$ . In addition, they are likely to be the most important part of the composition they belong to. As a result, we cannot simply discard longer strokes or truncate their length, as this would not make sense with regard to the composition. As a first step, we have decided to discard strokes with a linear distance smaller than 0.01, corresponding to a length along its path of 1% of the unit circle. This approach reduces the number of strokes of length 1, but it is not sufficient. Longer strokes are still too few compared to the shorter ones. Therefore, a second step involves splitting longer strokes into multiple sub-strokes (see Algorithm.2.5)<sup>68</sup>. Fig.2.26b shows the resulting spread of stroke lengths compared to their cumulative curvature. Maximum stroke length is 8 and the maximum curvature is equivalent to two complete circles. However, strokes of length 1 remain preponderant (see distribution in Fig.2.26c, dark gray) and will require additional handling during modeling. The final dataset contains 52370 strokes with a mean length of 1.49.

### Stroke encoding details

As a reminder, a composite cubic Bézier curve is a sequence of  $n$  components  $\mathbf{c}_i$  of 2-d control points  $[\mathbf{p}_{i,0}, \mathbf{p}_{i,1}, \mathbf{p}_{i,2}, \mathbf{p}_{i,3}]$ . The last control point of a component is also the same as the first of the following component, and as each stroke begins at the origin, we can limit required information per component to three 2-d points. Furthermore, because tracing a stroke is a continuous action, we can describe its inflections more efficiently using a differential approach. As shown in Fig.2.27a, each control point can be characterized as the difference from previous ones or local neighbors. More precisely, each component  $\mathbf{c}_i$  becomes:

$$\Delta_i = [\delta_i, \delta'_i, \delta''_i] = [\mathbf{p}_{i,3} - \mathbf{p}_{i,0}, \mathbf{p}_{i,1} - \mathbf{p}_{i,0}, \mathbf{p}_{i,2} - \mathbf{p}_{i,3}] \quad (2.18)$$

For modeling purposes, the dataset must be split into training and validation sets. We have chosen a splitting ratio of 0.1, meaning that 10% of the total dataset is kept for model validation. Partitioning is carried out randomly with the constraint of similar stroke length distribution for the two sub-datasets. Concerning data augmentation, at training time, strokes can be vertically and horizontally mirrored or specified in reverse order (see Fig.2.27b). Finally, model inputs have to be standardized before training. As we are mirroring strokes, input values are already centered. We only have to compute a standard deviation for  $\delta_{i,x}, \delta_{i,y}, \delta'_{i,x}, \delta'_{i,y}, \delta''_{i,x}, \delta''_{i,y}$ . To keep spatial homogeneity, we take the average of the standard deviation of  $\delta_{i,x}, \delta_{i,y}$  and similarly for the group of tangent values.

<sup>68</sup>We also apply the procedure described in Subsection.2.3.Composition lengths that limits the number of strokes per composition by discarding the shortest ones when they are too numerous.

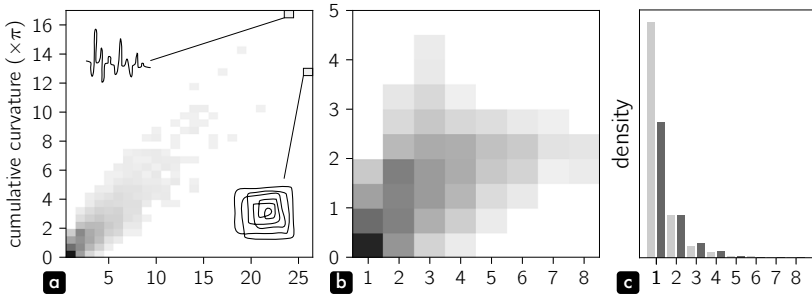


Figure 2.26: Panels **a** and **b** show the spread of stroke length compared with corresponding cumulative curvature: before the splitting procedure is applied to long strokes in **a**, and after in **b**. Panel **a** also presents two complex stroke samples. Panel **c** plots the distribution of stroke length: before the splitting procedure is applied to long strokes in light gray, and after in dark gray.

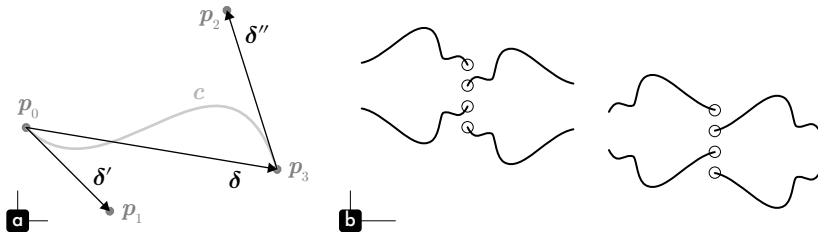


Figure 2.27: Strokes are encoded as sequences of components  $c$  defined by the differential vectors  $\delta, \delta', \delta''$ , as depicted in panel **a**. Panel **b** shows the 8 possible training strokes that can be obtained through data augmentation. A given stroke can be vertically and horizontally mirrored, or specified in reverse order (initial point is circled).

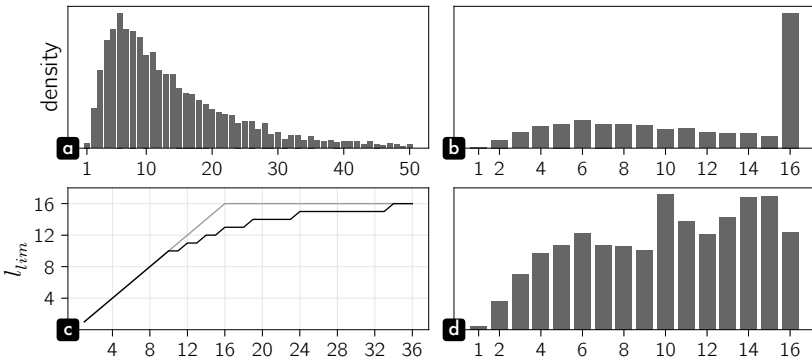


Figure 2.28: Composition length limitation. The raw density of composition lengths is shown in panel **a** (range limited to  $[1, 50]$ ). Discarding smaller strokes after 16 produces a problematic over representation of compositions of length 16 (panel **b**). The panel **c** presents the smoother adaptive limitation of Eq.2.19 (black line) compared to the basic approach (dark gray line). Panel **d** is the resulting density of this procedure.



## 2 Personal practice

### Composition lengths

The distribution of composition lengths takes a lognormal shape with a mode at 6 strokes (Fig.2.28a). Maximum length is 209 and the density is low beyond 30. To help the model, it is a good idea to constrain composition lengths to a tighter range. Empirically, we have decided to set the maximum length to 16. To apply such limitation, a basic approach would be to order all strokes of a composition by length (linear distance along the path) and discard the smaller ones. This procedure logically produces an over representation of compositions of length 16 (see Fig.2.28b). This behavior is then likely to push the model to only generate compositions of this fixed length. To avoid this behavior, we adopted an adaptive maximum length. Given a composition length  $l_{in}$ , a target maximum length  $l_{max}$  and a protected length  $l_{pro}$  (under which no stroke can be discarded), the length limit  $l_{limit}$  is set by:

$$l_{lim} = \begin{cases} l_{in} & \text{if } l_{in} \leq l_{pro} \\ \text{round}(l_{max} - (l_{max} - l_{pro}) \exp(\frac{l_{pro} - l_{in}}{l_{max} - l_{pro}})) & \text{otherwise} \end{cases} \quad (2.19)$$

In Fig.2.28c, this adaptive maximum length corresponds to the black curve, which is smoother than the basic approach is in gray. The resulting density (Fig.2.28d) is then compact and well-balanced. The optimal  $l_{pro}$  has been empirically set to 7. From Fig.2.28c, we notice that the actual *protected* length is 10, due to the rounding operation ( $l_{lim}$  can only be an integer). The final count is a dataset of 5238 compositions and a mean length of 10.00.

### Composition permutations

As stated before, a composition is a sequence of strokes without any pre-defined order. Modeling details will be given later in Section.3.3 and Section.3.4, but a *minima*, we need to define a procedure that shuffles compositions in a coherent manner. Strokes within a composition are not equivalent and may be organized hierarchically. A first level can be defined by group of strokes that are in contact. For instance, in Fig.2.29a, there are 3 different groups. This split is actually operated at the binary map level, before vectorization, so that stroke widths can be taken into account in the *skeleton disentanglement* process. The second level in the hierarchy is the identification of individual strokes. In Fig.2.29b, group 1 is then broken in 3. Finally, too complex (curvy) strokes that have been split into multiple shorter ones, constitute a third hierarchical level (group 0 in Fig.2.29c).

Continuing our example, without any hierarchy, permutation of 6 strokes would lead to 720 possibilities. It is possible that an artist could begin with strokes 1-0 and 1-1, then decides to produce 0, and finally adjust group 1 with 1-2. But from

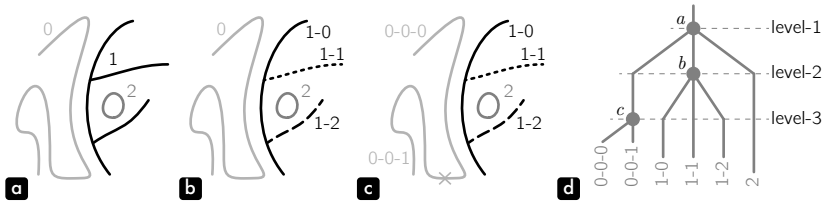


Figure 2.29: Composition permutations are operated on the branches of a 3-level hierarchical tree (panel a: groups of connected lines, panel b: individual strokes, panel c: split of longer strokes). Panel d shows the resulting tree.

the point of view of a spectator, his/her perceptual mechanism involves grouping effects, pushing strokes in connection to appear as one entity. For this reason, we assume that permutations preferably happen down a tree. A representation of this tree is given in Fig.2.29d, where permutations can be made at nodes  $a$ ,  $b$  and  $c$ . This reduces the number of possible permutations to 72 ( $a:6 \times b:6 \times c:2$ ). However, it is quite unlikely that an artist or a viewer will think about strokes 0-0-0 and 0-0-1 independently. So, our permutation algorithm should be able to set a maximal depth of permutation. For the composition dataset, it has been set to level-2.<sup>69</sup> For our example, the count of possible permutation is then reduced to 36 ( $a:6 \times b:6$ ).

Ideally, we would permute compositions at training time, like in any other data augmentation technique. The validation set would then consist of totally unseen drawings. However, our modeling objectives are overly ambitious compared with the proposed dataset statistics, i.e. my compositions are not stereotyped enough for its limited number of inputs ( $\sim 5k$ ). This variety carries artistic richness, but poses practical challenges for our modeling efforts. Previous work on drawings<sup>70</sup> had, for instance, 15 times more elements per specific categories (e.g. cars, cats, ...). As a result, we have decided to pre-compute permutations at the dataset level and to operate the training/validation split per group of shuffled versions of the same drawing.

The main drawback of this choice is the risk of overfitting. In practice, this is unlikely because our model is trained on incomplete compositions (see modeling details of Section.3.3 and Section.3.4, as well as corresponding results in Section.4.1). The second issue is that different compositions will be associated with different numbers of permutations depending on their hierarchical structure. Therefore, a maximum number of permutations has been set to 32. This limitation is operated down the

<sup>69</sup>In experiments with kanji, which are outside the scope of this manuscript, maximum depth has been set to -1. This negative value means that nodes only connected to tree leaves are protected from permutations. In our example, only node  $a$  would be able to produce permutations, i.e. 6 possibilities. It is a coherent behavior for kanji, because we want to shuffle composition of character roots, but not roots themselves.

<sup>70</sup>Ha and Eck, 2017.

## 2 Personal practice

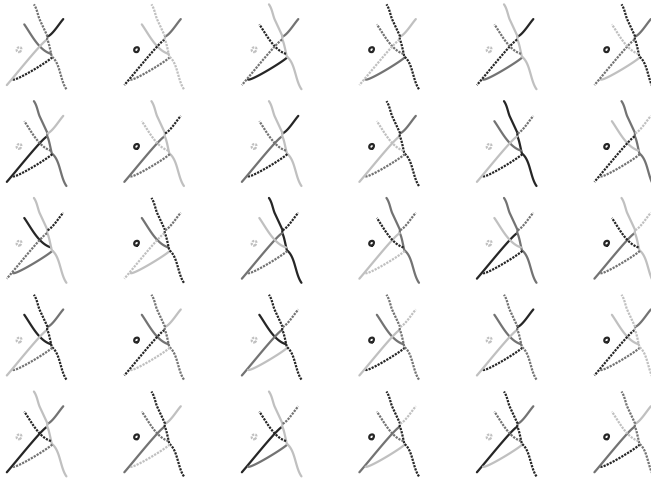


Figure 2.30: 30 permutations of the same composition. The permutation algorithm tries to optimally balance the remaining allowed permutations at every branch but, due to rounding operations, the maximum number of 32 is not always attained.

tree and tries to optimally balance the remaining allowed permutations at every branch. All implementation details can be found in Algorithm.2.6 and a whole set of permutations from the same drawing is presented in Fig.2.30. The final count is a dataset of 143190 compositions with mean length 10.60. This dataset has a larger mean length than the dataset without permutations (10.00), because compositions with fewer strokes do not usually reach the maximum number of 32 permutations.

### Composition encoding details

As compositions will be fed to the model in multiple orders, it would be unwarranted to encode the initial points  $\mathbf{p}_{0,0}$  of successive strokes in a differential way, like for the stroke dataset. The original absolute positioning from the center has been preferred. Then, all  $\mathbf{p}_{0,0}$  need to be standardized. We should first recenter all 2-d points by suppressing the mean, but this procedure would disrupt our specific positioning within the unit circle. We have chosen to scale all  $\mathbf{p}_{0,0}$  by the square root of the second non-central moments given by  $[\mathbf{p}_{0,0}^2]^{1/2}$ . Again, to keep spatial homogeneity, we averaged  $x$  and  $y$  dimensions. Finally, the dataset is split into a training set and a validation set with a similar validation ratio of 0.1. However, compositions labelled as *non-element* (subjective second class compositions, see Subsection.2.3.Manual work) are reserved for the training set only. Validation ratio is therefore corrected accordingly to match the target ratio of 0.1.

**Algorithm 2.1:** Drawing to binary map

---

**In:** · *img*, a linear image in the range  $[0, 1]$   
 · *mask*, a binary cleaning mask  
 · *inverted*, a bool stating if *img* has an inverted scale (None)  
 · *threshold*, maximum intensity of pixels considered as strokes (None)  
 · *kernelSize*, standard deviation of the Gaussian kernel (2.0)  
 · *minArea*, minimum area of black or white surfaces to keep (4)

**Out:** · *binMap*, a binary map

---

```

if inverted is None then
  | inverted ← median(img) < 0.5
if inverted then
  | img ← 1 - img
if img is colored then
  | img ← min(imgr, imgg, imgb)
if threshold is None then
  | threshold ← 0.5 percentile(img, 5) + 0.5 percentile(img, 95)
binMap ← img < threshold
binMap ← binMap * mask
binMap ← gaussianFilter(binMap, kernelSize) > 0.5
→ clean binMap from black and white dots smaller than minArea
return binMap

```

---



---

**Function** processIntersection(*intersection*, *excludeConLines*, *meanStkW*, *minAngle*):

```

if len(intersection) == 0 then
  | return discard intersection and break upper loop
if len(intersection) == 1 then
  | → add closest intersection-point to the last unconnected line extremity
  | return discard intersection and break upper loop
for each line connected to the intersection do
  | if excludeConLines and line in conLines then
  | | → bypass line
  | v ← compute orientation vector of the line from its end to the
  | | min(meanStkW, len(line) - 1) point along the line
  | for each linePair of the intersection do
  | | α ← compute angle between lines v
  | | if α ≥ minAngle and α > matchAngle then
  | | | match ← linePair
  | | | matchAngle ← α
if no match then
  | return break upper loop
  | → connect linePair of match
return continue upper loop

```

---

## 2 Personal practice

---

### Algorithm 2.2: Binary map to individual lines

---

**In:** · *binMap*, a binary map

· *minAngle*, minimal angle between mergeable lines at intersections ( $0.8\pi$ )

· *conLineRatio*, maximum connection-line length as a ratio of the estimated mean stroke width of the drawing (0.75)

**Out:** · *lines*, a list of individual lines (being sequences of 2-d points)

---

▽ extract skeleton and corresponding 2-d points

*skeleton* ← skeletonize strokes of *binMap* to 1-pixel centerlines

*meanStkW* ←  $[\text{sum}(\text{binMap})/\text{sum}(\text{skeleton})]$  ◁ mean stroke width is estimated by dividing the total number of stroke-pixels by the number of skeleton-pixels

*points* ← build a list of 2-d coordinates of every line-pixel from *skeleton*

**for** each *point* in *points* **do**

└ *neighbors[point]* ← list of adjacent line-*point* (8 at most)

▽ build lines

**for** each *point* in *points* **do**

└ **if**  $\text{len}(\text{neighbors}[\text{point}])$  in (1,2) **then** ◁ line-*point* and extremity-*point*

└└ *lines[line]* ←+ *point*

▽ build intersections

*intersections* ← clusterize all connected intersection-*point* ( $\text{len}(\text{neighbors}[\text{point}]) > 2$ ) and collect every adjacent *line*

→ discard any *line* of length 1 and merge every associated *intersection*.

*conLines* ← any *line* with length  $< \text{conLineRatio} * \text{meanStkW}$

→ merge *intersection* linked by a *conLine*

▽ process intersections excluding connection lines

**for** each *intersection* in *intersections* **do**

└ **while** 2 *line* have been connected at *intersection* **do**

└└ → **processIntersection**(*intersection*, *excludeConLines* = True, ...)

▽ process intersections including connection lines

**for** each *intersection* in *intersections* **do**

└ → discard any *conLine* being the last unconnected *line* at *intersection*

└ **while** 2 *line* have been connected at *intersection* **do**

└└ → **processIntersection**(*intersection*, *excludeConLines* = False, ...)

▽ final processing

→ discard all remaining *conLine* from *intersections*

→ add closest intersection-*point* to all unconnected *line* extremities remaining at each *intersection*

→ close (if not closed yet) any loop-*line* by duplicating its first *point* at the end

**return** *lines*

---

**Algorithm 2.3:** Composite cubic Bézier curves fitting

---

**In:** · *points*, a sequence of 2-d points  
 · *tolerance*, maximum distance between original points and fitted curve (1.0)  
 · *tgtDist*, distance (in index) along *points* to compute tangents (10)  
 · *tgtDistDiv*, dividing value of *points* length to set a maximum distance (in index) to compute tangents (10)  
 · *loopDist*, distance between *points* extremities to consider it as a loop (2.0)  
 · *optimSteps*, maximum number of optimization steps (100)

**Out:** · *compCurve*, a composite cubic Bézier curve

---

```

tgtDist ← max(1, min(tgtDist, len(points)/tgtDistDiv, len(points) - 2))
tgta ← normalize(points[tgtDist] - points[0])           ↙ starting tangent
tgtb ← normalize(points[-1] - points[-tgtDist - 1])   ↘ ending tangent
▽ improve closed curve tangents, if loop extremities are in opposition
if ||points[-1] - points[0]|| < loopDist and tgta · tgtb < 0 then
  | tgta ← normalize(tgta - 0.5tgtb)
  | tgtb ← normalize(tgtb - 0.5tgta)
return fitCubicBezier(points, tgta, tgtb, ...)

```

---



---

**Function** fitCubicBezier(*points*, *tgt<sub>a</sub>*, *tgt<sub>b</sub>*, *tolerance*, *tgtDist*, *tgtDistDiv*, *loopDist*, *optimSteps*):

```

▽ with 2 points, return a straight line
if len(points) = 2 then
  | β ← 1/3 ||points[-1] - point[0]||
  | return [[points[0], points[0] + β tgta, points[-1] + β tgtb, points[-1]]]
▽ initial attempt to fit the curve
u ← chordLengthParam(points)
nodes ← computeNodes(points, u, tgta, tgtb)
error ← computeMeanError(points, nodes, u)
if error < tolerance then
  | return [nodes]
▽ if error is not too high, try reparameterization
if error < 4 tolerance then
  for each optimSteps do
    | u ← paramOptim(nodes, points, u)
    | nodes ← computeNodes(points, u, tgta, tgtb)
    | error, splitIdx ← computeMeanError(points, nodes, u)
    | if error < tolerance then
      | | return [nodes]
▽ fitting failed, so split and fit recursively
tgtDist ← max(1, min(tgtDist, splitIdx, len(points) - 1 - splitIdx,
  len(points)/tgtDistDiv/2))
tgtmid ← normalize(points[splitIdx - tgtDist] - points[splitIdx + tgtDist])
compCurve ← fitCubicBezier(points[: splitIdx + 1], tgta, tgtmid, ...)
compCurve += fitCubicBezier(points[splitIdx :], -tgtmid, tgtb, ...)
return compCurve

```

---

## 2 Personal practice

---

---

**Function** `chordLengthParam`(*points*):

$\text{tangentNorms} \leftarrow$  norm of local difference of *points*  
 $u \leftarrow \text{cumsum}(\text{concatenate}([0], \text{tangentNorms}))$   
 $u \leftarrow u/u[-1]$   
**return**  $u$

---

---

---

**Function** `computeNodes`(*points*,  $u$ ,  $\text{tgt}_a$ ,  $\text{tgt}_b$ ):

$\nabla$  We try to minimize the distance between a Bézier curve  $c(u)$  and its corresponding original *points*.  $c$  control points are  
 $[points[0], points[0] + \beta_a \text{tgt}_a, points[-1] + \beta_b \text{tgt}_b, points[-1]]$ .  
 $\beta_a, \beta_b \leftarrow$  solve or find the best solution of the linear system  $c(u) - points = \mathbf{0}$   
 $\triangleleft$  please find details in the original paper appendices (Schneider, 1990)  
**return**  $[points[0], points[0] + \beta_a \text{tgt}_a, points[-1] + \beta_b \text{tgt}_b, points[-1]]$

---

---

---

**Function** `computeMeanError`( $points_{origin}$ ,  $nodes_{optim}$ ,  $u$ ):

$points_{optim} \leftarrow \text{evalBezier}(nodes_{optim}, u)$   
 $errors \leftarrow$  point-wise l2-norm between  $points_{origin}$  and  $points_{optim}$   
 $splitIdx \leftarrow$  index of  $\max(errors)$   
**return**  $\text{mean}(errors)$ ,  $splitIdx$

---

---

---

**Function** `paramOptim`( $points_{origin}$ ,  $nodes_{optim}$ ,  $u$ ):

$\nabla$  We try to find a better parameterization of  $u$  that pushes points of the fitted Bézier curve closer to their corresponding original points.  
 $c(u), c'(u), c''(u) \leftarrow \text{evalBezier}(nodes_{optim}, u)$  and derivatives  
**for** each  $i, p$  in  $u, points_{origin}$  **do**  
 $\nabla$  distance is minimal when  $p$  is perpendicular to  $c(i)$ , so we apply one step of Newton-Raphson's method to solve  $f(i) = (c(i) - p) \cdot c'(i) = 0$  with  
 $f'(i) = c'(i) \cdot c'(i) + (c(i) - p) \cdot c''(i)$   
 $i \leftarrow i - f(i)/f'(i) = i - ((c(i) - p) \cdot c'(i)) / (c'(i) \cdot c'(i) + (c(i) - p) \cdot c''(i))$   
**return**  $u$

---

**Algorithm 2.4:** Composite cubic Bézier curves simplification

**In:** · *compCurve*, a composite cubic Bézier curve  
 · *tolerance*, maximum distance between original and simplified curves (1.0)  
 · *optimSteps*, maximum number of optimization steps (100)

**Out:** · *compCurve*, simplified composite cubic Bézier curve

---

```

for each nodes in compCurve do
  | compPtsorigin +← pointsorigin ← evalBezier(nodes)
compCount ← len(compCurve) + 1
while len(compCurve) < compCount do
  | compCount ← len(compCurve)
  | for each consecutive pair of nodes and pointsorigin do
  | | compCurveoptim, compPtsoptim +←
  | |     simplifyNodePair(nodesa, nodesb, ptsa,origin, ptsb,origin, ...)
  | if len(compCurve) is odd then
  | | compCurveoptim, compPtsoptim +← last nodes and last points
  | compCurve, compPtsorigin ← compCurveoptim, compPtsoptim
  | if len(compCurve) is odd then
  | | compCurveoptim, compPtsoptim ← first nodes and first points
  | | for each next consecutive pair of nodes and pointsorigin do
  | | | compCurveoptim, compPtsoptim +←
  | | |     simplifyNodePair(nodesa, nodesb, ptsa,origin, ptsb,origin, ...)
  | | | compCurve, compPtsorigin ← compCurveoptim, compPtsoptim
return compCurve

```

---



---

**Function** simplifyNodePair(*nodes<sub>a</sub>*, *nodes<sub>b</sub>*, *pts<sub>a,origin</sub>*, *pts<sub>b,origin</sub>*, *tolerance*, *optimSteps*):

```

pointsorigin ← concatenate(ptsa,origin, ptsb,origin)
u ← chordLengthParam(pointsorigin)
tgta ← normalized input tangent from nodesa
tgtb ← normalized output tangent from nodesb
nodesoptim ← computeNodes(pointsorigin, u, tgta, tgtb)
error ← computeMeanError(pointsorigin, nodesoptim, u)
if error < 4 tolerance then
  | for each optimSteps do
  | | u ← paramOptim(nodesoptim, pointsorigin, u)
  | | nodesoptim ← computeNodes(pointsorigin, u, tgta, tgtb)
  | | error ← computeMeanError(pointsorigin, nodesoptim, u)
  | | if error < tolerance then
  | | | | return [nodesoptim], [pointsorigin]
return [nodesa, nodesb], [ptsa,origin, ptsb,origin]

```

---



## 2 Personal practice

---

### Algorithm 2.5: Splitting of long strokes

---

**In:** · *compCurve*, a composite cubic Bézier curve, i.e. a stroke  
· *protectedLength*, upper bound of the protected number of components (3)  
· *maxCurvature*, maximum cumulative curvature of a stroke ( $2.5\pi$ )

**Out:** · *compCurves*, a list of smaller composite cubic Bézier curves

---

**if**  $\text{len}(\text{compCurve}) \leq \text{protectedLength}$  **then**

└ **return** [*compCurve*]

**for** each *nodes* in *compCurve* **do**

└ *compCumCurvature*  $\leftarrow$  *evalCumulativeCurvature(nodes)*

**if** *compCumCurvature*  $\leq$  *maxCurvature* **then**

└ **return** [*compCurve*]

$n \leftarrow \lceil \text{compCumCurvature} / \text{maxCurvature} \rceil$

*compCurves*  $\leftarrow$  find  $n$  groups of curve components that best equalize their respective cumulative curvature (groups can eventually be  $>$  *maxCurvature*)

**return** *compCurves*

---

---

### Algorithm 2.6: Composition permutations

---

**In:** · *codes*, stroke hierarchical codes of a composition, e.g. [(0,0,0), (0,0,1), (1,0), (1,1), (1,2), (2)]  
· *maxPerms*, maximal number of permutation, if 0 no limitation (32)  
· *depthLimit*, depth where permutations begin to be fixed (2)  
· *seed*, seed of the random number generator

**Out:** · *permIndices*, a list of permutation indices

---

*indices*  $\leftarrow$  *range(len(codes))*

*permIndices*  $\leftarrow$  *treePermutations(codes, indices, maxPerms, depth = 0, ...)*

**return** *permIndices*

---

---

### Function *permutationsCount(codes, countLimit, depth, depthLimit)*:

└  $\nabla$  This function is recursive and early stops when *countLimit* is reached.

*countLimit*  $\leftarrow$   $\min(\text{countLimit}, 8! = 40300)$

*count*  $\leftarrow$  number of possible permutations for the tree defined by *codes* and respecting *depthLimit*

└ **return** *count*

---

---

### Function *productNormalize(x, targetProduct)*:

└ **if** *targetProduct*  $\leq 1.0$  **then**

└ **return** vector having the shape of *x* and filled with ones

**if**  $\sum \log(x) \leq \log(\text{targetProduct})$  **then**

└ **return** *x*

└ **return**  $\exp\left(\frac{\log(x)}{\sum \log(x)} \log(\text{targetProduct})\right)$

---

---



---

**Function** `permutationGivenIdx(sequence, i)`:

▽ This function is designed to avoid the computation of all permutations before selecting the  $i^{\text{th}}$  one.

└ **return**  $i^{\text{th}}$  permutation of *sequence*

---



---



---

**Function** `treePermutations(codes, indices, maxPerms, depth, depthLimit, seed)`:

*codesMaxDepth* ← maximum depth in *codes*

**if** *codesMaxDepth* > 0 **and** ((*depthLimit* > 0 **and** *depth* < *depthLimit*) **or** (*depthLimit* ≤ 0 **and** *codesMaxDepth* > -*depthLimit*) **then**

└ ▽ build tree at current level

*treeCodes, treeIndices* ← split *codes* and *indices* in groups having the same first code digit and remove this digit, e.g. *treeCodes* = [[(0,0), (0,1)], [(0), (1), (2)], []]

└ ▽ compute remaining *maxPerm* per branch of the tree

*nodePermsCount* = len(*treeCodes*)!

**for** each *branchCodes* in *treeCodes* **do**

└ *treePermsCount* +← *branchPermsCount* ←

└ **permutationsCount**(*branchCodes, maxPerms, depth, ...*)

*treeMaxPerms* ← **productNormalize**(*treePermsCount,*

*maxPerms/nodePermsCount*)

└ ▽ recursion in deeper levels

**for** each *branchCodes, branchIndices, branchMaxPerms* in

*treeCodes, treeIndices, treeMaxPerms* **do**

└ *treePermIndices* +← *branchPermIndices* ← **treePermutations**(  
└ *branchCodes, branchIndices, branchMaxPerms, depth + 1, ...*)

└ ▽ apply permutations

*maxPerms* ← min(*nodePermsCount, [maxPerms]*)

*selectedIndices* = range(*maxPerms*)

**if** *nodePermsCount* > *maxPerms* **then**

└ *selectedIndices* ← randomly choose *maxPerms* elements from  
└ *selectedIndices* using *seed*.

**for** each *i* in *selectedIndices* **do**

└ *permIndices* +← **permutationGivenIdx**(*treePermIndices, i*)

└ **return** *permIndices*

**return** [*indices*]

---



## 3 Model implementation

Model implementation is a critical step for transposing theoretical ideas into practical objects, i.e. computational tools, which in turn make it possible to carry out concrete experiments. As our stroke and composition datasets are now operational, this chapter focuses on the mathematical background underpinning our computational tools, and describes how they are articulated to materialize the paradigm detailed in Chapter.1. We also explain how artificial neural networks implement suitable probabilistic models, and how our chosen architectural designs address specific compositional questions. The resulting models are then treated in the order of their hierarchical nesting. As detailed earlier, my datasets are characterized by a rich compositional diversity that is actualized over a necessarily limited number of inputs. This limitation poses important functional challenges for neural networks, which required extra care during implementation. We finally endeavor to rationalize the meaning/significance of different hyperparameters, and to simplify the training procedure. However, the presentation of the results from these models is postponed to the second part of the manuscript (Part.II), alongside relevant perceptual experiments and artistic explorations.

### 3.1 Probabilistic models

A probabilistic model  $P$ , also called a statistical model, is the definition of a probability law over a set of possible observations e.g. every possible stroke or composition. These possible observations constitute the *sample* space, which will be represented by a continuous random vector  $\mathbf{x}$  in a potentially high-dimensional space  $\mathcal{X}$ . A probability law associates a probability  $p(\mathbf{x} = \mathbf{x}) \geq 0$  to any sample  $\mathbf{x}$ , and the probability over the entire space  $\mathcal{X}$  fulfills  $p(\mathbf{x} \in \mathcal{X}) = \int_{\mathcal{X}} p(\mathbf{x}) d_{\mathbf{x}} = 1$ . The model  $P$  specifying law  $p$  is a *good* model if it captures a coherent organization for  $\mathbf{x}$ , and confers an informative shape to the distribution of  $\mathbf{x}$ .

To control the shape of our model,  $P$  is parameterized by a vector  $\boldsymbol{\theta}$  in some space  $\Theta$ . Because  $\mathbf{x}$  is continuous,  $p(\mathbf{x}; \boldsymbol{\theta})$  denotes the probability density function of  $\mathbf{x}$ . Depending on the modeled phenomenon, it is usually not possible to compute the true  $\boldsymbol{\theta}$  numerically: the best we can do is obtain an approximation. We can learn this approximation from a dataset  $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ , consisting of  $N$  i.i.d (independent and identically distributed) *real* samples of  $\mathbf{x}$ . We regard  $p_{\mathcal{D}}(\mathbf{x})$  as the true distribution of the data we are trying to model.

### 3 Model implementation

#### Maximum likelihood

We must define a procedure for computing the approximated parameters  $\theta^*$ . In machine learning, a popular choice is *maximum likelihood*. “The intuition behind this framework, is that if the model is likely to produce training set samples  $\mathbf{x}_n$ , then it is also likely to produce similar samples  $\mathbf{x}$ , and unlikely to produce dissimilar ones.”<sup>1</sup> Under this assumption, our probabilistic model  $P$  also becomes a *generative model*. The objective of generative modeling is twofold: description of a given phenomenon, and construction of a tool that is able to produce new unseen coherent occurrences of said phenomenon. “A generative model simulates how the data is generated in the real world.”<sup>2</sup> This goal can be expressed as:<sup>3</sup>

$$\begin{aligned}\theta^* &= \arg \max_{\theta} \prod_{n=1}^N p(\mathbf{x}_n; \theta) \\ &= \arg \max_{\theta} \log \prod_{n=1}^N p(\mathbf{x}_n; \theta) \\ &= \arg \max_{\theta} \sum_{n=1}^N \log p(\mathbf{x}_n; \theta)\end{aligned}\tag{3.1}$$

To avoid the multiplication of probabilities, which leads to smaller and smaller values thus causing numerical problems, we convert product to summation via the logarithmic function. This is possible because the derivative of the logarithm is strictly positive, so that  $\arg \max_u f(u) = \arg \max_u \log f(u)$ .

*Maximum likelihood* can be thought of as a procedure that minimizes the distance between  $p(\mathbf{x}; \theta)$  and  $p_{\mathcal{D}}(\mathbf{x})$  as measured by the Kullback-Leibler divergence metric  $D_{\text{KL}}$ .

$$\begin{aligned}\theta^* &= \arg \min_{\theta} D_{\text{KL}}(p_{\mathcal{D}}(\mathbf{x}) \parallel p(\mathbf{x}; \theta)) \\ &= \arg \min_{\theta} \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [\log p_{\mathcal{D}}(\mathbf{x}_n) - \log p(\mathbf{x}_n; \theta)] \\ &= \arg \min_{\theta} \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [\log p_{\mathcal{D}}(\mathbf{x}_n)] - \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [\log p(\mathbf{x}_n; \theta)] \\ &= \arg \max_{\theta} \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [\log p(\mathbf{x}_n; \theta)] \\ &= \arg \max_{\theta} \sum_{n=1}^N p_{\mathcal{D}}(\mathbf{x}_n) \log p(\mathbf{x}_n; \theta) \\ &= \arg \max_{\theta} \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}_n; \theta)\end{aligned}\tag{3.2}$$

<sup>1</sup>Doersch, 2016.

<sup>2</sup>Kingma and Welling, 2019.

<sup>3</sup>I. Goodfellow et al., 2016, pp. 131–132.

We can remove  $\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\log p_{\mathcal{D}}(\mathbf{x}_n)]$  as it does not depend on  $\theta$ . In addition,  $p_{\mathcal{D}}(\mathbf{x}_n) = \frac{1}{N}$  by definition. So, up to a constant, the result is similar to Eq.3.1. In order to simplify the formulation of this objective, we define the maximization function  $\mathcal{L}(\mathbf{x}, \theta) = \log p(\mathbf{x}; \theta)$ .

#### Time series

Strokes and compositions can be considered as time series of  $T$  movements or events. Within this framework, the random vector  $\mathbf{x}$  is also defined as a temporal sequence:  $\mathbf{x} = \mathbf{x}_{1:T} = (\mathbf{x}_1, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T)$ . Because time is irreversible, an event at instant  $t$  is a consequence of all preceding events: any  $\mathbf{x}_t$  from a sequence cannot be considered as independent of  $\mathbf{x}_{1:t-1}$ . As a result,  $p(\mathbf{x}_{1:T})$  is as an ordered product of conditional distributions. For example, if  $T = 3$ , then  $p(\mathbf{x}_{1:3}) = p(\mathbf{x}_1)p(\mathbf{x}_2 | \mathbf{x}_1)p(\mathbf{x}_3 | \mathbf{x}_1, \mathbf{x}_2)$ . More generally:

$$p(\mathbf{x}_{1:T}) = p(\mathbf{x}_1, \dots, \mathbf{x}_T) = p(\mathbf{x}_1) \prod_{t=2}^T p(\mathbf{x}_t | \mathbf{x}_{1:t-1}) = \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{x}_{1:t-1}), \mathbf{x}_{1:0} = \emptyset \quad (3.3)$$

The maximization function for the time series becomes:

$$\begin{aligned} \mathcal{L}(\mathbf{x}_{1:T}, \theta) &= \log p(\mathbf{x}_{1:T}; \theta) \\ &= \log \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{x}_{1:t-1}; \theta), \mathbf{x}_{1:0} = \emptyset \\ &= \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{x}_{1:t-1}; \theta), \mathbf{x}_{1:0} = \emptyset \end{aligned} \quad (3.4)$$

#### Neural networks

Probabilistic models based on *maximum likelihood* do not imply the adoption of neural networks. In the family of bio-inspired machine learning tools, artificial evolution (genetic algorithms) and cellular automata are other forms of *Artificial Intelligence*<sup>4</sup>. Frameworks such as Bayesian networks or Support Vector Machines (SVMs) are also popular, and have been widely used before the recent popularity of deep learning<sup>5</sup>. Neural networks were also introduced early, and have been around since the early days of computers. However, it took five decades of developing more and more powerful computational resources before the ideas proposed by McCulloch and Pitts in 1943<sup>6</sup> could be used to solve real-life problems. Zip Code recognition, i.e. the automatic processing of handwritten digits, represented a

<sup>4</sup>Floreano and Mattiussi, 2008.

<sup>5</sup>Bishop, 2006.

<sup>6</sup>Rojas, 1996.

### 3 Model implementation

milestone in the rebirth of neural networks<sup>7</sup>. The related MNIST Database<sup>8</sup> created in 1998 remains an essential sandbox to experiment with neural networks. The contemporary widespread use of deep learning, supported by neural networks with a large and complex accumulation of layers, capitalizes upon the development of computing frameworks based on GPU (Graphic Cards) rather than CPU. These technological innovations have reduced training time by more than 10 folds, making it possible to handle larger datasets. A second factor in the development of deep learning is the rise of the internet and its huge pool of data. The ImageNet dataset<sup>9</sup> and its associated competition have driven the development of different architectures dedicated to image content classification, and more generally object recognition<sup>10</sup>. Some models of artificial vision have achieved super-human performance, and have triggered substantial interest in the cognitive sciences community. Connections with biological early brain processing have even been demonstrated<sup>11</sup>. We also have found similarities between deep convolutional neural networks and human judgments of the orientation of abstract paintings<sup>12</sup> (see Appendix.A.1).

Over a period of only ten years, deep learning has become an efficient and versatile computational tool. Driven by leaders of the digital industry like Google and Facebook, open source libraries such as TensorFlow<sup>13</sup> and PyTorch<sup>14</sup> provide high-level frameworks with automatic differentiation for the gradient computation and a transparent switch from CPU to GPU backends. Developers and researchers can then focus on network architecture rather than low-level considerations. A negative side effect of this success is that deep learning has become a field with an extremely high publication rate, difficult to track. Source codes of new models are usually freely available online, allowing the resolution of specific problems with reduced development efforts. However, if the main objective is not to solve a specific problem or fulfill a specific task with high accuracy, but rather to gain insights into a complex phenomenon and to understand the meaning of each architectural detail, it can be wise to take a step back: amidst yearly new architectures and dozens of variants, we have decided to stay with concepts that have a *longer* history in neural network research.

There is no space for an extensive introduction to neural networks, however we need to make the link with the maximization function in Eq.3.4. A simple neuron is a processing unit collecting a finite number of inputs. These inputs are linearly combined with individual weighting values. These weights are precisely the learned parameters, corresponding to fragments of  $\theta$ . Nonetheless, if a neuron were

---

<sup>7</sup>LeCun et al., 1989.

<sup>8</sup>LeCun et al., 1998.

<sup>9</sup>Russakovsky et al., 2015.

<sup>10</sup>He et al., 2015; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014.

<sup>11</sup>DiCarlo et al., 2012; Rajalingham et al., 2018; Yamins and DiCarlo, 2016.

<sup>12</sup>Lelièvre and Neri, 2019, 2021.

<sup>13</sup>Abadi et al., 2015.

<sup>14</sup>Paszke et al., 2019.

solely a linear combination of inputs, neural networks would not be far from linear regression algorithms. A neuron is therefore associated with a nonlinear function, e.g. an exponential or a sigmoid function. Alone, a neuron would not be sufficient to model complex functions and systems. When arranged over an array, neurons form a layer. If multiple layers are stacked and interconnected, we obtain a complete network. The size of each layer can be adjusted to match the requirements of input/output dimensionalities and the expected *modeling power*. Neural weights are randomly initialized and then information is fed forward to produce an output. Through Eq.3.4, this output provides an error estimation that can be exploited within the context of traditional gradient descent (or ascent) algorithms, thus optimizing  $\theta$  iteratively. In neural networks, this optimization step is denominated as error back-propagation, in opposition to the forward pass that produces the output. Error back-propagation corresponds to computing the partial derivatives of the error w.r.t. each weight by using the chain rule back to the input. As a result, the exact behavior of intermediary layers is shaped during the training without explicit control on the part of the modeler. In addition, intermediary layers remain unavailable to scrutiny based on the output alone. For this reason, these layers are usually called *hidden layers*.

#### **Recurrent neural networks**

The neural network architectures dedicated to vision outlined above belong to the Convolutional Neural Networks family (CNNs), i.e. approximately models which deal with discrete images of a fixed number of pixels. Because input and output dimensionalities are pre-defined, these approaches do not lend naturally to the handling of time series: even if we could standardize the length of input sequences beforehand, feeding them to a CNN would lead to the unwieldy expansion of first layers by the length of the series. A better idea is to reuse the same layers multiple times, with a recurrent connection, i.e. the output of a unit being fed to the same unit, a step in time further. This particular connection gives its name to a whole class of neural networks, the recurrent neural networks (RNNs). The difficulty is then to keep track of the gradient for error back-propagation. The problem can be formalized as a directed graph by unfolding/unrolling the network along its time dimension. Helpfully, this job is carried transparently by PyTorch, our deep learning toolbox of choice.

Early successful applications of RNNs to text and handwriting generation<sup>15</sup> relied on a special type of recurrent unit called Long Short-Term Memory (LSTM)<sup>16</sup>. This unit aims at improving model performance on long-term dependencies in sequential data (i.e. avoiding a *vanishing* gradient  $\rightarrow 0$ , or an *exploding* gradient  $\rightarrow \infty$ ). For this purpose, a memory cell is introduced in the recurrent unit to store important

---

<sup>15</sup>Graves, 2013.

<sup>16</sup>Gers et al., 2000; Hochreiter and Schmidhuber, 1997.



### 3 Model implementation

information. Over time, this memory cell is updated with new information provided by the input and the previous state. Elimination of unnecessary data is also driven by a forget gate. The memory cell finally influences the current state of the unit, its *output*. Another type of recurrent unit, the Gated Recurrent Unit (GRU)<sup>17</sup>, claims similar results in a variety of applications with fewer parameters. However, the LSTM unit remains a more versatile choice.

The simplest way to address time series is the pure generative RNN architecture presented in Fig.3.1a. It naturally involves a layer with a recurrent unit supported by a random vector denoted  $\mathbf{h}$ . A second random vector  $\mathbf{y}$  materializes a basic linear layer (fully connected), from which it is possible to draw some  $\mathbf{x}'$ .  $\mathbf{x}$  and  $\mathbf{x}'$  actually refer to the same random variable. The ' notation emphasizes the generative action of the model. In opposition to any  $\mathbf{x}$ ,  $\mathbf{x}'$  is specifically *not* a real item  $\mathbf{x}_n$  from the dataset  $\mathcal{D}$ . Fig.3.1b is an unfolded representation of Fig.3.1a along the time dimension. This graph gives a clearer overview of the computational dependencies. For instance, gray arrows show that  $\mathbf{x}'_t$  are used in place of normal  $\mathbf{x}_t$  inputs at evaluation, for the generation of new sequences. It also gives useful insight into the definition of the following deterministic functions:

$$\mathbf{h}_t = f_{\mathbf{h}}(\mathbf{x}_t, \mathbf{h}_{t-1}; \boldsymbol{\theta}_{\mathbf{h}}), \quad \mathbf{x}_0 = \mathbf{0}, \quad \mathbf{h}_{-1} = \mathbf{0} \quad (3.5)$$

$$\mathbf{y}_t = f_{\mathbf{y}}(\mathbf{h}_t; \boldsymbol{\theta}_{\mathbf{y}}) \quad (3.6)$$

where  $\boldsymbol{\theta}_{\mathbf{h}}$  and  $\boldsymbol{\theta}_{\mathbf{y}}$  are two parts of the model parameter  $\boldsymbol{\theta}$ .  $\boldsymbol{\theta}_{\mathbf{h}}$  is a vector of weights specifying the behavior of the LSTM unit, and  $\boldsymbol{\theta}_{\mathbf{y}}$  is a vector of weights of the linear layer. Functions  $f_{\mathbf{h}}$  and  $f_{\mathbf{y}}$  also include nonlinearities, such as the ReLU activation function<sup>18</sup>. To introduce the dependency of model  $P$  on  $\mathbf{h}$  and  $\mathbf{y}$ ,  $p(\mathbf{x}_t | \mathbf{x}_{1:t-1})$  can be marginalized against those random vectors. Beginning with  $\mathbf{h}$ :

$$\begin{aligned} p(\mathbf{x}_t | \mathbf{x}_{1:t-1}) &= \int_{\mathbf{h}_{t-1}} p(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{h}_{t-1}) p(\mathbf{h}_{t-1} | \mathbf{x}_{1:t-1}) d\mathbf{h}_{t-1} \\ &= p(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{h}_{t-1}) \\ &= p(\mathbf{x}_t | \mathbf{h}_{t-1}(\mathbf{x}_{1:t-1})) \\ &= p(\mathbf{x}_t | \mathbf{y}_{t-1}(\mathbf{h}_{t-1}(\mathbf{x}_{1:t-1}))) \end{aligned} \quad (3.7)$$

As  $\mathbf{h}_t$  is the result of the deterministic function  $f_{\mathbf{h}}(\mathbf{x}_t, \mathbf{h}_{t-1})$ , where  $\mathbf{h}_{t-1}$  can be recursively and deterministically defined given  $\mathbf{x}_{1:t-1}$ , then  $p(\mathbf{h}_{t-1} | \mathbf{x}_{1:t-1})$  follows a Dirac distribution, with its mode given by Eq.3.5. As a result, the integral over the hidden states can be replaced by a single point. Then, we make the dependency of  $\mathbf{h}_{t-1}$  on  $\mathbf{x}_{1:t-1}$  explicit.  $\mathbf{y}_{t-1}$  is introduced using the same procedure<sup>19</sup>. Finally, our maximization function can be expressed as:

<sup>17</sup>Cho et al., 2014.

<sup>18</sup>Clevert et al., 2015; Nair and Hinton, 2010.

<sup>19</sup>Parts of this demonstration come from Bayer and Osendorfer, 2015

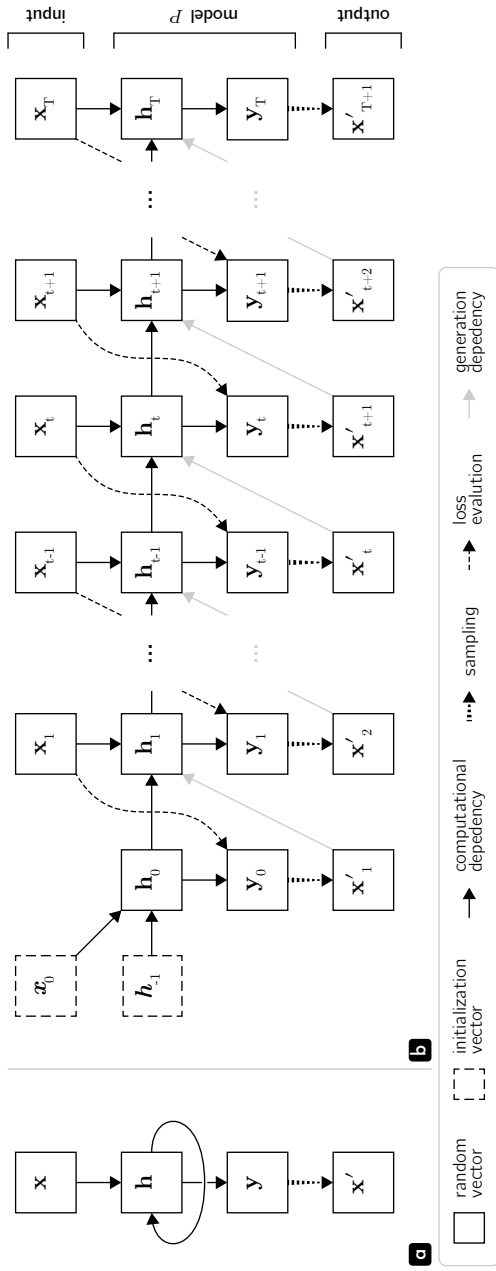


Figure 3.1: Simple generative RNN architecture. The graph in panel **b** is an unfolded representation of panel **a** along the time dimension, better highlighting computational dependencies.

### 3 Model implementation

$$\mathcal{L}(\mathbf{x}_{1:T}, \boldsymbol{\theta}) = \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{y}_{t-1}(\mathbf{h}_{t-1}(\mathbf{x}_{1:t-1}; \boldsymbol{\theta}_h); \boldsymbol{\theta}_y)) \quad (3.8)$$

This generative RNN architecture has been stated as the simplest way to address time series. Indeed, our experiments with compositions highlight one fundamental issue, rendering this model inapplicable in practice. Designed at first to extend handwriting materials, texts, or sounds with a coherent style, new outputs are actually conditioned on several previous time samples. There is no real need for the generation of the first line or word. But in our case, the generation of the first stroke is mandatory to produce completely new compositions. With this architecture, possibilities on the first stroke are too high and provoke a hazardous initialization of the composition. This causes difficulties when the network attempts to recover expressiveness and propose satisfactory outputs. Therefore, the recurrent unit needs to be conditioned on some information of the targeted composition. From an artistic point of view, we could call this target an *inspiring mental image*.

#### Variational Auto-Encoder

To ease introduction of Variational Auto-Encoders (VAEs), we will momentarily silence the temporal nature of  $\mathbf{x}$ . In order to mathematically express the idea of conditioning expressed above, we assume that  $\mathbf{x}$  is generated by some random process involving another continuous random variable  $\mathbf{z}$ .  $\mathbf{z}$  is defined over a high-dimensional space  $\mathcal{Z}$  and is connected to  $\mathbf{x}$  through the deterministic function  $\mathbf{x} = f_p(\mathbf{z}; \boldsymbol{\theta})$ , where  $f_p : \mathcal{Z} \times \Theta \rightarrow \mathcal{X}$ . Then,  $p(\mathbf{x}; \boldsymbol{\theta})$  can be considered as a marginal distribution against  $\mathbf{z}$ , with  $p(\mathbf{z})$  the prior distribution of  $\mathbf{z}$ .

$$\begin{aligned} p(\mathbf{x}; \boldsymbol{\theta}) &= \int_{\mathbf{z}} p(\mathbf{x} | \mathbf{z}; \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z} \\ &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [p(\mathbf{x} | \mathbf{z}; \boldsymbol{\theta})] \end{aligned} \quad (3.9)$$

The space covered by  $\mathbf{z}$  is “called *latent* because given just an output  $\mathbf{x}$  produced by the model, we do not necessarily know which settings of the latent variable  $\mathbf{z}$  have generated it.”<sup>20</sup> In other words, the latent space is at this stage a different, but inaccessible representation of  $\mathbf{x}$ . Then, we can design the latent space as a simplified and less intricate parameterization of  $\mathbf{x}$ . If  $\mathbf{z}$  were a compressed representation of  $\mathbf{x}$  with reduced dimensionality, it would force the model to extract essential aspects of  $\mathbf{x}$ , and could provide us with the opportunity to discover hidden intrinsic regularities. “This quest for disentangled, semantically meaningful, statistically independent and causal factors of variation in data is generally known as unsupervised representation learning.”<sup>21</sup>

---

<sup>20</sup>Doersch, 2016.

<sup>21</sup>Kingma and Welling, 2019.

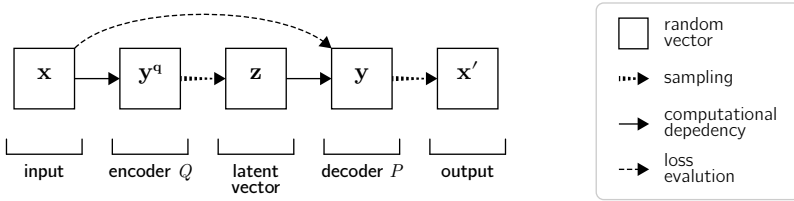


Figure 3.2: Simple VAE architecture.

As a result, we should impose as few *a priori* constraints on  $\mathbf{z}$  as possible, and only make minimal assumptions on its distribution. A simple choice is to set  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$ , with  $\mathbf{I}$  the identity matrix. Despite its generic nature, a normal distribution may not be optimal. Nonetheless, the model  $P$  is implemented using deep neural networks, so we suppose that it will have enough *power* to adapt the prior to whatever necessary internal distribution required by the model.

The efficiency of this model and its highly nonlinear hidden layers are associated with the downside that it becomes intractable for direct optimization using Eq.3.9. During the optimization procedure, if we naively sample some  $\mathbf{z}$  from  $p(\mathbf{z})$ ,  $p(\mathbf{x} \mid \mathbf{z})$  will be nearly 0 for most  $\mathbf{z}$ , and it will not help in training the model i.e. producing an efficient gradient for back-propagation. The idea is therefore to find a way to sample some  $\mathbf{z}$  that are likely to have produced a particular  $\mathbf{x}$ . To do so, let us introduce a second model  $Q$  with a conditional distribution  $q(\mathbf{z} \mid \mathbf{x})$ . This pdf is based on a deterministic function  $\mathbf{z} = f_q(\mathbf{x}; \phi)$ , parameterized by a vector  $\phi$  in some high-dimensional space  $\Phi$ , where  $f_q : \mathcal{X} \times \Phi \rightarrow \mathcal{Z}$ .  $q(\mathbf{z} \mid \mathbf{x}; \phi)$  will be an approximation of the intractable true posterior  $p(\mathbf{z} \mid \mathbf{x})$ .

In Fig.3.2, we notice that the proposed architecture is now structured like an *auto-encoder*. On the one hand, model  $Q$  *encodes*  $\mathbf{x}$  into  $\mathbf{z}$ , and on the other hand model  $P$  *decodes*  $\mathbf{z}$  to reconstruct  $\mathbf{x}$ . We will therefore refer to these models as the probabilistic encoder  $Q$  and decoder  $P$ .  $Q$  and  $P$  are also sometimes referred as recognition and generative models respectively. In addition, this framework is specified as *variational* because  $\mathbf{z}$  is enforced to follow a probability distribution. Even with a trained *basic* auto-encoder, sampling a  $\mathbf{z}$  to produce a new  $\mathbf{x}'$  is not possible: an existing  $\mathbf{x}$  has to be encoded to produce a functional  $\mathbf{z}$ . That is why *basic* auto-encoders are mostly used for compression purposes, that do not require the generative ability.

Concerning the optimization, besides the *maximum likelihood* on model  $P$ , we need to guarantee that the approximation of  $p(\mathbf{z} \mid \mathbf{x})$  by  $q(\mathbf{z} \mid \mathbf{x})$  is sufficient. In other words,  $D_{\text{KL}}(p(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z} \mid \mathbf{x}))$  has to be minimal. But this  $D_{\text{KL}}$ , seen as the expectation  $\mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{z} \mid \mathbf{x}) - \log q(\mathbf{z} \mid \mathbf{x})]$ , is computed with some  $\mathbf{z}$  sampled from  $p(\mathbf{z} \mid \mathbf{x})$ , which is intractable. We will therefore consider the Kullback-Leibler divergence in the opposite direction,  $D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z} \mid \mathbf{x}))$ .

### 3 Model implementation

If  $q(\mathbf{z} \mid \mathbf{x})$  were a perfect approximation of  $p(\mathbf{z} \mid \mathbf{x})$ , there would not be any difference, however  $D_{\text{KL}}$  is generally not symmetric between two distributions, and it has practical implications for model behavior<sup>22</sup>. Nonetheless, we define the following new maximization function:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) = \log p(\mathbf{x}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}; \phi) \parallel p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta})) \quad (3.10)$$

#### Variational lower bound

$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi)$  is then called the *variational lower bound*, or the *evidence lower bound* (ELBO), because the  $D_{\text{KL}}$  term is non-negative and will be very close to zero as  $q(\mathbf{z} \mid \mathbf{x})$  becomes efficient, meaning that  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\theta})$ .  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi)$  therefore places a lower bound on the original *maximum likelihood* (remember  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta})$ ). We render Eq.3.10 practically actionable by expressing  $D_{\text{KL}}$  as an expectation and applying Bayes rule.

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) &= \log p(\mathbf{x}; \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} [\log q(\mathbf{z} \mid \mathbf{x}; \phi) - \log p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta})] \\ &= \log p(\mathbf{x}; \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} [\log q(\mathbf{z} \mid \mathbf{x}) - \log p(\mathbf{x} \mid \mathbf{z}) - \log p(\mathbf{z}) + \log p(\mathbf{x})] \end{aligned} \quad (3.11)$$

But  $p(\mathbf{x})$  does not depend on  $\mathbf{z}$ . So, it can be moved outside the expectation and eliminated:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) = -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} [\log q(\mathbf{z} \mid \mathbf{x}; \phi) - \log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta}) - \log p(\mathbf{z}; \boldsymbol{\theta})] \quad (3.12)$$

Then, contracting parts of  $\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})}$  into a  $D_{\text{KL}}$ :

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} [\log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta})] - D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}; \phi) \parallel p(\mathbf{z}; \boldsymbol{\theta})) \quad (3.13)$$

#### Reparameterization trick

$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi)$  will be optimized using stochastic gradient ascent/descent. This implies that every operation must be differentiable in the model. “Stochastic gradient descent can handle stochastic inputs, but not stochastic units within the network!”<sup>23</sup> On this point,  $\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})} [\log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta})]$  is problematic. During the forward pass, we can average as many samples as needed to obtain a good estimate of the expectation, but the error cannot be back-propagated through the sampling operation of  $\mathbf{z}$ . To overcome this issue, we use the *reparameterization trick*<sup>24</sup>.

<sup>22</sup> Huszár, 2015 exposed this issue on a larger scale about *maximum likelihood*: “Minimizing  $D_{\text{KL}}(P_{\mathcal{D}} \mid P)$  corresponds to moment matching and has a tendency to find models  $P$  that cover all the modes of  $P_{\mathcal{D}}$ , at the cost of placing probability mass where  $P_{\mathcal{D}}$  has none. Minimizing  $D_{\text{KL}}(P \mid P_{\mathcal{D}})$  in this case leads to a mode-seeking behavior: the optimal  $P$  will typically concentrate around the largest mode of  $P_{\mathcal{D}}$ , at the cost of completely ignoring smaller modes.”

<sup>23</sup>Doersch, 2016.

<sup>24</sup>Kingma and Welling, 2013; Rezende et al., 2014.

First, we assert that  $q(\mathbf{z} \mid \mathbf{x})$  follows a normal distribution. Then, the family of normal distributions is closed under linear transformations. It means that we can decouple the sampling operation from the specific transformation operated by  $q(\mathbf{z} \mid \mathbf{x})$ . To be precise, we define  $q(\mathbf{z} \mid \mathbf{x}; \phi) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_\phi(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x})))$ , with vectors  $\boldsymbol{\mu}_\phi(\mathbf{x})$  and  $\boldsymbol{\sigma}_\phi^2(\mathbf{x})$  directly derived from  $\mathbf{y}^q$ . This way, we can first sample an  $\epsilon$  from a distribution  $p(\epsilon) = \mathcal{N}(\epsilon \mid \mathbf{0}, \mathbf{I})$  with  $\mathbf{I}$  the identity matrix, and then compute  $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \epsilon$ . As a result,  $\epsilon$  becomes an auxiliary variable with the independent marginal  $p(\epsilon)$ , which does not need to be learned. So:

$$\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta})] = \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[\log p(\mathbf{x} \mid \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \epsilon; \boldsymbol{\theta})] \quad (3.14)$$

During optimization, a complete iteration over the dataset  $\mathcal{D}$  is called an epoch. Multiple epochs are necessary to obtain a good approximation of  $\boldsymbol{\theta}$  and  $\phi$ . So, at the tiny scale of an individual  $\mathbf{x}_n$  from  $\mathcal{D}$ , we can sample  $\epsilon$  only a limited number of times  $n_\epsilon$  and still get a good estimation of the expectation.  $n_\epsilon = 1$  is usually considered as sufficient but, in practice, we have preferred  $n_\epsilon = 8$ .<sup>25</sup> However, in order to simplify our notation, in the following we consider  $n_\epsilon = 1$  and discard expectation, so that:

$$\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta})] \approx \log p(\mathbf{x} \mid \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \epsilon; \boldsymbol{\theta}) \quad (3.15)$$

Then, the practical maximization function from Eq.3.13 can be simplified to:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) \approx \log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}; \phi) \parallel p(\mathbf{z}; \boldsymbol{\theta})) \quad (3.16)$$

### Kullback-Leibler divergence of normal distributions

The second term of  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi)$  in Eq.3.16 is a  $D_{\text{KL}}$  between two multivariate normal distributions. It can be computed in closed form with:

$$\begin{aligned} D_{\text{KL}}(\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \parallel \mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) = \\ \frac{1}{2} \left( \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - K + \log\left(\frac{\det \boldsymbol{\Sigma}_1}{\det \boldsymbol{\Sigma}_0}\right) \right) \end{aligned} \quad (3.17)$$

where  $K$  is the dimensionality of the random variable of interest, i.e.  $\mathbf{z}$ . In the case of VAEs, it simplifies to:

$$\begin{aligned} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}_\phi^2)) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) \\ = \frac{1}{2} \left( \text{Tr}(\text{diag}(\boldsymbol{\sigma}_\phi^2)) + \boldsymbol{\mu}_\phi^\top \boldsymbol{\mu}_\phi - K - \log(\det(\text{diag}(\boldsymbol{\sigma}_\phi^2))) \right) \\ = \frac{1}{2} \left( \sum_{k=1}^K \sigma_{\phi k}^2 + \sum_{k=1}^K \mu_{\phi k}^2 - K - \log\left(\prod_{k=1}^K \sigma_{\phi k}^2\right) \right) \end{aligned} \quad (3.18)$$

<sup>25</sup>This multiple sampling of  $\epsilon$  to get a better estimate of  $\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{z}; \boldsymbol{\theta})]$  is related to importance weighted auto-encoders (IWAE), which claim to produce a tighter lower bound (Burda et al., 2016).

### 3 Model implementation

leading to,

$$D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}; \phi) \parallel p(\mathbf{z}; \theta)) = \frac{1}{2} \sum_{k=1}^K \sigma_{\phi_k}^2(\mathbf{x}) + \mu_{\phi_k}^2(\mathbf{x}) - 1 - \log(\sigma_{\phi_k}^2(\mathbf{x})) \quad (3.19)$$

## 3.2 Stroke model

Strokes are sequences of  $T$  components of cubic Bézier curves  $\mathbf{c}_t$ , and each component is a fixed vector of 6 values  $\Delta_t = [\delta_{t,x}, \delta_{t,y}, \delta'_{t,x}, \delta'_{t,y}, \delta''_{t,x}, \delta''_{t,y}]$  (see Subsection.2.3.Stroke encoding details). Strokes are then the basic elements of compositions, and it implies an additional constraint. RNN architecture supports compositions of variable length, i.e. stroke number, but each stroke must be defined as a vector of a constant dimensionality, no matter the complexity or length of the stroke. The main purpose of the stroke model is therefore to compress the representation of any stroke into a standardized format. Ideally, the entire expressive space of strokes should be captured and organized into a finite and reduced number of dimensions. This job is typically carried out by VAEs. In this section, we will thus explain how to combine RNNs with VAEs.

### Architecture

The stroke model architecture is presented in Fig.3.3. It is essentially based on a previous work<sup>26</sup> dedicated to figurative sketches collected by Google through the *Quick, Draw!* mini-game<sup>27</sup>. This model aims at reconstructing and generating drawings per class, e.g. cat or car only, and has then been extended to kanji<sup>28</sup>.

In Fig.3.3, we notice that the output of the recurrent unit  $\mathbf{h}^g$  is only used at  $t = T$ . It means that the latent random vector  $\mathbf{z}$  is dependent on the whole sequence of  $\mathbf{x} = \mathbf{x}_{1:T}$ . Even if the latent space is continuous and covers an infinite number of possible strokes, it does not care about incomplete strokes. This architecture is thus a space with circumscribed morphological possibilities, which is coherent with the idea of strokes as a minimal vocabulary for creative endeavors.

Concerning decoder  $P$ , the main difference from the simple generative RNN (see Fig.3.1) is that  $\mathbf{h}$  is also conditioned on  $\mathbf{z}$ . The previous state of  $\mathbf{h}_0$  is specifically computed by a dedicated layer represented by the random vector  $\mathbf{g}$ . The idea is to increase network power by overcoming the *initial stroke issue* at generation (see Subsection.3.1.Recurrent neural networks).

<sup>26</sup>The work of Ha and Eck, 2017 is seminal and has profoundly impacted our project. We should also mention Cho et al., 2014 (basic Auto-Encoder) and Bowman et al., 2016 for their work on representation and generation of sentences.

<sup>27</sup>Jongejan et al., 2016.

<sup>28</sup>Clanuwat et al., 2018.

### 3.2 Stroke model

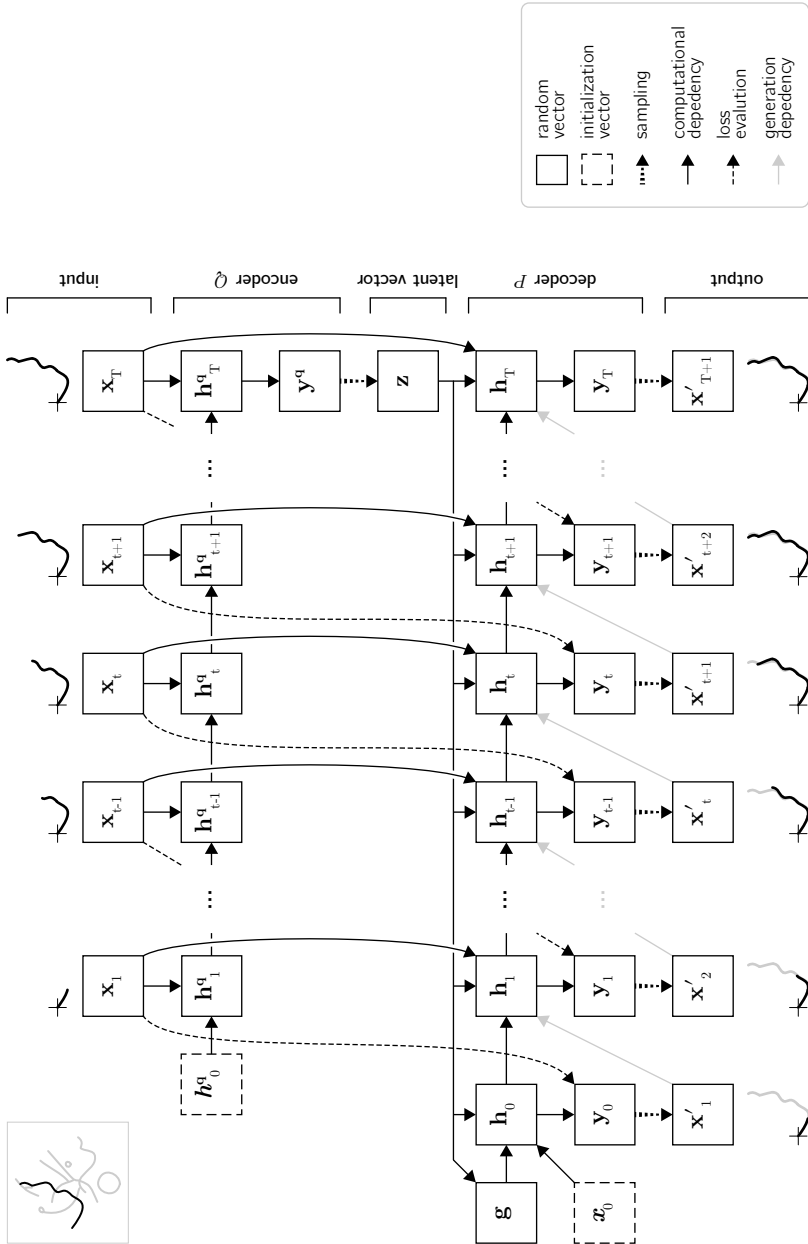


Figure 3.3: Stroke model architecture.



### 3 Model implementation

#### Optimization function

Like a typical VAE, the maximization function of the stroke model is the ELBO defined in Eq.3.16, but extended to time series (see Eq.3.3):

$$\begin{aligned} \mathcal{L}(\mathbf{x}_{1:T}, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \log p(\mathbf{x}_{1:T} | \mathbf{z}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}_{1:T}; \boldsymbol{\phi}) \parallel p(\mathbf{z}; \boldsymbol{\theta})) \\ &= \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{z}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}_{1:T}; \boldsymbol{\phi}) \parallel p(\mathbf{z}; \boldsymbol{\theta})) \end{aligned} \quad (3.20)$$

In order to implement encoder  $Q$  and decoder  $P$  with neural networks, respectively parameterized by  $\boldsymbol{\phi}$  and  $\boldsymbol{\theta}$ , we define the following deterministic functions:

$$\mathbf{h}_t^q = f_{\mathbf{h}^q}(\mathbf{x}_t, \mathbf{h}_{t-1}^q; \boldsymbol{\phi}_{\mathbf{h}}), \quad \mathbf{h}_0^q = \mathbf{0} \quad (3.21)$$

$$\mathbf{y}^q = f_{\mathbf{y}^q}(\mathbf{h}_T^q; \boldsymbol{\phi}_{\mathbf{y}}) \quad (3.22)$$

$$\mathbf{g} = f_{\mathbf{g}}(\mathbf{z}; \boldsymbol{\theta}_{\mathbf{g}}) \quad (3.23)$$

$$\mathbf{h}_t = f_{\mathbf{h}}(\mathbf{x}_t, \mathbf{z}, \mathbf{h}_{t-1}; \boldsymbol{\theta}_{\mathbf{h}}), \quad \mathbf{x}_0 = \mathbf{0}, \quad \mathbf{h}_{-1} = \mathbf{g} \quad (3.24)$$

$$\mathbf{y}_t = f_{\mathbf{y}}(\mathbf{h}_t; \boldsymbol{\theta}_{\mathbf{y}}) \quad (3.25)$$

By default,  $\mathbf{x}_0 = \mathbf{0}$ . We can think of this initialization vector as an empty canvas, an implicit part of any stroke (and later, any composition). Following a similar development as in Eq.3.7, our maximization function can be expressed as follows (parameters  $\boldsymbol{\theta}$  and  $\boldsymbol{\phi}$  are omitted on the r.h.s. for legibility):

$$\begin{aligned} \mathcal{L}(\mathbf{x}_{1:T}, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{y}_{t-1}(\mathbf{h}_{t-1}(\mathbf{x}_{1:t-1}, \boldsymbol{\epsilon}, \mathbf{y}^q(\mathbf{h}_T^q(\mathbf{x}_{1:T})))) \\ &\quad - D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}_{1:T}) \parallel p(\mathbf{z})) \end{aligned} \quad (3.26)$$

with  $K$  the dimensionality of  $\mathbf{z}$  (see Eq.3.19),

$$D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}_{1:T}) \parallel p(\mathbf{z})) = \frac{1}{2} \sum_{k=1}^K \sigma_{\phi_k}^2(\mathbf{x}_{1:T}) + \mu_{\phi_k}^2(\mathbf{x}_{1:T}) - 1 - \log(\sigma_{\phi_k}^2(\mathbf{x}_{1:T})) \quad (3.27)$$

#### Choice of $\log p(\mathbf{x}_t | \mathbf{y}_{t-1})$

The choice of  $\log p(\mathbf{x}_t | \mathbf{y}_{t-1})$  is critical, because it precisely shapes the capability of model  $P$ .  $\mathbf{x}_t$  is a random vector, so each of its components has to be materialized as such. As recalled at the beginning of this section,  $\mathbf{x}_t$  contains random vectors  $\boldsymbol{\Delta}_t = [\boldsymbol{\delta}_t, \boldsymbol{\delta}'_t, \boldsymbol{\delta}''_t]$ , which describe the movement of the pen from  $t-1$ . In addition, the model needs to signal when the generation of a stroke is ended. The model should be able to produce strokes of length 1, as well as strokes of length 8. We therefore introduce a new binary random variable  $\beta_t$ , controlling the state of the pen during the movement from  $t-1$  to  $t$ . If  $\beta_t = 1$ , the curve component is

visible, otherwise not, and the stroke is completed. In  $\mathbf{x}_t$ ,  $\beta_t$  will be materialized as a one-hot vector  $[\beta_{t,1}, \beta_{t,0}]$ . As a result:

$$\mathbf{x}_t = [\delta_{t,x}, \delta_{t,y}, \delta'_{t,x}, \delta'_{t,y}, \delta''_{t,x}, \delta''_{t,y}, \beta_{t,1}, \beta_{t,0}] \quad (3.28)$$

Before assigning a shape to  $\log p(\mathbf{x}_t | \mathbf{y}_{t-1})$ , we must verify the independence of the  $\mathbf{x}$  components. Pen movements  $\Delta_t$  do not carry any meaning when they are invisible i.e.  $\beta_t = 0$  or  $[\beta_{t,1}, \beta_{t,0}] = [0, 1]$ . Probabilities associated with pen movements are therefore conditioned on  $\beta_t$ , so that:

$$\log p(\mathbf{x}_t | \mathbf{y}_{t-1}) = \log p(\beta_t | \mathbf{y}_{t-1}) + \log p(\Delta_t | \mathbf{y}_{t-1}, \beta_t) \quad (3.29)$$

$p(\beta_t | \mathbf{y}_{t-1})$  is chosen to be a Bernoulli distribution with  $j_{\theta_t}$  probabilities. However, the model outputs raw logits, i.e. unnormalized log probabilities  $\widetilde{\log(j_{\theta_t})}$ , which have to be normalized with a softmax function.

$$\log j_{\theta_t,i} = \log \frac{\exp \widetilde{\log(j_{\theta_t,i})}}{\sum_{k=0}^1 \exp \widetilde{\log(j_{\theta_t,k})}} = \widetilde{\log(j_{\theta_t,i})} - \log \sum_{k=0}^1 \exp \widetilde{\log(j_{\theta_t,k})} \quad (3.30)$$

then,

$$\log p(\beta_t | \mathbf{y}_{t-1}) = \log \prod_{i=0}^1 j_{\theta_t,i}^{\beta_{t,i}} = \sum_{i=0}^1 \beta_{t,i} \log j_{\theta_t,i} \quad (3.31)$$

When  $\beta_t = 0$ , the stroke is ended, and we set  $\Delta_t = \mathbf{0}$ . Consequently,  $p(\Delta_t | \mathbf{y}_{t-1}, \beta_{t,0}) = 1$ . Otherwise,  $p(\Delta_t | \mathbf{y}_{t-1}, \beta_{t,1})$  is chosen to be a multivariate normal distribution with a diagonal covariance matrix. We set  $C = 6$  to be the dimensionality of  $\Delta_t$ .

$$\begin{aligned} \log p(\Delta_t | \mathbf{y}_{t-1}, \beta_{t,1}) &= \log \mathcal{N}(\Delta_t | \boldsymbol{\mu}_\theta(\mathbf{y}_{t-1}), \text{diag}(\boldsymbol{\sigma}_\theta^2(\mathbf{y}_{t-1}))) \\ &= \log \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^C \frac{(\Delta_{t,i} - \mu_{\theta i})^2}{\sigma_{\theta i}^2}\right)}{\sqrt{(2\pi)^C \prod_{i=1}^C \sigma_{\theta i}^2}} \\ &= -\frac{1}{2} \sum_{i=1}^C \left( \frac{(\Delta_{t,i} - \mu_{\theta i})^2}{\sigma_{\theta i}^2} \right) - \frac{1}{2} \log(2\pi)^C - \frac{1}{2} \log \prod_{i=1}^C \sigma_{\theta i}^2 \\ &= -\frac{1}{2} \sum_{i=1}^C \frac{(\Delta_{t,i} - \mu_{\theta i})^2}{\sigma_{\theta i}^2} + \log \sigma_{\theta i}^2 + \log 2\pi \end{aligned} \quad (3.32)$$

In the original RNN-VAE implementation for simple sketches<sup>29</sup>,  $p(\Delta_t | \mathbf{y}_{t-1}, \beta_{t,1})$  is chosen to be a mixture of multivariate normal distributions instead of a single multivariate normal distribution. This additional source of stochasticity in the model undoubtedly confers greater capability, and thus *accuracy*, however in practice we found that it acts primarily as a source of confusion. Even if our

<sup>29</sup>Clanuwat et al., 2018; Ha and Eck, 2017.

### 3 Model implementation

stroke model outputs are probabilistic (i.e. target points for the next portion of the curve are spread over a 2-d normal distribution), the variance/spread of these areas reflects uncertainty for a given  $\mathbf{z}$  at this time step, rather than different possible stroke endings. Theoretically, using a mixture distribution would confer the ability for a given  $\mathbf{z}$  to sprout different stroke alternatives, but we want  $\mathbf{z}$  to be a compressed and non-ambiguous version of  $\mathbf{x}$ . In fact, despite adopting probabilistic outputs, we expect our model to become nearly deterministic during training, i.e. we want the variance to be as small as possible. In Section.3.5, we will detail this strategy and several other training tricks.

## 3.3 Composition model

Compositions are sequences of  $T$  strokes, and they will also be symbolized by random vectors  $\mathbf{x}_{1:T}$  for simplicity. In addition, compositions do not have a specific order. Original compositions particularly exist in the dataset in multiple arrangements (see Subsection.2.3.Composition permutations). However, one objective of the composition model is to project every visually similar composition to the same location into the latent space. To achieve this goal with a regular RNN-VAE architecture, we can constrain the decoder  $P$  to reconstruct a given composition in an *absolute* order, no matter the input sequence. So, let us define a target sequence  $\mathbf{u}_{1:T}$  that corresponds to any unordered version  $\mathbf{x}_{1:T}$  of it. This *absolute* ordering is completely arbitrary, and we have chosen to sort strokes by decreasing linear length (length along the path).<sup>30</sup> Strokes that are too long and that are split during dataset formatting (see Subsection.2.3.Splitting of long strokes) are kept continuous, and reordered as a group.

Our composition model is primarily designed to implement the hyper-compositional object, addressing complete compositions only (see Section.1.3). However, we should anticipate the requirements of the model dedicated to the compositional plane (see Section.3.4). This model specifically uses a *mental image*, a target  $\mathbf{z}$  to constrain interrelations between strokes, i.e. conditional probabilities on the plane (see particularly Subsection.1.3.Probabilistic plane for details). To this end, the compositional objective of the artist may only be vaguely defined. One strategy to implement this idea is to encode  $\mathbf{z}$  from partial compositions, i.e. the first few strokes the artist already has in mind, or he/she has already drawn on the canvas. As a result, we define a secondary goal for our current composition model, i.e. the ability to project into latent space any incomplete version of  $\mathbf{x}$  to the same location

---

<sup>30</sup>An alternative ordering logic for strokes could have involved their starting position: going from top to bottom, and secondarily from left to right. This rule makes sense for kanji/hànzì, as it is close to their natural order. However, early experimental results showed that the decoder  $P$  is more accurate for the first strokes. An absolute order based on *visual importance*, i.e. linear length of strokes, is therefore more efficient.

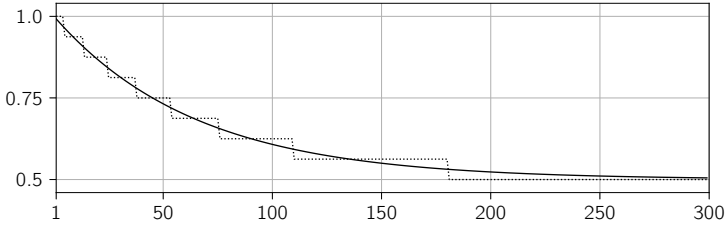


Figure 3.4: Conditioning length ratio, exponential decay. The solid line corresponds to the function of Eq.3.33 with  $C_{min} = 0.5$  and  $epoch_{min} = 300$ . The dotted line is the actual discrete ratio for a composition of length 16.

as the complete one. For this purpose, we define  $C \in [1, T]$ , a conditioning length on  $\mathbf{x}$ , so that inputs become  $\mathbf{x}_{1:C}$  ( $\mathbf{u}_{1:T}$  remains unchanged).

### Conditioning length, $C$

Conditioning length  $C$  can be considered as a data-augmentation technique because it adds some randomness to the input. At training,  $C$  is sampled from a discrete uniform distribution. Ideally, it should be in the range  $[1, T]$ , but such variability produces unreliable inputs, pushing the model to bypass the encoder  $Q$ . Therefore, we must specify a lower bound for sampling  $C$ . This minimal value  $C_{min}$  is defined as a ratio of  $T$ . In practice, we have chosen  $C_{min} = 0.5$ , so that the model receives at least half of the strokes. Of course, the model can still make predictions for shorter sequences, but at the price of an increased uncertainty on  $\mathbf{z}$  encodings. Nonetheless, even with the use of  $C_{min}$ , the encoder (and the whole model) faces difficulties during training with a low rate of improvement. Empirically, we found it more efficient to slowly decrease the lower bound ratio  $C_{ratio}$ , from 1 to  $C_{min}$ . This decay must not be too fast, nor too slow; when too fast, this extra-procedure is not helpful; when too slow, the network learns a representation too optimized for the current state of  $C_{ratio}$ . As the problem keeps evolving, the model suffers from a lack of plasticity. As a result, we have defined the lower bound ratio as an exponential decay, driven by the current  $epoch$  value (see Fig.3.4).

$$C_{ratio} = C_{min} + (1 - C_{min}) \exp\left(-epoch \frac{\log 100}{epoch_{min}}\right) \quad (3.33)$$

where  $epoch_{min}$  is the target epoch when  $C_{ratio}$  approximately reaches  $C_{min}$  (i.e.  $C_{ratio} = C_{min} + 0.01$ ). We used  $epoch_{min} = 300$ , which showed good results.  $C$  can only be an integer so, because of rounding effects,  $C_{min}$  is actually reached before  $epoch_{min}$ , 181 in our case (see dotted line in Fig.3.4).

### 3 Model implementation

#### Architecture

In Fig.3.5, we see that the architecture is similar to the stroke model. The two exceptions are on the input, limited to  $\mathbf{x}_{1:C}$ , and on the output, replaced by  $\mathbf{u}_{1:T}$ .

Concerning the encoder  $Q$ , one architectural point is not illustrated in Fig.3.5. The LSTM unit  $\mathbf{h}^q$  is actually bidirectional<sup>31</sup>. Basically, this means that  $\mathbf{h}_C^q$  is the concatenation of two random vectors  $\mathbf{h}_C^{q\rightarrow}$  and  $\mathbf{h}_1^{q\leftarrow}$ . For  $\mathbf{h}_t^{q\rightarrow}$ , time flows normally from 1 to  $C$  and we keep the last hidden state  $C$ . On the other hand, for  $\mathbf{h}_t^{q\leftarrow}$ , time is reversed, i.e.  $t = C : 1$  and the last hidden state is 1. We can view this procedure as a way to emphasize the unordered nature of  $\mathbf{x}$ . In practice, bidirectional LSTM units generally provide better performance, and it has been positively validated for the composition model. This architectural detail is also used in the stroke model, but we omitted it from our earlier description because it is essentially a training trick (it is transparent from an implementation point of view, and it is easily available in PyTorch).

#### Optimization function

As for the stroke model,  $P$  and  $Q$  are implemented with neural networks through the following deterministic functions:

$$\mathbf{h}_t^q = f_{\mathbf{h}^q}(\mathbf{x}_t, \mathbf{h}_{t-1}^q; \phi_{\mathbf{h}}), \mathbf{h}_0^q = \mathbf{0} \quad (3.34)$$

$$\mathbf{y}^q = f_{\mathbf{y}^q}(\mathbf{h}_C^q; \phi_{\mathbf{y}}) \quad (3.35)$$

$$\mathbf{g} = f_{\mathbf{g}}(\mathbf{z}; \theta_{\mathbf{g}}) \quad (3.36)$$

$$\mathbf{h}_t = f_{\mathbf{h}}(\mathbf{u}_t, \mathbf{z}, \mathbf{h}_{t-1}; \theta_{\mathbf{h}}), \mathbf{u}_0 = \mathbf{0}, \mathbf{h}_{-1} = \mathbf{g} \quad (3.37)$$

$$\mathbf{y}_t = f_{\mathbf{y}}(\mathbf{h}_t; \theta_{\mathbf{y}}) \quad (3.38)$$

Following a similar development to that adopted for Eq.3.7, our maximization function can be expressed as follows (parameters  $\theta$  and  $\phi$  are omitted on the r.h.s. for legibility):

$$\mathcal{L}(\mathbf{x}_{1:T}, \theta, \phi) = \sum_{t=1}^T \log p(\mathbf{u}_t \mid \mathbf{y}_{t-1}(\mathbf{h}_{t-1}(\mathbf{u}_{1:t-1}, \epsilon, \mathbf{y}^q(\mathbf{h}_C^q(\mathbf{x}_{1:C})))))) - D_{\text{KL}}(q(\mathbf{z} \mid \mathbf{x}_{1:C}) \parallel p(\mathbf{z})) \quad (3.39)$$

Even if only  $\mathbf{x}_{1:C}$  is used by the encoder  $Q$ ,  $\mathcal{L}$  is defined on the l.h.s. over  $\mathbf{x}_{1:T}$  because  $\mathbf{x}_{C+1:T}$  is still required to produce  $\mathbf{u}_{1:T}$ . In addition,  $\mathbf{u}_{1:T}$  does not appear in  $\mathcal{L}$ , as it is a reformulation of  $\mathbf{x}_{1:T}$ .

<sup>31</sup>Ha and Eck, 2017; Schuster and Paliwal, 1997.

### 3.3 Composition model

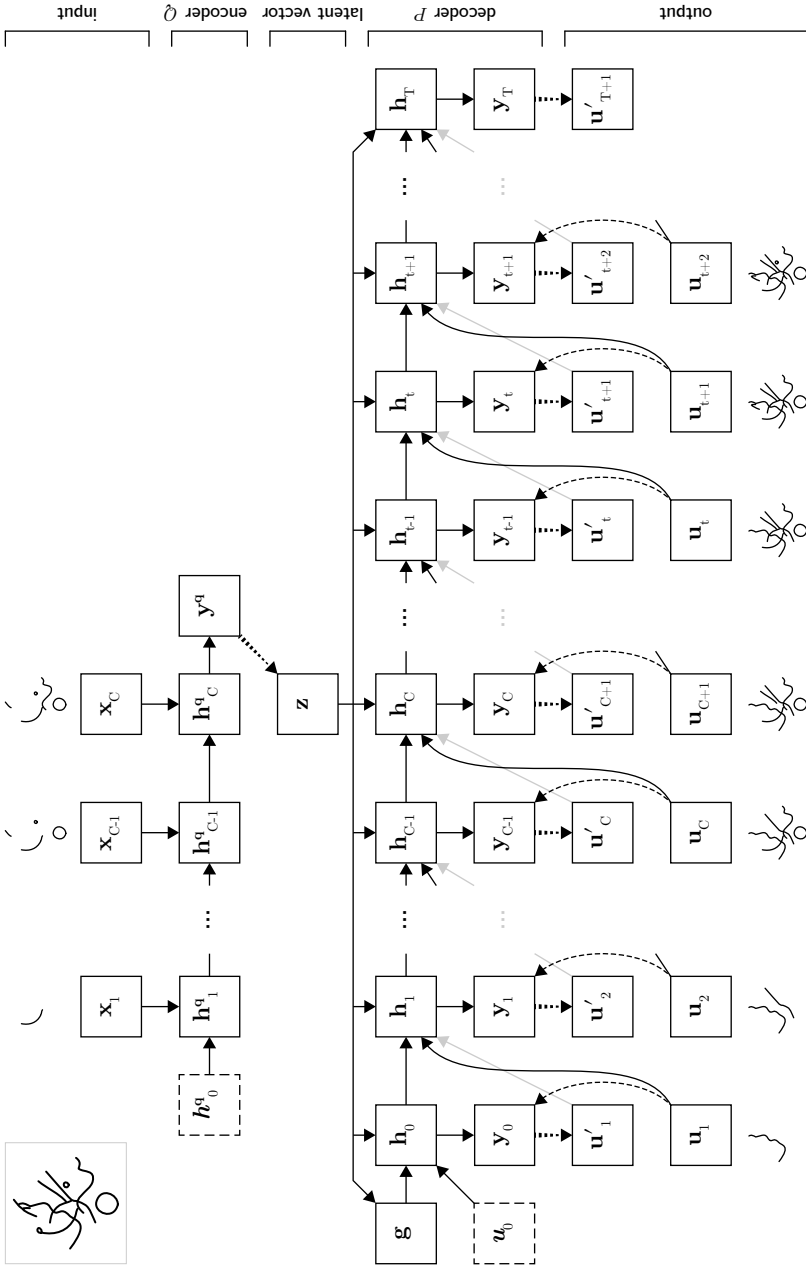


Figure 3.5: Composition model architecture. Legend available in Fig.3.3.

### 3 Model implementation

#### Choice of $\log p(\mathbf{u}_t \mid \mathbf{y}_{t-1})$

$\mathbf{u}_t$  and  $\mathbf{x}_t$  are random vectors describing a stroke. The first required piece of information is therefore represented by the initial position  $\mathbf{p}_t$  of the stroke (termed  $\mathbf{p}_{0,0}$  in Subsection.2.3.Composition encoding details). The shape of the stroke is then defined by the latent vector  $\mathbf{z}_{stroke}$  greedily encoded by the stroke model (i.e.  $\boldsymbol{\mu}_\phi$ ). For legibility, we represent this fixed-length random vector by  $\mathbf{s}_t$ . Finally, the random variable  $\beta_t$  is controlling the pen state. In similar fashion to the stroke model, when  $\beta_t = 0$ , no more stroke is visible and the composition is ended.

$$\mathbf{u}_t = \mathbf{x}_t = [\mathbf{p}_{t,x}, \mathbf{p}_{t,y}, \mathbf{s}_t, \beta_{t,1}, \beta_{t,0}] \quad (3.40)$$

Strokes do not have any meaning when  $\beta_t = 0$ , i.e.  $[\beta_{t,1}, \beta_{t,0}] = [0, 1]$ . Probabilities associated with  $\mathbf{p}_t$  and  $\mathbf{s}_t$  are thus conditioned on  $\beta_t$ , but mutually independent:

$$\log p(\mathbf{u}_t \mid \mathbf{y}_{t-1}) = \log p(\beta_t \mid \mathbf{y}_{t-1}) + \log p(\mathbf{p}_t \mid \mathbf{y}_{t-1}, \beta_t) + \log p(\mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_t) \quad (3.41)$$

$p(\beta_t \mid \mathbf{y}_{t-1})$  follows a Bernoulli distribution like in the stroke model (see Eq.3.31 for details). When  $\beta_t = 0$ , we set  $\mathbf{p}_t = \mathbf{s}_t = \mathbf{0}$ . Consequently,  $p(\mathbf{p}_t \mid \mathbf{y}_{t-1}, \beta_{t,0}) = p(\mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,0}) = 1$ . Otherwise,  $p(\mathbf{p}_t \mid \mathbf{y}_{t-1}, \beta_{t,1})$  and  $p(\mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1})$  are multivariate normal distributions with diagonal covariance matrices (see Eq.3.32).

Once again, the adoption of a mixture distribution to characterize outputs from the composition model could have been an option. Composition modeling is more complex than stroke modeling, and may require more capability. In addition, compared with strokes, it is likely that, at some point in the generative process, an artist is forced to choose among several options constrained by what is already present on the paper. A mixture distribution would precisely fit this idea. However, it is not the desired aspect of composition targeted by the current model. For now, we do not want to quantify the probabilities associated with alternatives offered at some time step. This task will be carried out by the next model (see below). The main purpose of the present model is to position compositions within a suitable space. We try to organize the set of compositions as a whole, as a hyper-compositional object. Stochasticity must therefore be associated with the encoding of  $\mathbf{z}$  by  $Q$  only. Depending on the *completeness* of this composition, the encoder projects it to a specific location of the latent space with more or less uncertainty. As soon as some  $\mathbf{z}$  is sampled, the reconstruction should be as deterministic as possible. A single multivariate normal distribution is thus sufficient, and we expect that during training the output variance becomes negligible.

### 3.4 Compositional plane model

The goal of this model is to study and predict variabilities at the scale of an individual drawing, at the scale of the compositional plane. We now consider

that an artist can face multiple alternatives at an instant  $t$  of the compositional process. The set of possible strokes that may be enacted to complete the drawing are naturally conditioned on previous choices, i.e. existing strokes on the canvas. Additionally, they are constrained by a target location in the space of regularities learned at the scale of his/her complete works, the hyper-composition (see Subsection.1.3.Probabilistic plane). This target *mental image* is represented by the latent random variable  $\mathbf{z}$  encoded by the composition model depicted in the previous section. The compositional plane model is thus a complementary model primarily designed to provide a spatial metric of individual stroke probabilities (experimental tools will be detailed in Section.4.3).

#### Architecture

In Fig.3.6, encoder  $Q$  is grayed out because it is reused from the composition model. At training, its parameter  $\phi$  is fixed, and do not need to be optimized. Input sequences  $\mathbf{x}$  are again possibly incomplete, with a conditioning length  $C \in [1, T]$ .  $Q$  is already used to address incomplete compositions, so an exponential decay of  $C_{ratio}$  from 1 to  $C_{min}$  is not necessary.  $C_{ratio}$  is fixed to  $C_{min} = 0.5$ . Incomplete compositions typically produce more uncertain outputs from  $Q$  than complete ones, and  $\mathbf{z}$  is not greedily sampled to reflect this information. It can be thought as the uncertainty of the artist about his/her compositional objective. In practice, we reverse this scheme, and force the model to learn the reconstruction of a specific final composition  $\mathbf{x}$  from a wide range of *mental images*  $\mathbf{z}$  (8 in our case). We also notice that absolute sequences  $\mathbf{u}$  are no longer required as an optimization target because compositions must be created and explored in any order. In addition, once a  $\mathbf{z}$  is chosen, *learned* stochasticity can only be associated with model outputs. We therefore assign more generative power to  $\mathbf{y}$  with a mixture distribution. We can also drop the denomination of *decoder* for  $P$  and simply use *model*  $P$ .

#### Optimization function

Encoder  $Q$  is frozen, so the model  $P$  only requires the following new deterministic functions:

$$\mathbf{g} = f_{\mathbf{g}}(\mathbf{z}; \theta_{\mathbf{g}}) \quad (3.42)$$

$$\mathbf{h}_t = f_{\mathbf{h}}(\mathbf{x}_t, \mathbf{z}, \mathbf{h}_{t-1}; \theta_{\mathbf{h}}), \mathbf{x}_0 = \mathbf{0}, \mathbf{h}_{-1} = \mathbf{g} \quad (3.43)$$

$$\mathbf{y}_t = f_{\mathbf{y}}(\mathbf{h}_t; \theta_{\mathbf{y}}) \quad (3.44)$$

Following a similar development to that adopted for Eq.3.7, the maximization function can be expressed as follows ( $\theta$  is omitted on the r.h.s. for legibility):

$$\mathcal{L}(\mathbf{x}_{1:T}, \theta) = \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{y}_{t-1}(\mathbf{h}_{t-1}(\mathbf{x}_{1:t-1}, \mathbf{z}))) \quad (3.45)$$



### 3 Model implementation

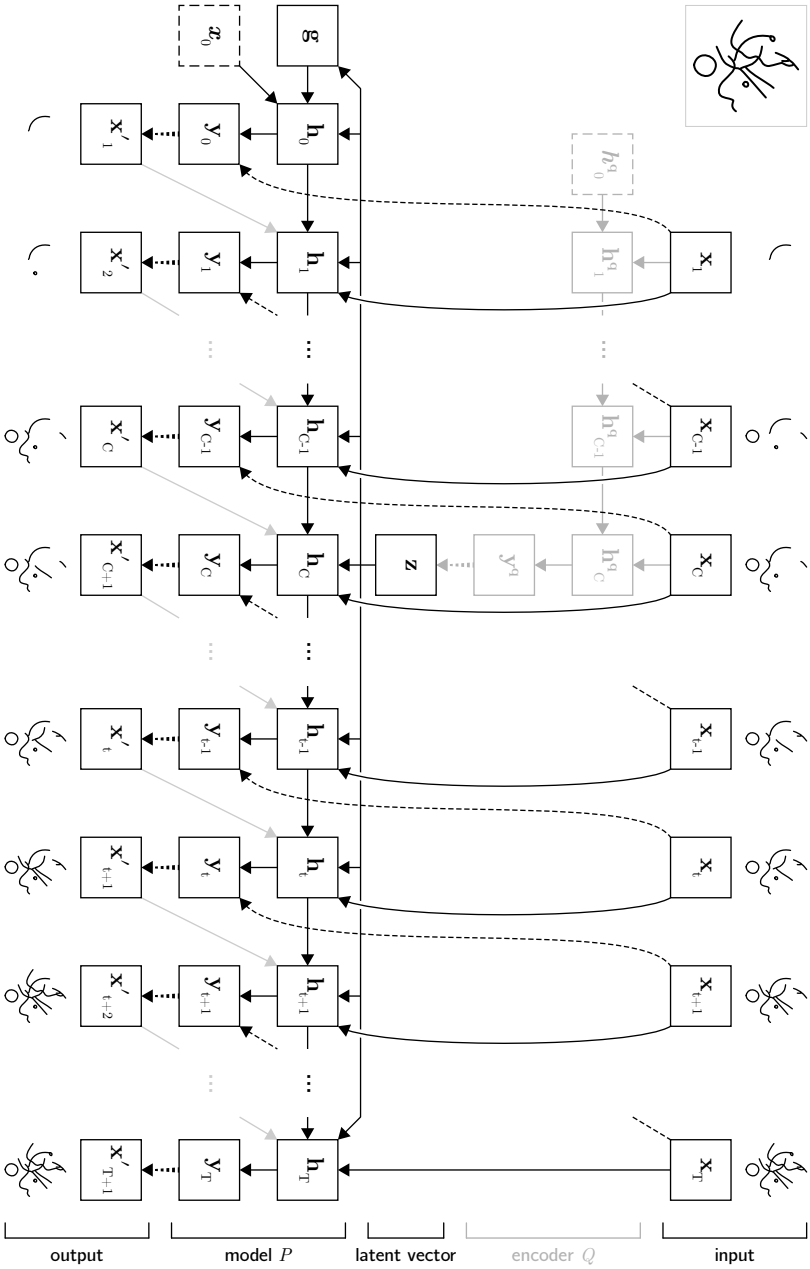


Figure 3.6: Compositional plane model architecture. Legend available in Fig.3.3.

**Choice of  $\log p(\mathbf{x}_t \mid \mathbf{y}_{t-1})$**

Similarly to the definition of  $\mathbf{x}_t$  given in the composition model,  $\mathbf{x}_t = [\mathbf{p}_{t,x}, \mathbf{p}_{t,y}, \mathbf{s}_t, \beta_{t,1}, \beta_{t,0}]$  (see details in Subsection.3.3.Choice of  $\log p(\mathbf{u}_t \mid \mathbf{y}_{t-1})$ ). In addition,  $\mathbf{p}_t$  and  $\mathbf{s}_t$  are kept conditioned on  $\beta_t$ , so that:

$$\log p(\mathbf{x}_t \mid \mathbf{y}_{t-1}) = \log p(\beta_t \mid \mathbf{y}_{t-1}) + \log p(\mathbf{p}_t, \mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_t) \quad (3.46)$$

$p(\beta_t \mid \mathbf{y}_{t-1})$  is still chosen to follow a Bernoulli distribution (see Eq.3.31 for details). When  $\beta_t = 0$ ,  $\mathbf{p}_t = \mathbf{s}_t = \mathbf{0}$  and  $p(\mathbf{p}_t, \mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,0}) = 1$ .

Otherwise,  $p(\mathbf{p}_t, \mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1})$  is now a mixture distribution. At each time step, the artist is confronted with multiple alternatives. In addition, if strokes of shape  $a$  and  $b$  are possible, each of them is expected at its dedicated position; shape  $a$  at location  $a$ , shape  $b$  at location  $b$ , and not shape  $a$  at location  $b$ . As a result,  $\mathbf{p}_t$  and  $\mathbf{s}_t$  are not directly independent. They are independent, only given an *alternative* index. We therefore introduce a random variable  $m$  following a categorical distribution. Naturally,  $\sum_{m=1}^M p(m) = 1$ , with  $M$  the number of alternatives, i.e. the number of mixtures (in practice,  $M = 8$ ). Then,  $p(\mathbf{p}_t, \mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1})$  can then be considered as the marginal distribution against  $m$ .

$$\begin{aligned} \log p(\mathbf{p}_t, \mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}) &= \log \sum_{m=1}^M p(m) p(\mathbf{p}_t, \mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m) \\ &= \log \sum_{m=1}^M p(m) p(\mathbf{p}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m) p(\mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m) \\ &= \log \sum_{m=1}^M \exp \left( \log p(m) + \log p(\mathbf{p}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m) \right. \\ &\quad \left. + \log p(\mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m) \right) \end{aligned} \quad (3.47)$$

$p(\mathbf{p}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m)$  and  $p(\mathbf{s}_t \mid \mathbf{y}_{t-1}, \beta_{t,1}, m)$  are finally chosen to be multivariate normal distributions with diagonal covariance matrices (see Eq.3.32).

### 3.5 Practical model training

Previous sections of this chapter have covered theoretical designs of stroke and composition models. These models need to be trained, and this aspect of model development represents a whole different problem in itself: even if a neural network architecture is ideal for solving a specific problem, small details of how training is implemented can dramatically change the results. Some of these training tricks are demonstrably efficient through experimentation. However, they remain poorly justified in the associated publications and/or they are difficult to

### 3 Model implementation

generalize to different datasets. In addition to these difficulties, the accumulation of hyperparameters represents another important issue that is caused by the multiplication of heuristics, hacks, and regularizers. These parameters are termed *hyper* because they are tweaked outside the main training procedure. They should be adjusted until an optimal combination is found. Such procedure is time-consuming, so we try to limit their number, and to give them tangible meanings when necessary. The other important goal is to obtain a continuous and expressive latent space, i.e. with qualitative homogeneity. How to judge and control this aspect objectively is a particularly difficult question. This section describes our efforts to ease the training protocol, and to better understand and justify the insights provided by our models.

#### *Disentanglement of latent space*

The procedure of disentangling the latent space involves obtaining a space of  $\mathbf{z}$  that decomposes  $\mathbf{x}$  into its *true* independent components. For instance, if  $\mathbf{x}$  consisted of centered pictures of colored dices, the natural components would be dice rotation angles and RGB channel intensities. These 6 independent dimensions would be sufficient to describe any  $\mathbf{x}$ . But what if  $\mathbf{x}$  cannot be trivially decomposed into adequate components? How do we decide whether our model has achieved an acceptable level of disentanglement?

The commonly assumed notion of disentanglement is quite restrictive for complex models where the true generative factors are not independent, very large in number, or where it cannot be reasonably assumed that there is a well-defined set of *true* generative factors [...]. To this end, we introduce a generalization of disentanglement, decomposition, which at a high-level can be thought of as imposing a desired structure on the learned representations. [...] We characterize the decomposition of latent spaces in VAEs to be the fulfillment of two factors:

- a.** An *appropriate* level of overlap in the latent space, ensuring that the range of latent values capable of encoding a particular data point is neither too small, nor too large. [...]
- b.** The aggregate encoding  $q(\mathbf{z})$  matching the prior  $p(\mathbf{z})$ , where the latter expresses the desired dependency structure between latents.<sup>32</sup>

This extended definition of disentanglement is interesting because it translates a fuzzy objective into two concrete sub-goals. However, in order to incorporate this definition into our framework, we must link points **a.** and **b.** to our maximization function. To do so, we will investigate  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$  under training conditions with  $\mathbf{x} \sim p_{\mathcal{D}}$ , i.e. all  $\mathbf{x}_n$  from  $\mathcal{D}$ , and transformed into a minimization objective. Parameters  $\boldsymbol{\theta}$  and  $\boldsymbol{\phi}$  will now be omitted for legibility on the r.h.s. of equations. In the most general form, we have:

---

<sup>32</sup>Mathieu et al., 2019.

$$\begin{aligned}
 -\mathbb{E}_{\mathbf{x} \sim p_D} \mathcal{L}(\mathbf{x}_n, \boldsymbol{\theta}, \phi) &= -\frac{1}{N} \sum_{n=1}^N \sum_{t=1}^{T_n} \log p(\mathbf{x}_{n,t} | \mathbf{x}_{n,1:t-1}, \mathbf{z}) \quad \left. \vphantom{-\frac{1}{N} \sum_{n=1}^N \sum_{t=1}^{T_n} \log p(\mathbf{x}_{n,t} | \mathbf{x}_{n,1:t-1}, \mathbf{z})} \right\} \mathcal{L}_{\mathbf{x}} \\
 &\quad + \frac{1}{N} \sum_{n=1}^N D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}_{n,1:T_n}) \| p(\mathbf{z})) \quad \left. \vphantom{\frac{1}{N} \sum_{n=1}^N D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}_{n,1:T_n}) \| p(\mathbf{z}))} \right\} \mathcal{L}_{D_{\text{KL}}}
 \end{aligned} \tag{3.48}$$

We notice that  $T$ , the abstract maximum length of  $\mathbf{x}$ , must be separately specified for each  $\mathbf{x}_n$  as  $T_n$ . For ease of following developments, we have also split the minimization objective into sub-terms  $\mathcal{L}_{\mathbf{x}}$  and  $\mathcal{L}_{D_{\text{KL}}}$ .

### Paradoxical $\mathcal{L}_{D_{\text{KL}}}$

First, there is at training a conflict between the two terms of  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi)$ . Basically,  $\mathcal{L}_{\mathbf{x}}$  enforces a good reconstruction of the samples, while  $\mathcal{L}_{D_{\text{KL}}}$  is supposed to guarantee a good generative model, i.e. to force the latent space  $\mathbf{z}$  to follow the chosen prior distribution. However, when  $\mathcal{L}_{\mathbf{x}}$  becomes efficient, we usually observe a negative impact on  $\mathcal{L}_{D_{\text{KL}}}$ . In some situations, the opposite happens. So, it appears impossible to minimize both  $\mathcal{L}_{\mathbf{x}}$  and  $\mathcal{L}_{D_{\text{KL}}}$  simultaneously. Secondly, if  $\mathcal{L}_{D_{\text{KL}}}$  were exactly 0, the encoder  $Q$  would be ignored by the decoder  $P$ , rendering the training procedure partially pointless. In order to better understand the paradoxical behavior of  $\mathcal{L}_{D_{\text{KL}}}$ , we will decompose this term<sup>33</sup>.

First, we make the subscript  $n$ , representing the index of inputs  $\mathbf{x}_{n,1:T_n}$ , as an explicit random variable  $\mathfrak{n}$ . For ease of notation, we will omit time subscripts. By definition, we have  $p(\mathfrak{n}) = q(\mathfrak{n}) = \frac{1}{N}$  and  $q(\mathbf{z} | \mathbf{x}_n) = q(\mathbf{z} | \mathfrak{n})$ . Then, we can think of  $q(\mathbf{z})$  as marginalizing against  $\mathfrak{n}$ , so that  $q(\mathbf{z}) = \sum_{n=1}^N q(\mathbf{z}, n)$ . As a result:

$$\begin{aligned}
 \mathcal{L}_{D_{\text{KL}}} &= \frac{1}{N} \sum_{n=1}^N \int_{\mathbf{z}} q(\mathbf{z} | \mathbf{x}_n) \log \frac{q(\mathbf{z} | \mathbf{x}_n)}{p(\mathbf{z})} d\mathbf{z} \\
 &= \int_{\mathbf{z}} \frac{1}{N} \sum_{n=1}^N q(\mathbf{z} | \mathfrak{n}) \log \frac{q(\mathbf{z} | \mathfrak{n})}{p(\mathbf{z})} d\mathbf{z} \\
 &= \int_{\mathbf{z}} \frac{1}{N} \sum_{n=1}^N \frac{q(\mathbf{z}, n)}{q(\mathfrak{n})} \log \frac{q(\mathbf{z})q(\mathbf{z} | \mathfrak{n})}{q(\mathbf{z})p(\mathbf{z})} d\mathbf{z} \\
 &= \int_{\mathbf{z}} \log \left( \frac{q(\mathbf{z})}{p(\mathbf{z})} \right) \sum_{n=1}^N q(\mathbf{z}, n) d\mathbf{z} + \int_{\mathbf{z}} \sum_{n=1}^N q(\mathbf{z}, n) \log \frac{q(\mathbf{z} | \mathfrak{n})}{q(\mathbf{z})} d\mathbf{z} \\
 &= \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} + \int_{\mathbf{z}} \sum_{n=1}^N q(\mathbf{z}, n) \log \frac{q(\mathbf{z}, n)}{q(\mathbf{z})q(\mathfrak{n})} d\mathbf{z} \\
 &= D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z})) + \mathbb{I}_{q(\mathbf{z}, \mathfrak{n})}[\mathbf{z}, \mathfrak{n}]
 \end{aligned} \tag{3.49}$$

<sup>33</sup>Demonstration is adapted from Hoffman and Johnson, 2016.

### 3 Model implementation

where  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]$  is the mutual information between an index  $n$  and its corresponding code  $\mathbf{z}$ . In other words, it indicates how much uncertainty can be eliminated about one of the random variable, knowing the other. In our case, it indicates the extent to which  $\mathbf{z}$  is deterministically defined by  $\mathbf{n}$ , i.e. any known  $\mathbf{x}_n$ . If encoding by  $Q$  is fully deterministic,  $\mathbf{z}$  and  $\mathbf{n}$  can be used indifferently and  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}] = -\sum_{n=1}^N q(n) \log q(n) = \mathbb{H}_{q(\mathbf{n})}[\mathbf{n}]$ . In this scenario, the maximal information carried by  $\mathbf{n}$  corresponds to its entropy,  $\log N$ . In the opposite scenario when  $\mathbf{z}$  and  $\mathbf{n}$  are independent,  $Q$  is completely ignored because uninformative, and mutual information is 0. Mutual information is therefore bounded by  $0 \leq \mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}] \leq \log N$ .

As a result,  $\mathcal{L}_{D_{\text{KL}}}$  combines two independent elements. Interestingly, these elements correspond to the disentanglement objectives expressed earlier: **a.** “an *appropriate* level of overlap in the latent space” is guaranteed if the mutual information  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]$  lies somewhere between its two bounds (at least distinct from 0). An appropriate level of ambiguity in the encoding by  $Q$  allows the latent space to smoothly map from one data point to another. **b.** “ $q(\mathbf{z})$  matching the prior  $p(\mathbf{z})$ ” is achieved if  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  becomes as small as possible. This is why  $\mathcal{L}_{D_{\text{KL}}}$  targets a paradoxical objective: the first part does not have to be 0, while the second part must be 0.

The second problem is that none of these terms are directly tractable at training. Indeed,  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  requires the marginalization of  $q(\mathbf{z} \mid \mathbf{n})$  over the whole dataset. It would be inefficient, if not impossible, to compute a gradient step with an entire dataset as input. It would require storage of large vectors exceeding GPU/CPU memory capacities. That is actually why all implemented training procedures are by mini-batch, i.e. randomly chosen subsets of  $\mathcal{D}$ .

Nevertheless,  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  can be computed at the end of each epoch (i.e. a training session over all mini-batches). Then, by difference with  $\mathcal{L}_{D_{\text{KL}}}$ , we can obtain a value of the mutual information  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]$ . So, despite being direct training tools, these two components are informative metrics for monitoring model behavior during optimization. For this purpose, we will express mutual information as a ratio  $\hat{\mathbb{I}}$ , independently of  $\mathcal{D}$  size:

$$\hat{\mathbb{I}} = \frac{\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]}{\log N} \quad (3.50)$$

As we cannot directly control  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]$  or  $\hat{\mathbb{I}}$ , our best strategy is to leverage the whole  $\mathcal{L}_{D_{\text{KL}}}$  as a proxy for  $\hat{\mathbb{I}}$ , while finding some additional regularizers on  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  that are accessible at mini-batch level.<sup>34</sup>

<sup>34</sup>This general idea is supported by Mathieu et al., 2019 and Kumar et al., 2018. Nonetheless, we would like to mention some other approaches.

T. Q. Chen et al., 2018 further split  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  into two terms: the total correlation  $D_{\text{KL}}(q(\mathbf{z}) \parallel \prod_{k=1}^K q(z_k))$ , and a dimension-wise  $D_{\text{KL}}(\sum_{k=1}^K D_{\text{KL}}(q(z_k) \parallel p(z_k)))$ . Scaling hyperparameters are then

Leveraging  $\mathcal{L}_{D_{\text{KL}}}$ 

Let us introduce a parameter  $\lambda$  to modulate the  $D_{\text{KL}}$  term of the ELBO<sup>35</sup>.

$$-\mathbb{E}_{\mathbf{x} \sim p_D} \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi) = \mathcal{L}_{\mathbf{x}} + \lambda \mathcal{L}_{D_{\text{KL}}} \quad (3.51)$$

In the literature, factor  $\lambda$  was first introduced to prevent the model from being stuck in a trivial initial state<sup>36</sup>.  $\mathcal{L}_{D_{\text{KL}}}$  can actually be made artificially close to zero.  $q(\mathbf{z} \mid \mathbf{x}_{n,1:T_n})$  can perfectly match  $p(\mathbf{z})$ , if  $\mathbf{z}$  is not forced to give meaningful information to the decoder  $P$ . In very ill-posed situations, most of the correcting gradient from  $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \phi)$  can be directed to  $\mathcal{L}_{D_{\text{KL}}}$ , which converge rapidly to zero. Then, both reconstruction and generative performance stay stationary to an extremely low level. The basic idea to overcome this issue is to anneal  $\mathcal{L}_{D_{\text{KL}}}$  during the first steps, and then, when the decoder  $P$  becomes efficient enough, to slowly set  $\lambda$  back to 1. Such training procedure is illustrated in Fig.3.7a.  $\lambda$  is usually chosen to follow a sigmoid-like shape over training steps. Another reported option<sup>37</sup> is to stop the optimization of  $\mathcal{L}_{D_{\text{KL}}}$  when it becomes too low. By clamping this term below a certain threshold, no gradient is then back-propagated to improve it. The downside of both approaches is the addition of yet another hyperparameter in the form of a temporal increasing rate or a clipping value.

A more efficient use of the  $\lambda$  parameter<sup>38</sup> is to leverage between a good reconstruction and an expressive latent space (see Fig.3.7b). In the extreme case, when  $\lambda = 0$ , the model actually becomes a simple auto-encoder with the best possible reconstruction accuracy, but no generative abilities.  $\lambda$  is then left as a

---

assigned to each component (including  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]$ ). As a result, it is possible to independently control each *contradictory* aspect of  $\mathcal{L}_{D_{\text{KL}}}$ . However, all these terms are still not accessible at mini-batch level, and the authors rely on biased estimates. For Mathieu et al., 2019, this approximation is not sufficiently qualitative, unless batch-size becomes ridiculously big.

With InfoVAE, S. Zhao et al., 2017 even proposed to completely exclude  $\mathbb{I}_{q(\mathbf{z}, \mathbf{n})}[\mathbf{z}, \mathbf{n}]$  from the optimization function. For these authors, it is not a problem if the encoder  $Q$  becomes fully deterministic, as long as  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z})) \rightarrow 0$ . In a dataset with a relative limited number of inputs (like ours), the lack of overlap between latents would be problematic for obtaining a smooth generative space. Besides, the authors use Maximum-Mean Discrepancy (MMD) with a kernel trick to minimize  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  at mini-batch level. This is an alternative metric to the Kullback-Leibler divergence and its asymmetric issues. We plan to investigate this tool in future work.

An alternative strategy, proposed by Rezende and Viola, 2018, involves constraining the model  $P$  with  $\mathcal{L}_{\mathbf{x}}$ . The authors argue that “in many practical cases it is much easier to decide on useful constraints in the data-domain, such as a desired reconstruction accuracy, rather than information constraints”. It is true that deciding an appropriate value for  $\hat{\mathbb{I}}$  is difficult and abstract, with no direct or qualitative visual impact on model outputs (despite extended investigations by Alemi et al., 2018). However, in our case, the expected accuracy of the data-space is particularly subject to artistic and subjective evaluations. Therefore, we have decided to keep a strategy that involves manipulating information within the model.

<sup>35</sup>We remark that, as soon as  $\lambda \neq 1$ , our optimization function deviates from the *true* ELBO. Nonetheless, it is of little concern if the general capacity of the model is improved.

<sup>36</sup>Bowman et al., 2016.

<sup>37</sup>In Ha and Eck, 2017, clamping of  $\mathcal{L}_{D_{\text{KL}}}$  is actually used together with initial annealing.

<sup>38</sup>Actually, the control parameter is denominated  $\beta$  in the  $\beta$ -VAE architecture from (Higgins et al., 2016; Higgins et al., 2017).

### 3 Model implementation

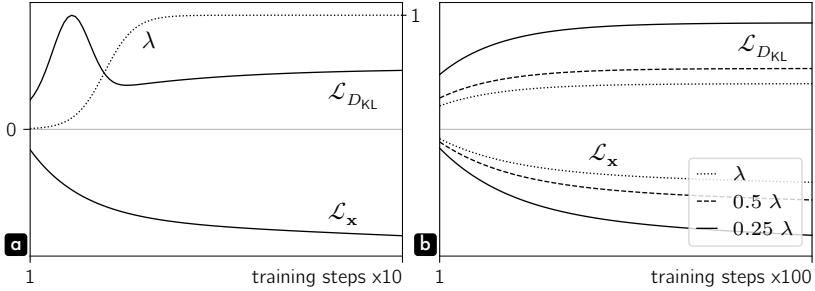


Figure 3.7: Illustration of  $\mathcal{L}_{\mathbf{x}}$  and  $\mathcal{L}_{D_{\text{KL}}}$  training errors, with the variation of  $\lambda$ . In panel **a**,  $\lambda$  follows a sigmoid-like function, annealing  $\mathcal{L}_{D_{\text{KL}}}$  during the initial training steps. Meanwhile,  $\mathcal{L}_{\mathbf{x}}$  is minimized with eased constraints. Panel **b** presents three optimization scenarios with different constant values of  $\lambda$ , and emphasizes the contradictory influences of the two ELBO terms.

hyperparameter, and determining its appropriate value is very difficult in practice. The desired tradeoff is a subjective appreciation, where  $\lambda$  can take extremely different values depending on the dataset (e.g. 1, 5, 20 or 250)<sup>39</sup>.

While prototyping experiments, we found that  $\lambda$  is actually very dependent on the dimensionalities involved in the network. A simple workaround is thus to normalize the different sources of contingent variability between models. On the one hand, we can normalize  $\mathcal{L}_{\mathbf{x}}$  by  $\hat{T} = \frac{1}{N} \sum_{n=1}^N T_n$ , the averaged length of  $T_n$  in the dataset. On the other hand,  $\mathcal{L}_{D_{\text{KL}}}$  can be divided by the dimensionality  $K$  of  $\mathbf{z}$ .

$$-\mathbb{E}_{\mathbf{x} \sim p_D} \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{\hat{T}} \mathcal{L}_{\mathbf{x}} + \frac{\lambda}{K} \mathcal{L}_{D_{\text{KL}}} \quad (3.52)$$

#### Adaptive constraint on $\mathcal{L}_{D_{\text{KL}}}$

In Eq.3.52, we have normalized the ELBO terms to reduce the variability of  $\lambda$  when we modify model parameters. However, it does not address the first scenario described above (trivial initial state), neither provides a comprehensive parameter to control  $\hat{\mathbb{I}}$ . Therefore, we propose a method to adaptively constrain  $\mathcal{L}_{D_{\text{KL}}}$  via a parameter  $\lambda'$ . In order to apply this constraint as uniformly as possible across all dimensions of  $\mathbf{z}$ , i.e. to encourage the balance of information across dimensions of the latent space,  $\lambda'$  is defined separately for each dimension. This redefinition of the term  $\mathcal{L}_{D_{\text{KL}}}$  also includes the normalizing value  $K$ , so that:

$$\mathcal{L}'_{D_{\text{KL}}} = \frac{1}{N K} \sum_{n=1}^N \sum_{k=1}^K \lambda'_k D_{\text{KL}}(q(z_k | \mathbf{x}_{n,1:T_n}) \parallel p(z_k)) \quad (3.53)$$

<sup>39</sup>Higgins et al., 2017.

with

$$\lambda'_k = \left\langle D_{\text{KL}}(q(z_k | \mathbf{x}_{n,1:T_n}) \parallel p(z_k)) \right\rangle^\gamma \quad (3.54)$$

The decorator  $\langle u \rangle$  represents the value of  $u$ , but detached from the optimization tree of the neural network. The gradient is not back-propagated through it. In other words,  $\left\langle D_{\text{KL}}(q(z_k | \mathbf{x}_{n,1:T_n}) \parallel p(z_k)) \right\rangle$  is no longer an optimizable parameter, but becomes a simple scalar with fixed numerical value. In addition, the strength of this adaptive constraint is controlled by the exponent  $\gamma \geq 0$ .

The logic behind this procedure is to add some nonlinearity to the  $D_{\text{KL}}$  values, by virtually bringing them to a power  $\geq 1$ . As a result, when  $D_{\text{KL}}$  values are small, the importance of  $\mathcal{L}'_{D_{\text{KL}}}$  is lowered in the minimization objective. The reconstruction part of the ELBO  $\mathcal{L}_{\mathbf{x}}$  takes the lead and forces the encoding of  $\mathbf{z}$  to be more deterministic. Reconstruction improves and the mutual information ratio  $\hat{\mathbb{I}} \rightarrow 1$ . As a consequence,  $D_{\text{KL}}$  and  $\mathcal{L}'_{D_{\text{KL}}}$  values increase in the ELBO. The gradient corrects the encoder to support better *overlap* within the latent space and  $\hat{\mathbb{I}} \rightarrow 0$ . These scenarios alternate and the training procedure finally reaches a steady state. We have explored stronger nonlinearities with higher  $\gamma$  values. They allow  $\mathcal{L}'_{D_{\text{KL}}}$  to asymptote more quickly, but do not improve the results. Therefore, we adopted a default value of  $\gamma = 1$  for all models.

Despite still requiring a  $\lambda$  hyperparameter, we believe that our method makes the training protocol easier. Compared with the original procedure,  $\lambda$  has a stronger effect on the value of the steady state of  $\mathcal{L}'_{D_{\text{KL}}}$ . Actually, it directly controls the final value of  $\hat{\mathbb{I}}$ . In addition, thanks to the adopted normalization of terms, the same  $\lambda$  value produces nearly equivalent  $\hat{\mathbb{I}}$  with different models. For instance, concerning our stroke and composition models, which have different  $K$  and  $\hat{T}$ , setting  $\lambda = 4$  guaranteed  $\hat{\mathbb{I}} \approx 0.14$  in both case. Nevertheless, this approach only works because we also regularize  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$ .

### Regularizing $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$

A regularizer is an additional function that is optimized together with the main maximization function. It does not have any natural justification like the ELBO development (see Section.3.1), but it empirically helps the back-propagated gradient to fulfill particular objectives. In our case, we introduce a regularizer  $D_{\text{Cov}}(q(\mathbf{z}), p(\mathbf{z}))$ , enforcing  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z})) \rightarrow 0$ . This regularizer is added to the optimization function with a scaling parameter  $\lambda_{D_{\text{Cov}}}$ .

$$-\mathbb{E}_{\mathbf{x} \sim p_D} \mathcal{L}(\mathbf{x}_n, \boldsymbol{\theta}, \phi) = \frac{1}{\hat{T}} \mathcal{L}_{\mathbf{x}} + \lambda \mathcal{L}'_{D_{\text{KL}}} + \lambda_{D_{\text{Cov}}} D_{\text{Cov}} \quad (3.55)$$

Despite multiple propositions<sup>40</sup>, the simplest and most robust way to push  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z})) \rightarrow 0$  at mini-batch level is to match the moments of the

<sup>40</sup>Among others: T. Q. Chen et al., 2018; S. Zhao et al., 2017



### 3 Model implementation

two distributions. The first moment of  $q(\mathbf{z})$  is  $\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  and must be optimized to  $\mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[\mathbf{z}] = \mathbf{0}$ . However,  $\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}$  is not accessible within a mini-batch. We cannot sample over the entire  $\mathbf{z}$ . So, let us introduce a sampling on  $\mathbf{x} \sim p_{\mathcal{D}}$  by the law of total expectation:

$$\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}] = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}_n)}[\mathbf{z}]] = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)] \quad (3.56)$$

Minimizing the element-wise mean squared error of  $\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)]$  would be a bad regularizer. It would essentially push every  $\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)$  towards 0, effectively centering  $q(\mathbf{z})$  but rendering  $\mathbf{z}$  completely uninformative, no matter the variance. A solution is therefore to look at higher-order moments<sup>41</sup>. Our regularizer focuses on the covariance of  $q(\mathbf{z})$  (second-order central moment). Using the law of total covariance:

$$\begin{aligned} \text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}] &= \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\text{Cov}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}_n)}[\mathbf{z}]] + \text{Cov}_{\mathbf{x} \sim p_{\mathcal{D}}}[\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}_n)}[\mathbf{z}]] \\ &= \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\text{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n))] + \text{Cov}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)] \end{aligned} \quad (3.57)$$

with

$$\text{Cov}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)] = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)^{\top}] - \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)]\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)]^{\top} \quad (3.58)$$

As  $\text{Cov}_{\mathbf{z} \sim p(\mathbf{z})}[\mathbf{z}] = \mathbf{I}$ , we also expect  $\text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  to converge to the identity matrix. We notice that off-diagonal values of  $\text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  are only caused by  $\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)$ . From Eq.3.58, we see that pushing off-diagonal values to 0 would reproduce the issue emphasized above in relation to optimization of the first moment, i.e. pushing every  $\boldsymbol{\mu}_{\phi}(\mathbf{x}_n) \rightarrow \mathbf{0}$ . Here, it can be avoided if the diagonal values of  $\text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  are enforced to be 1. But this goal is shared with the optimization of  $\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n)$ . In other words, if  $\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n) = \mathbf{1}$ , then the encoder  $Q$  becomes uninformative, even though  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z})) = 0$ .<sup>42</sup> Thanks to our  $\mathcal{L}'_{D_{\text{KL}}}$  optimization function,  $\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n)$  is in practice maintained to reasonable values, neither too small nor too high. Nonetheless, it is not pertinent to allow  $\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\text{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n))]$  to compensate for an incomplete decorrelation of  $\text{Cov}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)]$ . We therefore directly optimize  $\text{Cov}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)]$ , and let  $\mathcal{L}'_{D_{\text{KL}}}$  prevent  $\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n)$  from being  $\mathbf{0}$ . For instance, trained stroke and composition models converged with this method to a value of  $\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[\text{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x}_n))] \approx 0.05\mathbf{I}$ , which renders this term negligible in  $\text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  while still guaranteeing good overlap of the encoded  $\mathbf{x}_n$  within latent space.

Finally, we can define our regularizer  $D_{\text{Cov}}(q(\mathbf{z}), p(\mathbf{z}))$  as the element-wise mean squared error between the truncated  $\text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  and  $\text{Cov}_{\mathbf{z} \sim p(\mathbf{z})}[\mathbf{z}] = \mathbf{I}$ :

<sup>41</sup>A covariance regularizer is proposed and demonstrated in Kumar et al., 2018. The authors also raise the possibility of using third and higher moments. We hope to explore this approach in future work.

<sup>42</sup>This possibility is highlighted in Kumar et al., 2018. The authors advise the use of this raw definition of  $\text{Cov}_{\mathbf{z} \sim q(\mathbf{z})}[\mathbf{z}]$  in situations where the dimensionality of  $\mathbf{z}$  is higher than the *true* number of data components, because it does not force the latent space to use all its available dimensions. In our case, this scenario is very unlikely. Our stroke and composition latent spaces are over-constrained to retain good model interpretability and practical application to psychophysical experiments.

$$D_{\text{Cov}}(q(\mathbf{z}), p(\mathbf{z})) = \frac{1}{K^2} \sum_{k=1}^K \sum_{k'=1}^K \left( \text{Cov}_{\mathbf{x} \sim p_{\mathcal{D}}}[\boldsymbol{\mu}_{\phi}(\mathbf{x}_n)] - \mathbf{I} \right)_{kk'}^2 \quad (3.59)$$

In summary, this regularizer enforces decorrelation between the different dimensions of  $\mathbf{z}$  and a good match between  $q(\mathbf{z})$  and the prior  $p(\mathbf{z})$ , while being accessible within a mini-batch procedure. Nonetheless, its computation still requires a descent batch-size to provide a useful gradient. Even though batch-sizes between 2 and 32 have been reported to give consistent optimal results for multiple architectures<sup>43</sup>, we set our batch-size to 64 as a compromise. Finally,  $D_{\text{Cov}}(q(\mathbf{z}), p(\mathbf{z}))$  is in practice very small, so we set  $\lambda_{D_{\text{Cov}}} = 2 \times 10^3$ .

### Issue with the optimization of $p(\beta_t \mid \mathbf{x}_{1:t-1}, \mathbf{z})$

$\mathcal{L}'_{D_{\text{KL}}}$  and  $D_{\text{Cov}}(q(\mathbf{z}), p(\mathbf{z}))$  are the result of several necessary improvements on the original ELBO. However, the reconstruction term  $\mathcal{L}_{\mathbf{x}}$  also presents issues that need to be addressed. We first decompose  $\mathbf{x}_t$  into the pen state  $\beta_t$  and other dependent variables, generically indicated by  $\Psi_t$ . For instance,  $\Psi_t$  refers to  $\Delta_t$  for the stroke model, while being  $[\mathbf{p}_t, \mathbf{s}_t]$  for the composition models.

$$\begin{aligned} \frac{1}{\hat{T}} \mathcal{L}_{\mathbf{x}} &= -\frac{1}{N\hat{T}} \sum_{n=1}^N \sum_{t=1}^{T_n} \log p(\beta_{n,t} \mid \mathbf{x}_{n,1:t-1}, \mathbf{z}) \quad \left. \vphantom{\sum_{n=1}^N} \right\} \mathcal{L}_{\beta} \\ &\quad -\frac{1}{N\hat{T}} \sum_{n=1}^N \sum_{t=1}^{T_n} \log p(\Psi_{n,t} \mid \mathbf{x}_{n,1:t-1}, \mathbf{z}, \beta_{n,t}) \quad \left. \vphantom{\sum_{n=1}^N} \right\} \mathcal{L}_{\Psi} \end{aligned} \quad (3.60)$$

With  $t$  ranging from 1 to  $T_n$ ,  $\beta_{n,t}$  is always 1 for all dataset samples. In this context, the model  $P$  simply learns to output 1 without specifying when the generation of a stroke or a composition should be terminated. To overcome this issue,  $p(\beta_{n,t} \mid \mathbf{x}_{n,1:t-1}, \mathbf{z})$  must be evaluated after  $t = T_n$ . We define an upper bound  $T_{\text{max}}$ , being the maximum  $T_n$  within the dataset  $\mathcal{D}$  (or in the current mini-batch +1). Normalization of this term is then modified accordingly.

$$\mathcal{L}_{\beta} = -\frac{\lambda_{\beta}}{N\hat{T}_{\text{max}}} \sum_{n=1}^N \sum_{t=1}^{T_{\text{max}}} \log p(\beta_{n,t} \mid \mathbf{x}_{n,1:t-1}, \mathbf{z}) \quad (3.61)$$

At training, we observed that  $\mathcal{L}_{\beta}$  was small compared with  $\mathcal{L}_{\Psi}$ . To correct this discrepancy, a scalar  $\lambda_{\beta}$  has been added to  $\mathcal{L}_{\beta}$  and empirically set to 10.

### Balancing model resources for the optimization of $p(\Psi_t \mid \mathbf{x}_{1:t-1}, \mathbf{z}, \beta_t)$

$\mathcal{L}_{\Psi}$ , the second part of the reconstruction term concerning  $p(\Psi_t \mid \mathbf{x}_{1:t-1}, \mathbf{z}, \beta_t)$ , is normalized by  $N\hat{T}$ . In so doing, we implicitly assign to each sample instant the same

<sup>43</sup>Masters and Luschi, 2018.

### 3 Model implementation

weight within the optimization process. However, our stroke model is for instance quite unbalanced in terms of length  $T$  (see Fig.2.26c in Subsection.2.3.Splitting of long strokes). So, before re-allocating our model resources differently, we explicitly defined the weight assigned to each sample instant with the matrix  $\mathbf{W}$ .

$$\mathcal{L}_{\Psi} = - \sum_{n=1}^N \sum_{t=1}^{T_n} W_{n,t} \log p(\Psi_{n,t} | \mathbf{x}_{n,1:t-1}, \mathbf{z}, \beta_{n,t}) \quad (3.62)$$

Naturally,  $\mathbf{W}$  sums to 1 and is a sparse matrix, with zeros where  $t > T_n$  and  $\frac{1}{NT}$  otherwise. Because this matrix is difficult to represent and manipulate, we aggregate the weights of elements of the same length within a lower triangular matrix  $\mathbf{w}$ . With  $\eta_T$  being the number of samples  $\mathbf{x}_n$  of length  $T$  within our dataset  $\mathcal{D}$ , the *even weighting* procedure described above is formalized by:

$$w_{T,t} = \begin{cases} \frac{\eta_T}{NT} & \text{for } T \in [1, T_{max}] \text{ and } t \in [1, T] \\ 0 & \text{otherwise} \end{cases} \quad (3.63)$$

Fig.3.8a,b display matrix  $\mathbf{w}$  of the original *even balancing* and its marginalization against  $T$  and  $t$  (stroke dataset in panel a, composition dataset in panel b). We remark that the distribution of the length of elements in each dataset, imposed by  $\eta_T$ , primarily affects lines of  $\mathbf{w}$ . Then,  $\mathbf{w}_T$  approximately reflects this distribution, going in opposite directions for strokes and compositions. On the contrary,  $\mathbf{w}_t$  is only decreasing (this is necessarily the case because it results from column summation).  $\mathbf{w}_T$  and  $\mathbf{w}_t$  represent the importance of each variable in the final error computation. By extension, these distributions correlate with the dedicated model resources. For instance, in the stroke model, strokes of length  $T \geq 5$  are under-represented. On this point, we do not really want to push the model to learn to generate these strokes more often than in the dataset  $\mathcal{D}$ ; however, we definitely expect the generation of such strokes to present equal quality w.r.t strokes of length 1. In the *equal weighting* condition, we can imagine that the model will mostly fail to represent longer strokes on their last stroke steps  $t$  (indeed it does so experimentally). There is no reason that the model would miss reproductions at  $t = 1$  and  $t = 2$ , as they are prominently represented within the dataset because of shorter strokes. Thus, our objective in re-allocating model resources is to assign more weight to later stroke steps  $t$ , rather than to entire strokes of longer  $T$ .

If element lengths  $T$  were uniformly distributed (i.e. equal  $\eta_T$ , uniform  $\mathbf{w}$ ),  $\mathbf{w}_t$  would then be an equally spaced decreasing staircase (increasing for  $\mathbf{w}_T$ ). For the composition dataset in Fig.3.8b, we are actually close to this ideal scenario. We will therefore try to impose such theoretical  $\mathbf{w}_t$ , while retaining the original  $\mathbf{w}_T$  ( $\mathbf{w}_{T|t}$  to be precise) as much as possible. We define  $\mathbf{w}'$  as the weighting matrix corresponding to this goal.

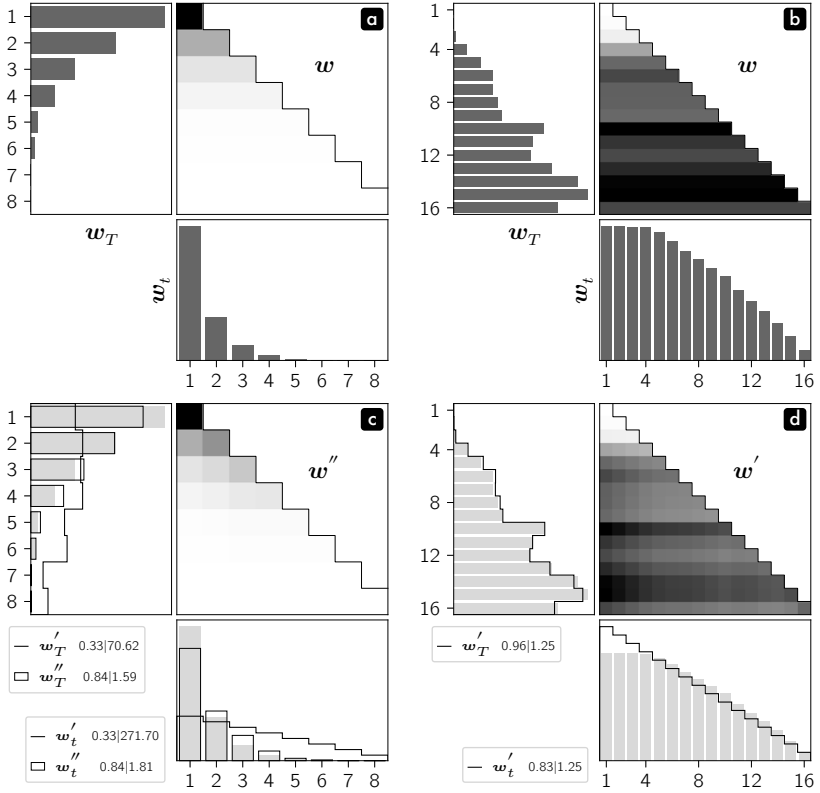


Figure 3.8: Panels **a**, **c** (strokes) and **b**, **d** (compositions) show the weighting matrices  $w$  and their marginalization against  $T$  and  $t$ . Giving each sample instant an equal weighting in the model (**a**, **b**) produces very unbalanced marginals. In order to better allocate model resources for the optimization of  $p(\Psi_t | \mathbf{x}_{1:t-1}, \mathbf{z}, \beta_t)$ , we have designed an improved weighting matrix  $w'$  (**c**, **d**) and a smoothed version  $w''$  (**c**). See Eq.3.64 and Eq.3.65 for details. In legends,  $u_{min} | u_{max}$  values indicate the ranges of distribution ratios against the baseline, indicated by light gray bars.

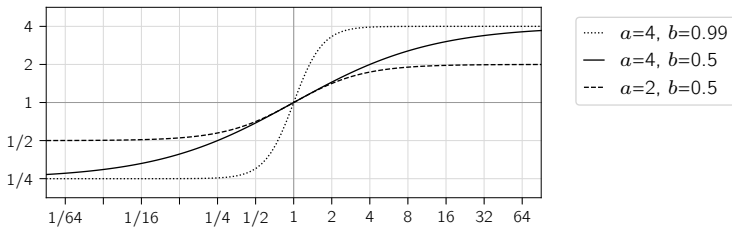


Figure 3.9: Examples of the smoothing function defined in Eq.3.66, for different parameters  $a$  and  $b$ .

### 3 Model implementation

$$\begin{aligned}
 \mathbf{w}' &= \mathbf{w}_{T|t} \mathbf{w}'_t = \frac{\mathbf{w}}{\mathbf{w}_t} \mathbf{w}'_t = \frac{\mathbf{w}}{\sum^T \mathbf{w}} \mathbf{w}'_t \\
 &= \frac{\eta_T}{N\hat{T} \sum_{u=1}^T \frac{\eta_{T_{max}+1-u}}{N\hat{T}}} \frac{T_{max} + 1 - t}{\sum_{u=1}^{T_{max}} u} \\
 &= 2 \frac{\eta_T}{\sum_{u=1}^T \eta_{T_{max}+1-u}} \frac{T_{max} + 1 - t}{T_{max}^2 + T_{max}}
 \end{aligned} \tag{3.64}$$

Distributions resulting from this procedure are illustrated in Fig.3.8c,d (continuous black line in  $\mathbf{w}_T$  and  $\mathbf{w}_t$  sub-plots). For the composition dataset (Fig.3.8d),  $\mathbf{w}'$  is close to the baseline  $\mathbf{w}$ . Density ratios of  $\mathbf{w}'_t$  compared with  $\mathbf{w}_t$  are in the range [0.83, 1.25]. This range is even smaller for  $\mathbf{w}'_T$ , which retains the shape of its original distribution. On the opposite end, for the stroke dataset shown in Fig.3.8c, the baseline distribution is so far from the theoretical one that  $\mathbf{w}'_t$  ratios range between 0.33 and 271.70. The shape of  $\mathbf{w}'_T$  is also drastically modified. In fact, multiplying a weight by 217.70 is counter-productive, and penalizes model faithfulness to the dataset. A maximum magnitude of 2 (i.e. ratios in the range [0.5, 2.0]) is in practice sufficient to make last steps  $t$  of longer strokes minimally represented within the model, particularly in latent space. We therefore introduce a smoothing function  $f$  that produces a new corrected weighting matrix  $\mathbf{w}''$ :

$$\mathbf{w}'' \propto \mathbf{w} f\left(\frac{\mathbf{w}'_t}{\mathbf{w}_t}\right) \tag{3.65}$$

where the actual  $\mathbf{w}''$  must be normalized afterwards because of nonlinearities introduced by  $f$ , defined as:

$$f(u) = \exp\left(\log(a) \tanh\left(\frac{\log(u) \operatorname{arctanh}(b)}{\log(a)}\right)\right) \tag{3.66}$$

where  $a$  is the output ratio upper bound ( $1/a$  for the lower bound), and  $b$  is a smoothing parameter. For instance, if we look at the solid line in Fig.3.9, where  $a = 4$  and  $b = 0.5$ , an original ratio of 4 produces an output ratio of 2. With  $b = 0.99$  (short dotted lines), the upper bound  $a$  is already almost reached when input ratio is equal to  $a$ . By default,  $b$  is set to 0.5 and, for the stroke model, we have used  $a = 1.5$ . From Fig.3.8c, we notice that  $\mathbf{w}''$  produces an effective multiplication range of [0.84, 1.81] on  $\mathbf{w}'_t$ . Accordingly,  $\mathbf{w}'_T$  presents minimal changes in its distribution. Concerning the composition model,  $\mathbf{w}'$  is already appropriate, so no smoothing was necessary.

#### Balancing $\Delta_t$ components in the stroke model

For the stroke model, we have  $\Psi_t = \Delta_t = [\delta_t, \delta'_t, \delta''_t]$ . Each component is independent, so that  $\mathcal{L}_\Psi$  can be decomposed as  $\mathcal{L}_\Psi = \mathcal{L}_\delta + \mathcal{L}_{\delta'} + \mathcal{L}_{\delta''}$ . In

so doing, we assign equal weighting to each component (the target point of this portion of the curve, and the two associated tangents). However, even if tangents represent an important aspect of strokes (e.g. for continuity, see Subsection.2.2.Bézier curves), they can be less accurate than target points without causing substantial degradation of visual appearance. Target points, lying on the stroke, are more determinant of general size and shape. We therefore choose to give an equivalent weight to  $\delta$  (2 dimensions) and  $\delta' + \delta''$  (4 dimensions). In other words, we assign an equal amount of model resources to  $\delta$  and the combination of the two tangents that must share. This empirical adjustment dramatically improves perceived stroke model accuracy, and corresponds to the following balance:

$$\mathcal{L}_{\Psi} = \frac{3}{2}\mathcal{L}_{\delta} + \frac{3}{4}\mathcal{L}_{\delta'} + \frac{3}{4}\mathcal{L}_{\delta''} \quad (3.67)$$

#### Probabilistic training and deterministic validation

Unlike the compositional plane model, the key feature of which is to produce a predictive field of next possible strokes given the previous ones and a latent code  $\mathbf{z}$ , for the other two *true* auto-encoders (stroke and composition models), a point  $\mathbf{z}$  in latent space must correspond to a unique reconstruction  $\mathbf{x}'$ . The model can express some recognition uncertainty in the variance of  $p(\mathbf{z} | \mathbf{x})$ , but it should not do so in the generation of  $\mathbf{x}'$ . At this stage, it seems legitimate to ask the following question: why not make the decoder  $P$  fully deterministic?

For demonstration purposes, we focus on  $\mathcal{L}_{\Psi}$ . If the decoder  $P$  were deterministic, output normal distributions should be replaced by Dirac distributions with modes being  $\mu_{\theta}$ . This approach would require fewer parameters to be optimized, as we could discard all  $\sigma_{\theta}^2$ . The alternative to *maximum likelihood* would then be a  $L^2$  norm between  $\mu_{\theta}$  and  $\Psi$ . The training behavior associated with such a loss function is illustrated in Fig.3.10. We observe that it rapidly decreases to 0 (dotted line). After a few epochs, the  $L^2$  norm would effectively stop driving the system to improve. Typically, a more constantly decreasing error is preferable. Some kind of adaptive scaling would be helpful, and this is precisely achieved by a probabilistic approach with  $\mathcal{L}_{\Psi}$  (solid line). Combining Eq.3.32 and Eq.3.62, we have:

$$\mathcal{L}_{\Psi} = \sum_{n=1}^N \sum_{t=1}^{T_n} \frac{W_{n,t}}{2} \sum_{i=1}^C \frac{(\Psi_{n,t,i} - \mu_{\theta_i})^2}{\sigma_{\theta_i}^2} + \log \sigma_{\theta_i}^2 + \log 2\pi \quad (3.68)$$

In this equation,  $\frac{1}{\sigma_{\theta}^2}$  acts as a scaling value for  $(\Psi - \mu_{\theta})^2$ , actually being  $L^2(\mu_{\theta}, \Psi)$ . The term  $\log \sigma_{\theta}^2$  decreases (thin dotted line), drives  $\sigma_{\theta}^2 \rightarrow 0$  and inversely increases the importance of  $L^2(\mu_{\theta}, \Psi)$ . In other words, the more confident the network becomes about its outputs, the higher the expected precision on  $L^2(\mu_{\theta}, \Psi)$ . Using a probabilistic output instead of a deterministic design can be considered as an optimization trick that produces better training gradients.

### 3 Model implementation

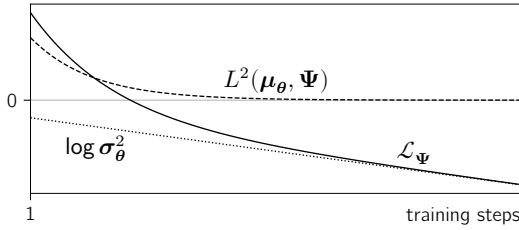


Figure 3.10: Probabilistic vs. deterministic optimization scenarios.

Nonetheless, monitoring  $L^2(\mu_\theta, \Psi)$  at training provides a good estimate of the *true* reconstruction improvements, because it corresponds to the model at evaluation time, where  $\mathbf{x}'$  is greedily sampled. In this case, output values are simply  $\mu_\theta$  for normal distributions, and the pen state, being a categorical distribution, is given by  $\arg \max \log(j_\theta)$ .

There is a final argument in favor of probabilistic outputs at training: the captured uncertainty is an additional type of information about strokes or compositions that the model can provide at low cost (see Section.4.3). These measurements have already proved useful for predicting some human perceptual behaviors (see Subsection.5.3.Perceptual scale prediction from Fisher information).

#### Limiting output variance $\log \sigma_\theta^2$

In theory,  $\log \sigma_\theta^2$  can be any real number. Nonetheless, from Eq.3.68 we have seen that  $\log \sigma_\theta^2$  strongly modulates the magnitude of  $L^2$  reconstruction error. In practice, we observe that the model pushes all  $\log \sigma_\theta^2$  to high values during the very first epoch. This way, the model attempts to minimize risk in its prediction. During optimization,  $\log \sigma_\theta^2$  then decreases proportionally to the increasing precision of  $\mu_\theta$  estimation. While this behavior applies on average, it hides huge discrepancies at the scale of individual predictions. Indeed, some output variance remains uninformative, while most modeling *power* is directed toward *easier* (more likely) arrangements of the dataset. As a result, we want to guarantee a minimal level of output selectivity by defining an upper bound to  $\log \sigma_\theta^2$ .

We know that data are standardized, so that  $\sigma_\theta = 1$  is a coherent upper bound. However, we empirically find this limit too permissive, as it does not really constrain the latent space to take minimal account of every data point. We therefore chose  $\sigma_\theta < \frac{1}{2}$ , leading to  $\log \sigma_\theta^2 < -2 \log 2$ . Naively clamping model outputs would be a bad idea for gradient back-propagation. We prefer to pass it through a negative softplus function, *vertically* shifted according to the upper bound, such as:

$$\log \sigma_\theta^2 = -\log(1 + \exp(\text{input} + a)) - 2 \log 2 \quad (3.69)$$

**Algorithm 3.1:** Backward Clamp

**In:** · *clampMin*, lower bound clamping value (-8)  
 · *clampMax*, upper bound clamping value (None)

**Function forward**(*tensor*):

└ → save *tensor* for backward pass  
 └ **return** *tensor*

**Function backward**(*gradient*):

└ **if** *clampMin* is not None **then**  
 └ └ *gradient*[(*tensor* < *clampMin*) and (*gradient* > 0)] = 0  
 └ **if** *clampMax* is not None **then**  
 └ └ *gradient*[(*tensor* > *clampMax*) and (*gradient* < 0)] = 0  
 └ **return** *gradient*

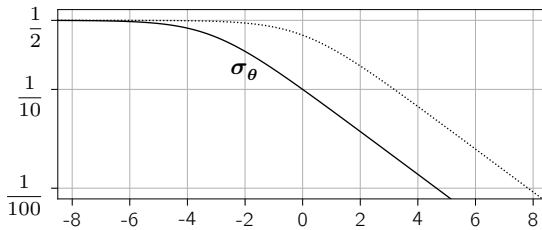


Figure 3.11: Limiting function of output variance  $\sigma_\theta$  is plotted with solid line, and the *horizontally* unshifted function is represented by the dotted line.

with  $a = \log 24$ . It is good practice to keep raw layer outputs centered around 0. For this purpose, the negative softplus function is also shifted *horizontally*, so that *input* = 0 corresponds to  $\sigma_\theta = \frac{1}{10}$ , and  $\log \sigma_\theta^2 = -2 \log 10$ . In Fig.3.11, the final function is plotted with a solid line, and the *horizontally* unshifted version is represented by the dotted line.

This procedure is very efficient and produces a significant positive effect on the model  $L^2$  accuracy. However, we added one last improvement. As stated earlier, during the first training epoch the model output variance is pushed to the upper bound. This behavior is smoothed by the negative softplus function, but in some cases the model tries hard to make values stick to the bound, with *input*  $\approx -30$ . It then takes a *long time* to get these values back to around 0. We speculate that this happens because all available modeling *power* has already been allocated to *easier* data points. We therefore designed a new neural network unit, called *Backward Clamp*. A complete description is reported in Algorithm.3.1.

A normal *Clamp* clips *tensor* within the range [*clampMin*, *clampMax*], and sets *gradient* = 0 for any value of *tensor* outside the bounds. Our unit differs by being the identity in the forward pass, and by adding a requirement during the backward pass: that the gradient does not point in a direction that would push



### 3 Model implementation

values beyond the bounds. This procedure restricts occurrences of  $gradient = 0$  to required situations only, while offering an easier way back into the bounds for extreme values.

#### Exposure bias

In all models with recurrent units (Fig.3.1, Fig.3.3, Fig.3.5, Fig.3.6), gray arrows in  $P$  highlight that  $\mathbf{x}'_t$  is used instead of  $\mathbf{x}_t$  to feed  $\mathbf{h}_t$  at evaluation. During inference, the model is therefore challenged not only with inputs it has never seen (like those within the validation set), but also possibly with inputs sampled from its own predictions. The approximated distribution of  $\mathbf{x}$  implemented by model  $P$  is by definition not guaranteed to effectively match dataset properties. There may be significant discrepancies, producing a model with poor generative abilities, especially on long sequences. Early *mistakes* seriously affect model stability.  $P$  faces difficulties recovering to known distributions, usually resulting in the generation of low-quality and unlikely sequences. This phenomenon is known as the *exposure bias*.

Naively, we could only rely on sampled  $\mathbf{x}'$  at training, and never show *true*  $\mathbf{x}$ . This procedure is feasible, but gives insufficient results in practice or, at best, a very slow learning rate. During the first epochs,  $\mathbf{x}'$  is mostly uninformative and pushes the model to bypass previous elements of the sequence. In VAEs, emphasis is then put on  $\mathbf{z}$  and the initialization of  $\mathbf{h}_{-1}$  by  $\mathbf{g}$ . That is why the standard method of training models on time series still involves ground truth  $\mathbf{x}_n$ . This is sometimes called *teacher forcing*.

Several complicated methods have been explored to address *exposure bias*. Among them, we find reinforcement learning<sup>44</sup> and adversarial discrimination<sup>45</sup>. This last one, called *professor forcing*, introduces a discriminator similar to the GAN architecture<sup>46</sup>, which is conjointly trained to distinguish *true* from *sampled*  $\mathbf{x}$ . The model  $P$  is therefore pushed to fool the discriminator by producing  $\mathbf{x}'$  closer to  $\mathbf{x}$ . This approach is promising, however this complex architecture comes with new issues. For instance, finding the right alternation rate between the training of the main model and the discriminator is not trivial.

A simpler and more comprehensive method is *schedule sampling*<sup>47</sup>. This approach uses either a *true* or a *sampled*  $\mathbf{x}$  depending on a Bernoulli distribution with probability  $p$ . The method is said *scheduled*, because  $p$  is decreased from 1 to 0 based on an exponential decay or similar function: the model slowly changes from training on ground truth inputs toward sampled inputs. Despite reported

---

<sup>44</sup>Ranzato et al., 2016.

<sup>45</sup>Lamb et al., 2016.

<sup>46</sup>Generative Adversarial Networks (GANs) have been introduced by I. Goodfellow, 2016; I. J. Goodfellow et al., 2014.

<sup>47</sup>S. Bengio et al., 2015.

performance improvements, this technique has received critical remarks. One paper demonstrates the theoretical difference between the two distributions optimized in *schedule sampling* (when  $p$  is 1 or 0), and thus exposes an inconsistent training objective:

We suggest that scheduled sampling works by pushing models towards a trivial solution of memorizing distribution of symbols conditioned on their position in the sequence, rather than on the prefix of preceding symbols. In RNN terminology, this would mean that the optimal architecture under [schedule sampling] uses its hidden states merely to implement a simple counter, and learns to pay no attention whatsoever to the content of the sequence prefix.<sup>48</sup>

We would like to moderate this criticism. Experimentally, we do see a change of regime along epochs, especially when the chosen decay function is not well adapted to the situation, but regime differences are not as extreme as expressed theoretically. We believe that the convergence to a middle point is possible, and helps to produce models more resilient to both regimes. Experimentally setting the probability of the Bernoulli to a fixed value  $p = 0.5$  consistently produces better results at evaluation: better than the original teacher forcing, and better than a pure self-sampling approach. We speculate that this procedure works by helping the model to take into equal account  $\mathbf{z}$  and previous inputs.

In the original article on *scheduled sampling*, the authors do not back-propagate gradients through sampled  $\mathbf{x}'$ . Working on language modeling, their input word space was discrete. Without an easy reparameterization trick for categorical distribution, the issue had been left unresolved. It was subsequently addressed with a trick called *Gumbel softmax*<sup>49</sup>. Similarly to the reparameterization trick for normal distributions (see details in Subsection.3.1.Reparameterization trick), *Gumbel softmax* makes use of an external source of stochasticity which does not have to be learned, i.e. does not require back-propagation. This auxiliary variable is sampled from a Gumbel distribution, corresponding to  $\mathbf{G} = -\log(-\log \mathbf{u})$ , with  $\mathbf{u} \sim \text{uniform}(0, 1)$ . Combined with log probabilities  $\log(\mathbf{j}_{\theta_t})$  of categorical distributions, samples can be computed as follows:

$$\beta_{t,i} = \frac{\exp((\log(\mathbf{j}_{\theta_t,i}) + \mathbf{G}_{t,i})/\tau)}{\sum_{k=0}^1 \exp((\log(\mathbf{j}_{\theta_t,k}) + \mathbf{G}_{t,k})/\tau)} \quad (3.70)$$

We remark that this is basically a softmax function modulated by a  $\tau$  parameter. As  $\tau \rightarrow 0$ , this function tends to the argmax function providing the primarily expected one hot vector definition of samples from a categorical distribution. In the literature, it is proposed to progressively decrease  $\tau$  as a function of epochs from 1 toward a small non-zero value. We have chosen not to modulate  $\tau$  and fixed it to 1, with no significant loss.

<sup>48</sup>Huszár, 2015.

<sup>49</sup>Goyal et al., 2017; Jang et al., 2017; Maddison et al., 2017.

### 3 Model implementation

The equation above is defined for pen states  $\beta$ , but it can be used similarly on mixture weights  $m$ . However, in both cases, we still need one hot outputs at inference, or during the sampling of  $\mathbf{x}'$ . There exists a variant termed *straight through Gumbel softmax*, where a greedy argmax function is used in the forward pass, while the *Gumbel softmax* is utilized for back-propagation only. We point out that this procedure is biased because the original distribution is not completely guaranteed. Nevertheless, the continuous gradient offered by this method remains highly beneficial.

#### Miscellaneous

We detail a few more remaining implementation details. First, neural network units such as  $\mathbf{h}$ ,  $\mathbf{y}$ ,  $\mathbf{g}$ , and their encoder counterparts are generic denominations. In practice, they can be materialized by one or more successive layers of different sizes. This is summed up in Fig.3.12.

In particular, all LSTM units are double layered. Although other strategies exist to make RNNs deeper<sup>50</sup>, this solution is effective and easy to implement.

When two layers are stacked, such as in  $\mathbf{h}^q$ ,  $\mathbf{h}$  and  $\mathbf{g}$ , we add dropout units in between these layers. Dropout is a technique to limit overfitting in the model, and therefore to improve its generalization<sup>51</sup>. The idea is to randomly zero some connections between layers at training. Each connection has  $p$  chance to be canceled from a Bernoulli distribution. In so doing, the average magnitude of the outputs generated by dropout units is scaled by  $1 - p$ . Because the units act as an identity function at evaluation, we need to compensate for the loss by  $\frac{1}{1-p}$  at training. We globally chose  $p = 0.1$ .

ReLU activation function is used by default for all intermediary linear layers (fully connected). Output layers are left without activation function, except for  $\mathbf{g}$  using a  $\tanh$  function, and some part of  $\mathbf{y}$  concerning  $\beta$ ,  $m$  and  $\sigma_{\theta}^2$ , using respectively softmax functions (see Eq.3.30) and a shifted negative softplus function (see Subsection.3.5.Limiting output variance  $\log \sigma_{\theta}^2$ ).

Neural network weights are initialized randomly before the first training iteration. For instance, linear layers parameters are usually sampled from a normal distribution. Results can be improved when  $\sigma = gain \sqrt{\frac{2}{size_{in} + size_{out}}}$ .  $gain$  is chosen depending on the associated activation function ( $gain = \sqrt{2}$  for ReLU activation,  $gain = \frac{5}{3}$  for  $\tanh$ , and  $gain = 1$  otherwise)<sup>52</sup>. Concerning LSTM units, the initialization

---

<sup>50</sup>Pascanu et al., 2014.

<sup>51</sup>Hinton et al., 2012.

<sup>52</sup>Glorot and Bengio, 2010.

### 3.5 Practical model training

	x	e	d	$\mathbf{h}^q$	$\mathbf{y}^q$	$\mathbf{g}$	$\mathbf{h}$	$\mathbf{y}$
stroke model	$\Delta+\beta$	64	128	x eb eb	g(ebn z)	s(z dn dn)	x+z d d	d g $\Delta+\beta$
comp. model	$\rho+z+\beta$	256	512	x eb eb	g(ebn Z)	s(Z dn dn)	x+Z d d	d g( $\rho+z$ )+ $\beta$
comp. plane	$\rho+z+\beta$		512			s(Z dn dn)	x+Z d d	d m+mg( $\rho+z$ )+ $\beta$
$\Delta$	6	differential cubic Bézier parameters			b	2	multiplier for bidirectional LSTM	
p	2	stroke initial position			g	2	normal distribution parameters $\mu$ and $\sigma$	
$\beta$	2	pen states, Bernoulli distribution			n	2	number of layers of recurrent units	
z	6	stroke latent space $K$			s	2	hidden and cell states of the LSTM unit	
Z	16	composition latent space $K$			m	8	number of mixtures	

Figure 3.12: Table of layer sizes. i|o materializes a layer, with its input and output sizes.

procedure is replicated from the sketch-rnn model implemented in the Magenta Project<sup>53</sup>. Finally, bias parameters for linear layers are initialized to 0.

Gradient descent optimization is performed with the Adam algorithm<sup>54</sup>. The initial learning rate is  $10^{-4}$ , and a scheduler reduces this learning rate by a factor of 0.9 when the network validation accuracy stays stable across 10 epochs. When a reconstruction  $L^2$  error is available (stroke model and composition model), this is the value tracked by the scheduler. A minimal learning rate is set to  $10^{-5}$ .

To improve back-propagation in recurrent neural networks, and limit their associated vanishing or exploding gradient issues, it has been proposed to clip these gradients<sup>55</sup>. Such procedure can be operated by clamping the gradient at some threshold, or on the norm of all parameters taken as a single vector. The gradient is then scaled by  $\min\left(\frac{\text{clipVal}}{\|\text{gradient}\|}, 1.0\right)$ . We opted for the latter option with a clipping value of 1.0.

Finally, every random process in Python and its scientific libraries (Numpy/Scipy<sup>56</sup>) were seeded, so that the different outcomes are reproducible, i.e. reproducible training sessions and results. In addition, PyTorch and CUDA<sup>57</sup> were set up to choose deterministic algorithms instead of faster optimized ones. Despite these efforts and the use of the same exact parameters and computing devices, there is still a minor source of randomness across different runs of the training procedure. This is due to a *bug* in the implementation of some components of the RNN unit in CUDA. Nonetheless, the magnitude of this issue is negligible compared with the trends found in the results.

<sup>53</sup>Sketch-rnn is the model developed by Ha and Eck, 2017. For more details on the Magenta Project: <https://magenta.tensorflow.org>

<sup>54</sup>Kingma and Ba, 2017.

<sup>55</sup>Y. Bengio et al., 2012; Pascanu et al., 2012.

<sup>56</sup>Please find more information at: <https://numpy.org> and <https://scipy.org>

<sup>57</sup>CUDA is the GPU computing library underlying PyTorch functionalities, and developed by NVIDIA. Please find more information at: <https://developer.nvidia.com/cuda-toolkit>



## *Part II*

# *Composition exploration*



## 4 Model results and tools

The second part of this manuscript is dedicated to the exploration of the functional models. We will cover raw results, psychophysical experiments investigating the homogeneity of the latent space, and finally artistic inquiries.

In this chapter, we primarily focus on models training logs, straightforward reconstructive accuracy, and generative abilities of the models. In a second time, we discuss the optimal way to travel in the latent space, i.e. interpolations between  $\mathbf{z}$  samples. Finally, the different models have several sources of stochasticity, so we detail how each probabilistic aspect can represent a quantitative tool to conduct measurements, and ultimately study composition.

### 4.1 Reconstruction and generation

Controlling the performance of our models is not an easy task. On the contrary to usual machine learning procedures, we are not solely focused on the reconstruction accuracy. The artistic relevance and *expressiveness* of the latent space are the most expected qualities of the model. As a result, there are frictions between quantitative and qualitative evaluations. However, even if the artistic feeling of model outputs are maybe more important for me as the modeler, checking a limited number of generated outputs can be misleading. The constructed space is very large, and a few samples may not be representative. In addition, different visual representations of the same data, e.g. with or without dynamic, can provoke very different artistic judgments. Therefore, we searched for the best monitoring values, while at the same time we tried to find the best model architectures and hyperparameters. There is no space here to make the history of changes in training monitoring values, and the evolution of ways to interpret them. We will only describe the current state of our training method.

#### *Training logs*

During prototyping and training phases of a model, it is important to have access to quantitative values to monitor the improvements given by different designs or changes in hyperparameters. So, besides the raw optimized values (Fig.4.1c,d,e,h, Fig.4.2c,d,e,h and Fig.4.3b,c), we track other important values, as defined in the



#### 4 Model results and tools

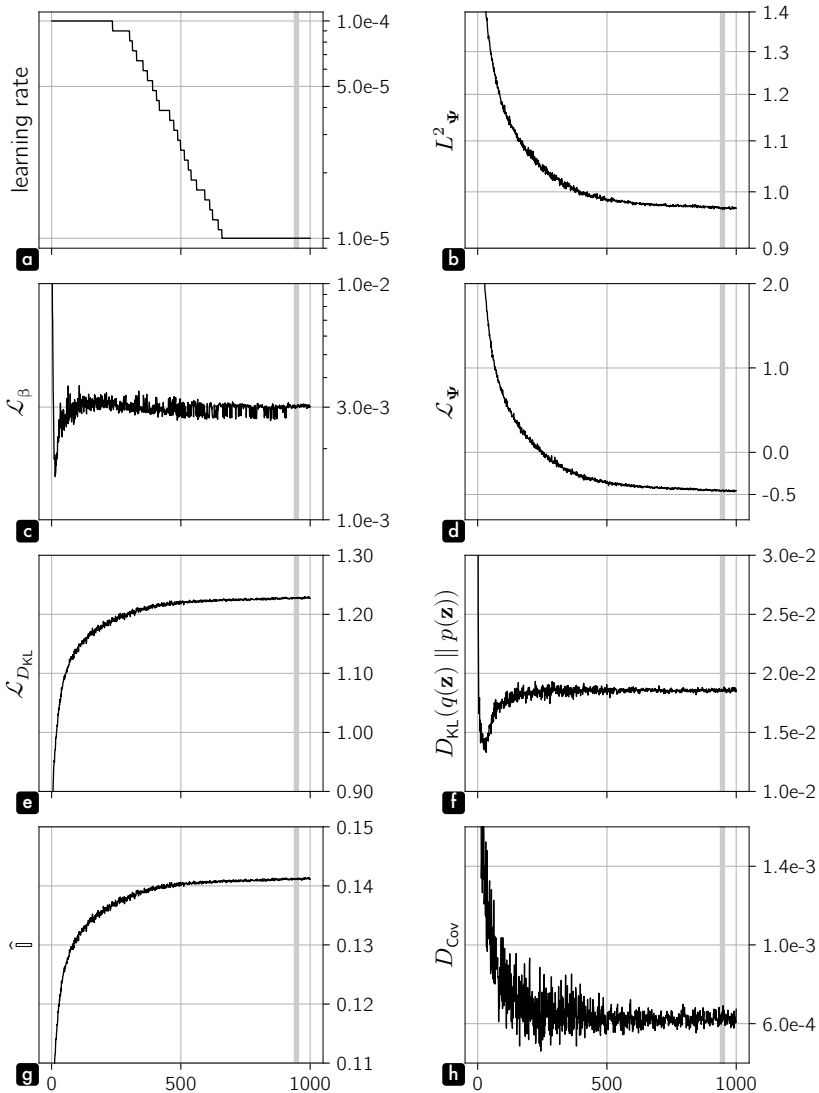


Figure 4.1: Training logs of the stroke model. Only values from panels **c**, **d**, **e** and **h** are directly optimized. Other values are only controlling measurements. The best epoch, considered as the fully trained state of the model, is highlighted with a gray vertical line.

## 4.1 Reconstruction and generation

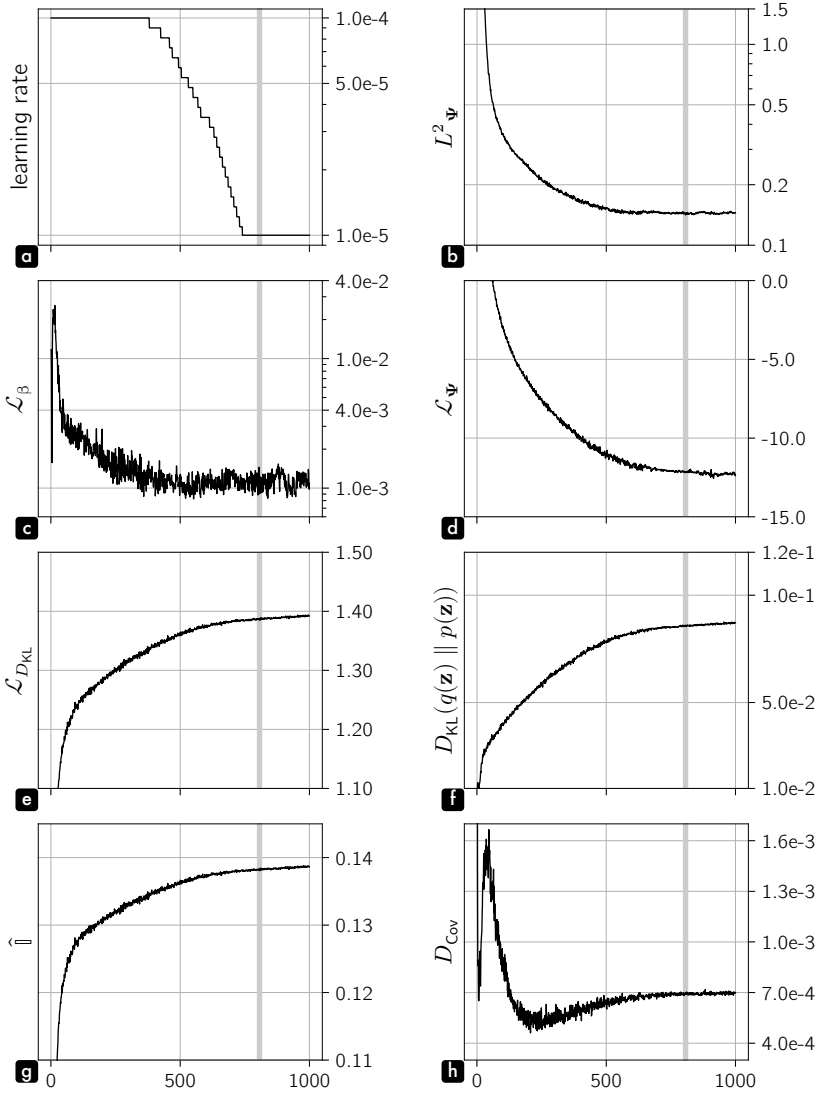


Figure 4.2: Training logs of the composition model. Only values from panels **c**, **d**, **e** and **h** are directly optimized. Other values are only controlling measurements. The best epoch, considered as the fully trained state of the model, is highlighted with a gray vertical line.

#### 4 Model results and tools

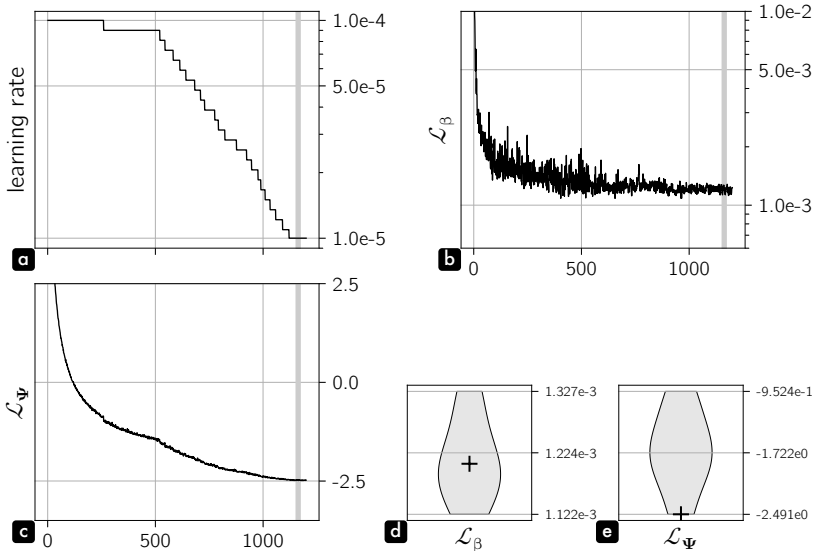


Figure 4.3: Training logs of the compositional plane model are presented in panels **a**, **b** and **c**. The best epoch, considered as the fully trained state of the model, is highlighted with a gray vertical line. Panels **d** and **e** show compositional plane model score distributions. The selected instance is marked with a cross.

previous chapter (Fig.4.1a,b,f,g, Fig.4.2a,b,f,g and Fig.4.3a). Of course, all these measurements are done on the validation dataset.

The main use of these training logs is to check the stability of the training procedure. Inter-epoch variability of the model state should be as smooth as possible, while still improving. The optimization process is highly nonlinear, and the solution space may be far from perfectly convex. Therefore, a high variability can be a sign of a too high learning rate, and can eventually lead to exploding gradients. On the opposite, a too low learning rate may keep the model state in a local minimum with poor performance. Another source of high variability could be a too small batch size. In this case, the random ordering of data could influence too dramatically the optimization. However, this is unlikely to happen in our case as the batch size is 64 for stroke and composition models to guarantee good estimates of  $D_{\text{Cov}}$ . In logs presented in Fig.4.1, Fig.4.2 and Fig.4.3, variables appear very stable, except for  $D_{\text{Cov}}$  and  $\mathcal{L}_\beta$ . In fact,  $\mathcal{L}_\beta$  apparent fuzziness is more likely the result of a scaling effect, as this variable is quickly converging after a few epochs. Then, even if the plotted  $D_{\text{Cov}}$  is estimated on the whole batch at validation, its behavior is the result of a less reliable estimate per mini-batch at training. Nevertheless, it does not prevent the variable from showing a long term convergence (see Fig.4.1h).

The second use of these logs is to check whether the training procedure does not overfit, and is really trained until full convergence. Overfitting is typically observed when a variable presents a U shape. It means that the model still improved its accuracy on the training set, but did not generalize discoveries on the validation set. It has to be particularly monitored for the reconstruction loss. In our case, it concerns  $\mathcal{L}_\Psi$ ,  $\mathcal{L}_\beta$  and  $L_\Psi^2$ . On stroke and composition models, the main variable, driving the learning rate scheduling and the best epoch selection is  $L_\Psi^2$ . From Fig.4.1a,b and Fig.4.2a,b, we observe that the best epochs, highlighted with gray vertical bars, are in both cases selected after that the learning rate reached the chosen lower bound, and are located in a nearly flat regions of  $L_\Psi^2$ . Other variables also appear quite stable, demonstrating that models are trained to full convergence. Concerning the compositional plane model (Fig.4.3a,b,c),  $L_\Psi^2$  is not accessible as the output is a mixture distribution.  $\mathcal{L}_\Psi$  is then the driving variable. The log behavior is also acting as expected. Note that no value is monitored concerning the latent space for this model, as the encoding of  $\mathbf{z}$  is inherited and fixed from the composition model.

### Score distributions

For legibility in previous logs, we have only plotted the best training run of each model. Indeed, with the exact same hyperparameters, model performance can significantly vary. This is due to the stochasticity in the neural networks weights initialization, in the mini-batch selection, and in the different data augmentation processes. Thus, in order to attest of the robustness of the results, it is customary to run the whole training procedure with different *seeds*. These seeds are some sort of keys, generating reproducible sequences of random numbers. In Fig.4.4, Fig.4.5 and Fig.4.3d,e, we plot the distribution of monitored values at best epochs for each seed. The violin plots are constructed over 6 seeds and the location of the selected one (shown in previous logs) is marked with a cross. First, we remark that results are in general similar across seeds. The shape of the different distributions are not showing problematic outliers. The most uneven distribution concerns the  $D_{\text{Cov}}$  value of the composition model (Fig.4.5g). Nonetheless, the selected seed is in the denser part. Globally, these plots guarantee that our trained models are not outliers, resulting from very lucky situations.

The second interesting aspect of these distributions is to highlight the rules guiding our selection of the most successful seed. A simple logic would be to choose the seed producing the best value for the driving variable, involved in the selection of the best epoch. This logic applies for the compositional plane model, but concerning stroke and composition models, our main interest is on constructing a good latent space, i.e. with independent dimensions and a density close to the chosen prior. These concerns correspond to Fig.4.4e,g and Fig.4.5e,g. We remark that for both models we chose the best compromise between the smallest  $D_{\text{Cov}}$

## 4 Model results and tools

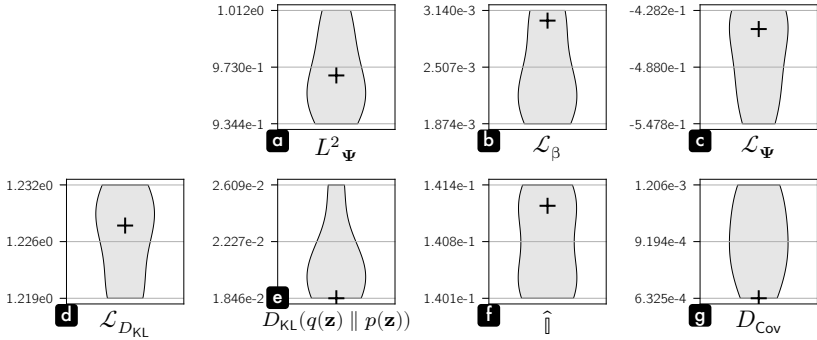


Figure 4.4: Score distributions of the stroke model. The selected instance is marked with a cross.

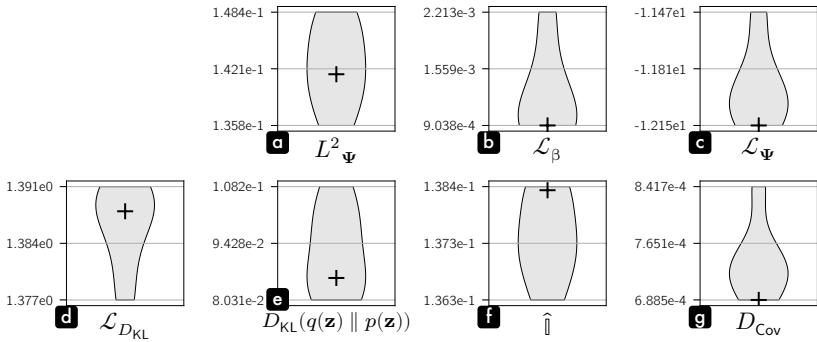


Figure 4.5: Score distributions of the composition model. The selected instance is marked with a cross.

and the smallest  $D_{KL}(q(\mathbf{z}) \parallel p(\mathbf{z}))$ , even if these seeds do not correspond to the best reconstruction score  $L^2_\Psi$  (see Fig4.4a and Fig4.5a). Then, why not directly choosing  $D_{Cov}$  or  $D_{KL}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  for the best epoch selection? The answer is that  $D_{KL}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  is pushed toward a small value, but still increase over the training (e.g. Fig.4.2f). In addition,  $D_{Cov}$  has an unpredictable general behavior (e.g. non-monotonic in Fig.4.2h), that is difficult to handle as a driving variable.

### Latent space dimensionality

Another difficult hyperparameter to adjust is the dimensionality  $K$  of  $\mathbf{z}$ . We have not found an objective way to determine a right value. Fig.4.6 and Fig.4.7 show the influence of  $K$  on the monitoring variables. For the stroke model, we have investigated  $K = [5, 6, 7, 8, 10]$  and for the composition model,  $K = [12, 16, 20]$ . The main issue is that the reconstruction error, e.g.  $L^2_\Psi$ , expectedly improves

## 4.1 Reconstruction and generation

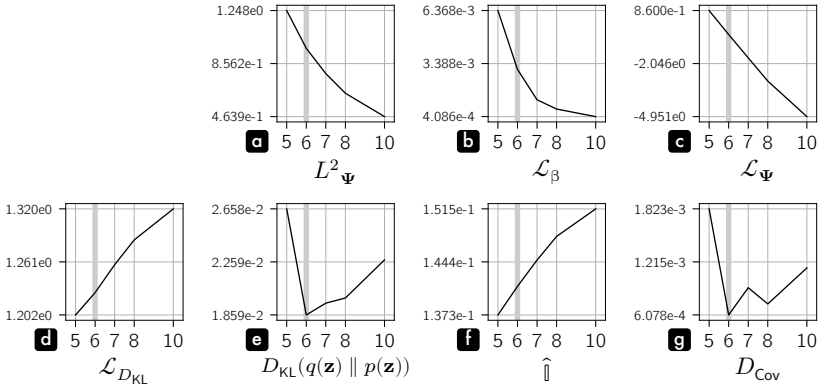


Figure 4.6: Scores w.r.t the dimensionality  $K$  of  $\mathbf{z}$  for the stroke model. The selected  $K$  is highlighted with a gray vertical line.

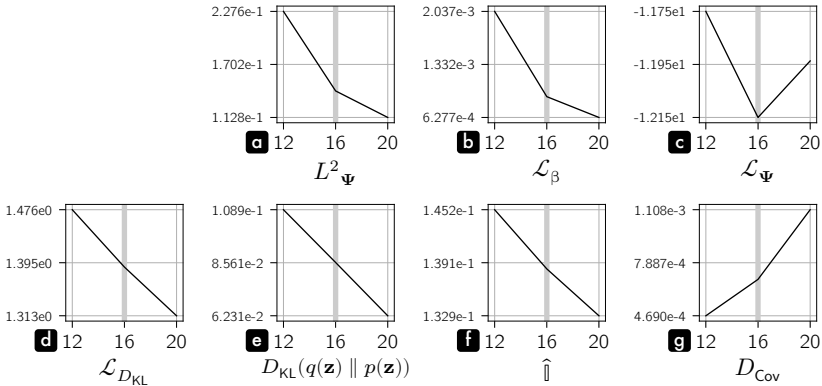


Figure 4.7: Scores w.r.t the dimensionality  $K$  of  $\mathbf{z}$  for the composition model. The selected  $K$  is highlighted with a gray vertical line.

with more dimensions, while more dimensions increases the difficulty to understand the latent space. So, in both cases, the choice of  $K$  results from a complicated tradeoff.

Concerning the stroke model, we will see in the next subsection that  $K = 10$  would be very beneficial for the reconstruction accuracy of longer strokes (see Fig.4.9). But, this improvement seems to result from some sort of overfitting. Qualitatively, we observed that it also produced unstable and fuzzy transitions (i.e. interpolations). We actually face another aspect of the very unbalanced length distribution in the stroke dataset (see Fig.2.26c), which may explain Fig.4.6e,g,

## 4 Model results and tools

where minimal values of  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  and  $D_{\text{Cov}}$  indicate an optimal  $K = 6$ . Even if  $K = 6$  does not capture all the expected expressiveness of strokes, the dataset imposes this limit for a smooth and coherent latent space. The only alternative would be to increase the diversity of long stroke in the dataset with new drawings, but it is possible in future works only.

However, the above logic does not stand for the composition model. In Fig.4.7e,g, we see that  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  decreases with  $K$ , while  $D_{\text{Cov}}$  increases.  $K = 16$  is therefore not optimal, but a subjective compromise. A higher number of dimensions would also have rendered the exploration of the latent space even more difficult with psychophysical experiments<sup>1</sup>.

### Reconstruction and prediction

The most immediate visual evaluation of the model is the reconstruction of some ground truth data. We encode/decode some  $\mathbf{x}_n$ , and compare the result with the original. To avoid the variability of the encoding, we greedily sample  $\mathbf{z}$  from  $q(\mathbf{z} \mid \mathbf{x})$ . It basically corresponds to  $\mu_\phi$  given by the encoder  $Q$ . Some reconstructions of strokes are shown in Fig.4.8. Reconstructions are in black over their ground truth in gray. The result is generally good. Unsurprisingly, the biggest errors happen on longer and more complicated strokes. Fig.4.9 also displays some reconstructions of strokes, but in a compositional context. Fig.4.9a is using the selected  $\mathbf{z}$  dimensionality  $K = 6$ . We observe that the lack of precision on longer strokes is really obvious and can, in some situations, completely disrupt the composition. As a result, even if the composition model were perfect, the stroke embedding would limit the possible visual reconstruction anyhow. However, we repeat that the choice of a lower  $K$  is made consciously to preserve a smooth latent space. A dimensionality  $K = 10$ , as presented in Fig.4.9b, is tempting for its far better results on longer strokes, but it would lead to interpolation issues at the compositional level.

Fig.4.10a displays reconstructions from the composition model. Resulting examples do not seem very accurate, but it is difficult to unravel errors from the composition model and intrinsic limitations of the stroke model. To overcome this issue, we can replace the ground truth underlay by the best possible reconstruction as depicted in Fig.4.9a. This is plotted in Fig.4.10b and the composition model accuracy is then very satisfactory. Only the two lower right compositions present noticeable errors. They possess a smaller number of strokes, and it may reflect a comparable issue as for the stroke dataset, i.e. compositions with fewer strokes are under-represented at the dataset level (see Fig.2.28d).

---

<sup>1</sup>In comparison, the paper associated with sketch-rnn (Ha & Eck, 2017) used  $K = 128$ . Even if their motivation is mostly driven by the reconstruction accuracy, we find this dimensionality impractical for latent space investigations. In addition, we believe that such very high  $K$  is very likely to produce models overfitting the dataset.

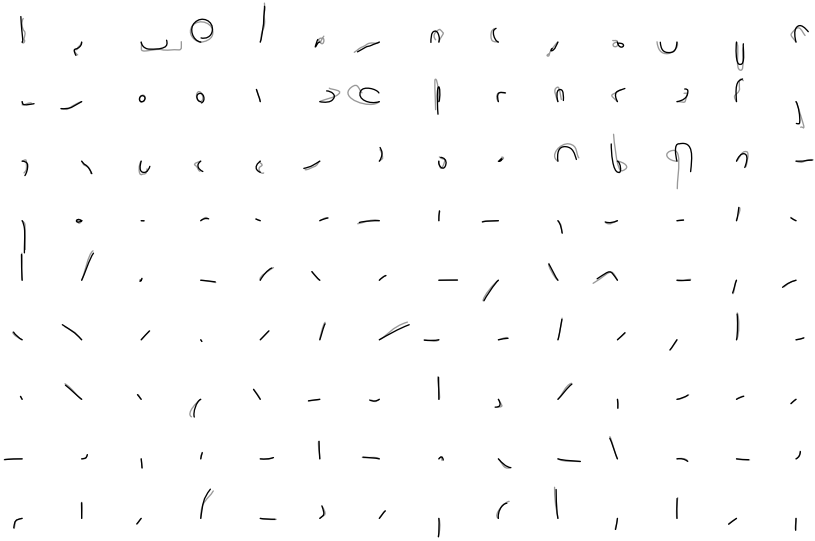


Figure 4.8: Reconstruction of strokes. Ground truth and reconstructed strokes are respectively in gray and black.

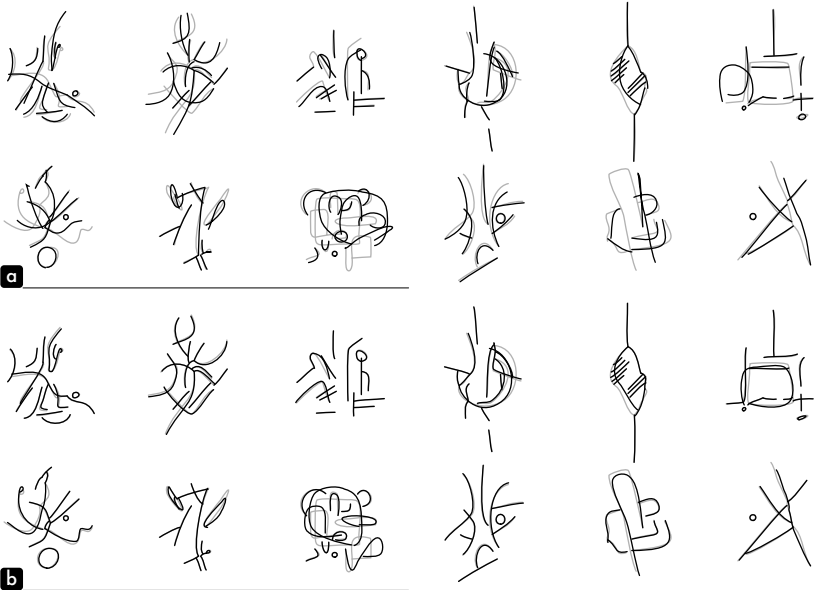


Figure 4.9: Panels **a** and **b** show reconstructions of strokes in a compositional context with  $K \in [6, 10]$ . Ground truth and reconstructed strokes are respectively in gray and black.



## 4 Model results and tools

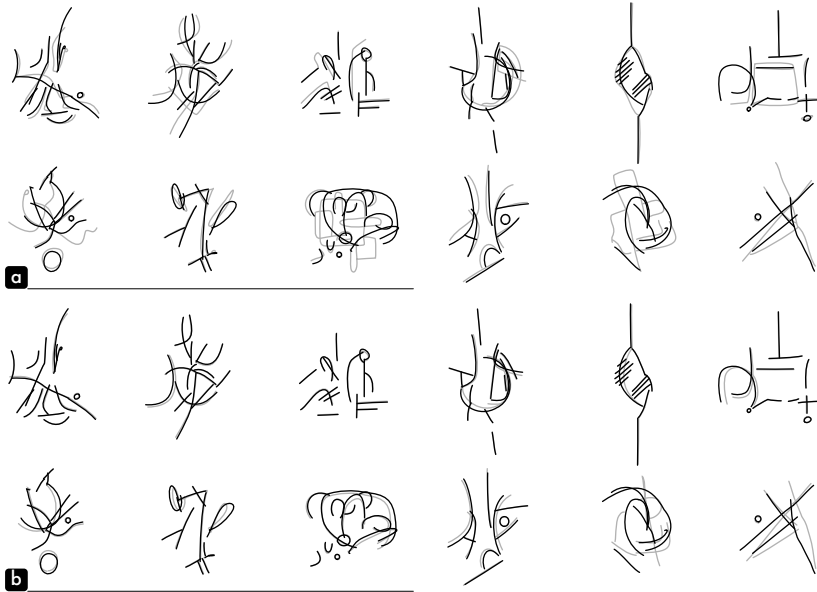


Figure 4.10: Reconstruction of compositions. In panel **a**, ground truth and reconstructed compositions are respectively in gray and black. In panel **b**, the ground truth is replaced by the best possible reconstruction, including stroke model limitations, as reproduced in Fig.4.9a.

Concerning the compositional plane model, an input  $\mathbf{x}_n$  does not have a unique reconstruction. Model output is a mixture distribution, designed to investigate different *next stroke* alternatives given a context, i.e. previous strokes and a  $z$  encoded by the composition model (symbolizing a target *mental image* of the final work). Therefore, instead of showing reconstructions of complete input sequences  $\mathbf{x}_{n,1:T_n}$ , we display the prediction of two strokes conditioned on an input sequence  $\mathbf{x}_{n,1:C_n}$ , with  $C_n = T_n - 2$  (see Fig.4.11a). Fig.4.11b shows different alternatives of such last strokes predictions. Most results seem pertinent, ranging from unusual to interesting. Nonetheless, the model is sometimes failing to propose *good* ending strokes. The biggest issue (very noticeable in Fig.4.11b top row) is a tendency to nearly repeat previous strokes. Despite multiple adjustments during the prototyping phase, this artifact is still present. It is probably an architectural problem, which will require in depth future investigations.

### Latent space

Generative abilities of the different models are more important to us than their reconstructive accuracy. One aspect to guarantee coherent, diverse and *interesting*

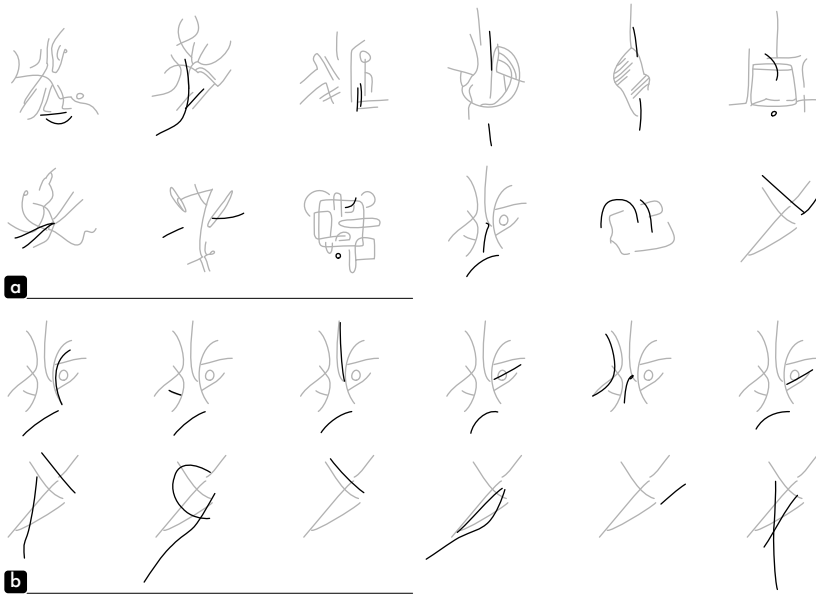


Figure 4.11: Compositional plane model predictions of some alternatives of ending strokes (in black). These predictions are based on the conditioning strokes shown in gray. Panel **a** displays one example for different compositions, and panel **b** different alternatives for the same composition.

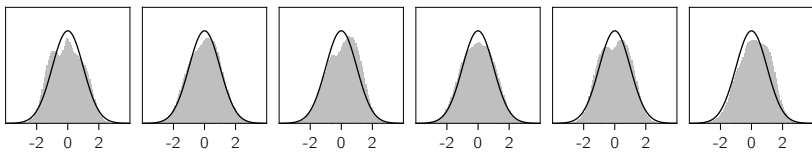


Figure 4.12: Stroke latent space distribution per dimension (in gray) and corresponding expected standard normal distribution (black line).

new samples is the fitting accuracy of the latent space to the expected prior, i.e. a multivariate standard normal distribution.  $D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}))$  is a good general quantitative measure, but we can also investigate actual densities per dimension. For strokes in Fig.4.12, we notice that the matching of the prior is excellent. On the contrary, concerning the compositions, some dimensions present in Fig.4.13 bimodal distributions. A trivial explanation could be that there is not enough dimensions to capture all compositional regularities, and that some of them have to be joined together. However, this effect still appears with higher  $K$ . A better interpretation could be the existence of *binary* regularities, e.g. features which could be either horizontal or vertical, but not in an intermediary

## 4 Model results and tools

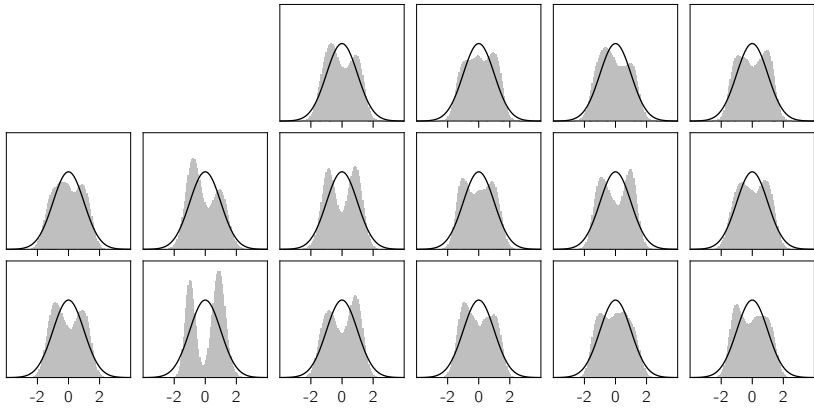


Figure 4.13: Composition latent space distribution per dimension (in gray) and corresponding expected standard normal distribution (black line).

orientation. The empirical demonstration of this statement is difficult. As we only matched the first two moments of  $q(\mathbf{z})$  and  $p(\mathbf{z})$  with the  $D_{Cov}$  regularizer (see Subsection.3.5.Regularizing  $D_{KL}(q(\mathbf{z}) \parallel p(\mathbf{z}))$ ), bimodal distributions can still be well centered around zero and present a reasonable unit variance. Only the optimization of higher moments could improve this issue, and validate (or not) the existence of *binary* regularities. This investigation is left for future works.

### Generation

Judging qualitative aspects of generated strokes and compositions is challenging. Whether it fits or not my compositional expectations is very subjective. From this perspective, I am generally very satisfied with the results. While randomly sampling in the latent space, there is still few areas producing fuzzy outputs, but given the relatively small size of my personal dataset, it can be considered as a good achievement. Examples displayed on the next pages may help you to build your own opinion.

For each model, there are two types of figure. A first set shows independent random samples, while a second set presents improvisations around a given sample. For stroke and composition models, we randomly explore the neighboring latent space of a given sample, i.e.  $\mathbf{z} + \mathcal{N}(\mathbf{z}_{noise} \mid \mathbf{0}, \sigma \mathbf{I})$  with  $\sigma = \exp(-2)$ . For the compositional plane model, the variability comes directly from different output alternatives, given the same sample  $\mathbf{z}$ .

Generated of strokes in Fig.4.14 and Fig.4.15, particularly present a rich variability and expressiveness. They produce a feeling of simplicity compared to real strokes

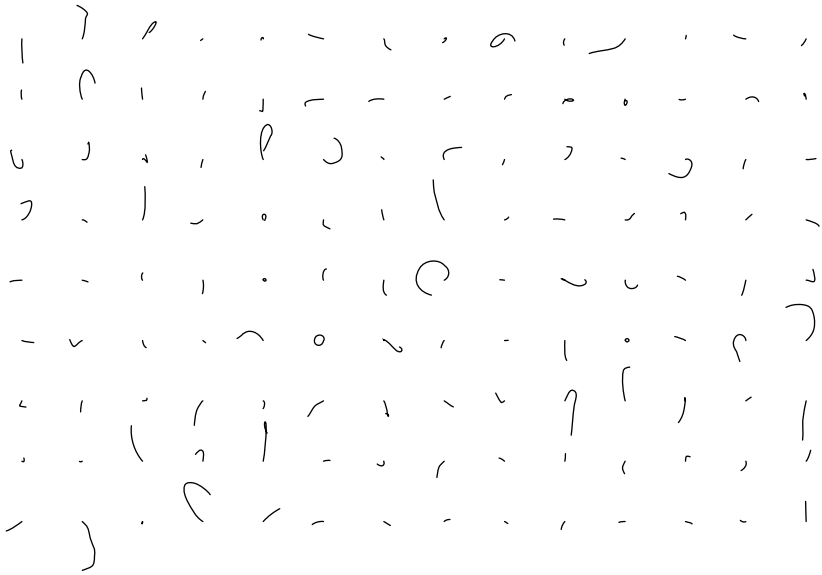


Figure 4.14: Generation of strokes.

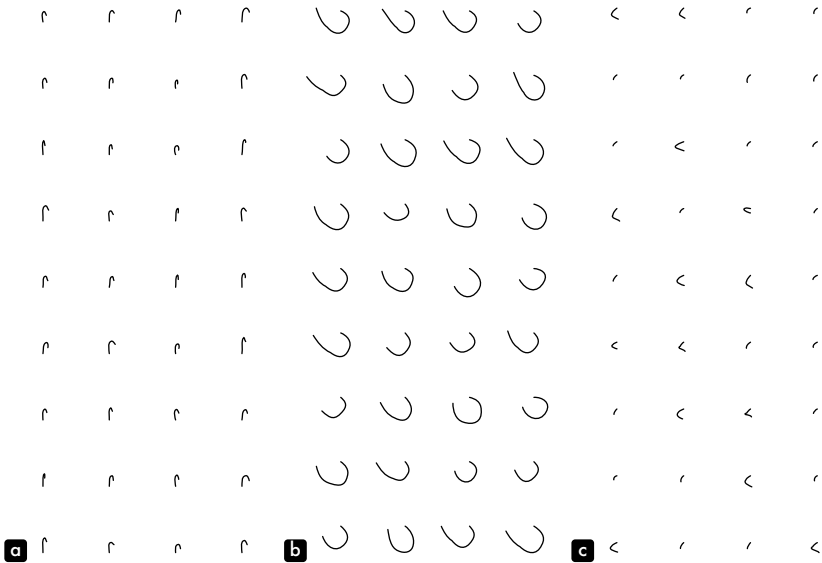


Figure 4.15: Generation of strokes, local explorations.

#### 4 Model results and tools

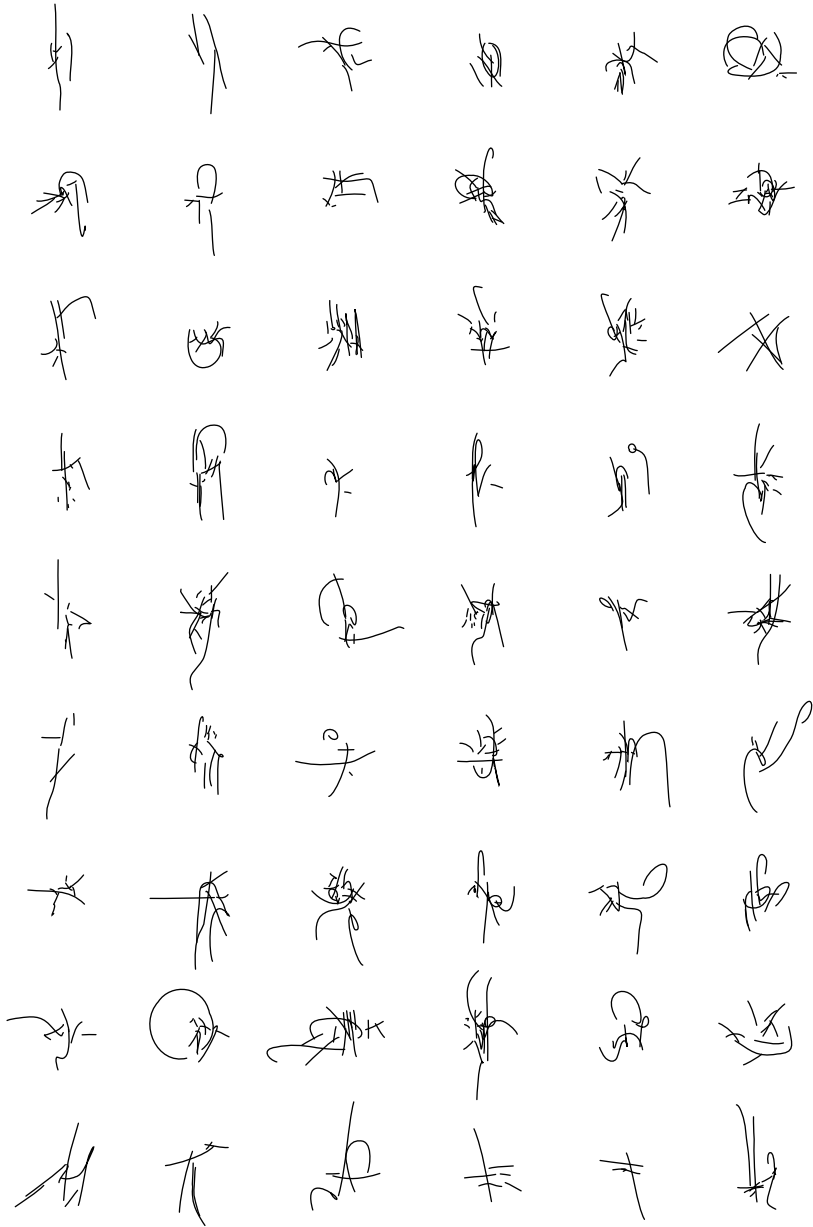


Figure 4.16: Generation of compositions.

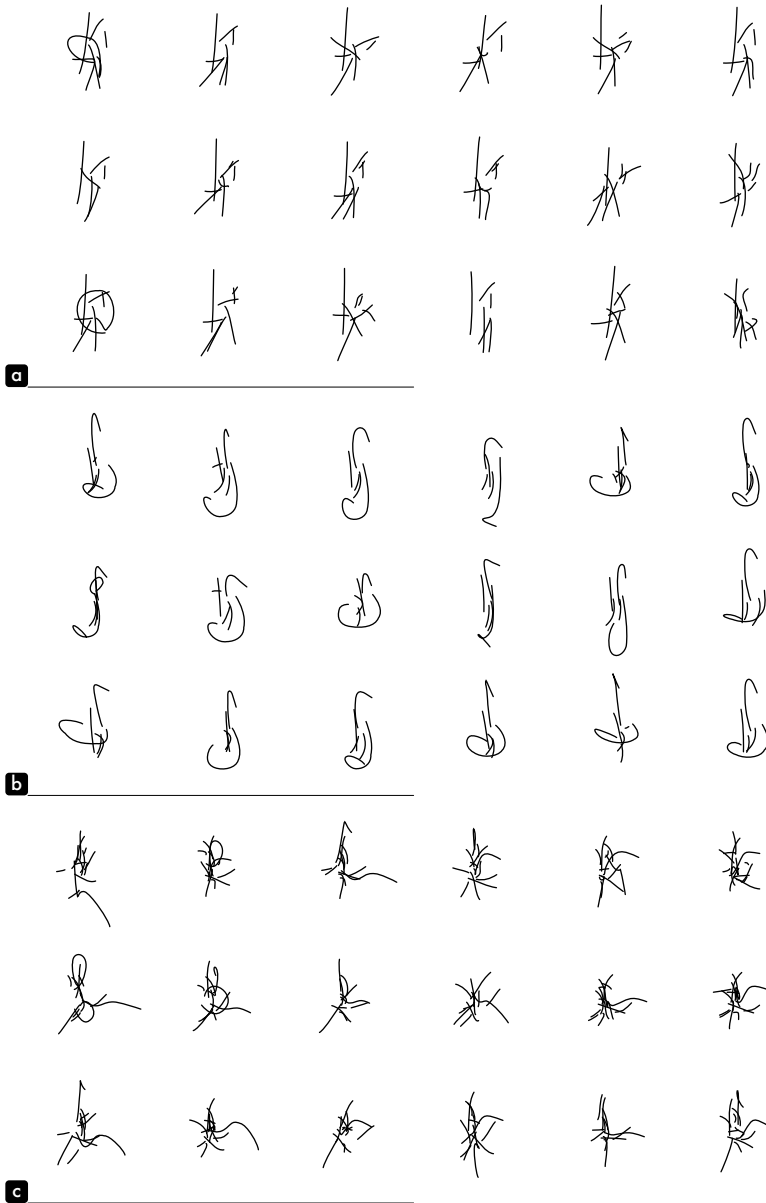


Figure 4.17: Generation of compositions, local explorations.

#### 4 Model results and tools

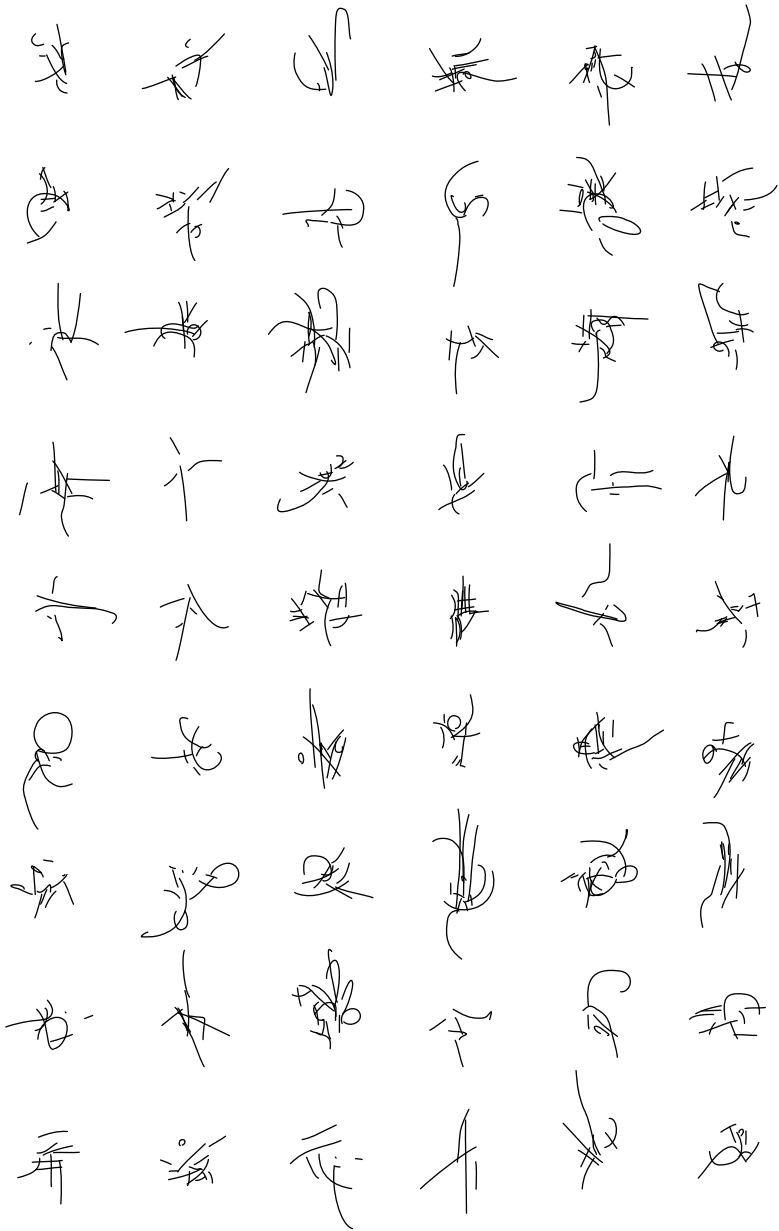


Figure 4.18: Generation of compositions from the compositional plane model.

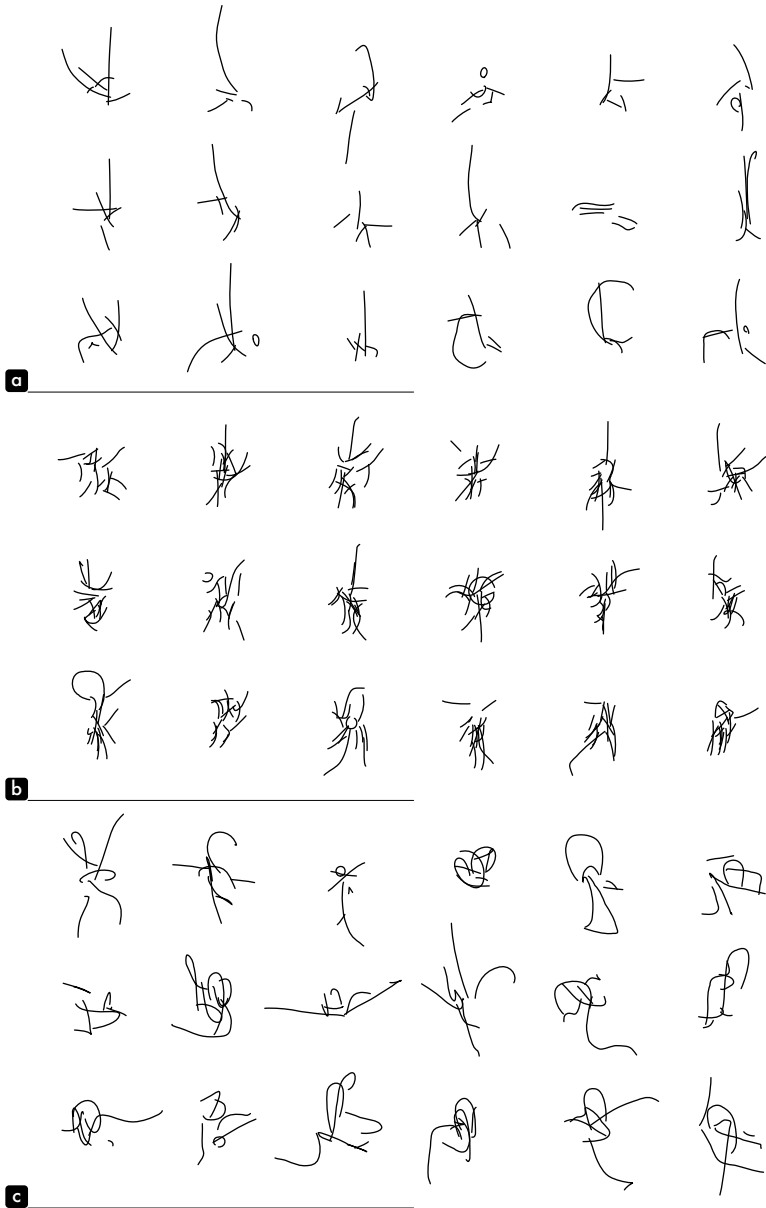


Figure 4.19: Generation of compositions from the compositional plane model, family explorations.



## 4 Model results and tools

from the dataset, but they are still pertinent to me. The only detail that could be perceived as an issue is the discrete nature of the number of components per cubic Bézier curves. In some location of the latent space, sharp frontiers then necessarily arise and provoke dramatic visual changes. For instance, in Fig.4.15c exploring local samples around a chosen  $z$ , strokes with a second component seem to belong to a different family, even if the first component of each stroke is similar. This is an open issue, mainly dependent on the chosen stroke parameterization.

The compositional latent space also appears rich and expressive (see Fig.4.16 and Fig.4.17). Nonetheless, smaller strokes of these compositions are sometimes grouped in little regions, and look messy. This is presumably an artifact of model indecision. It resembles to the *infamous* blurry outputs of most pixel-based VAEs. This effect is said to be due to the asymmetry in the  $D_{KL}$ , as described in the previous chapter. The *maximum likelihood* objective chosen in VAEs tends to over-generalize data density, rather than to focus on regions of maximum intensity (see Footnote.22, p.114).

Finally, generated compositions from the compositional plane model generally present *sharper* compositional patterns (see Fig.4.18). They seem to offer more pronounced compositional targets. At the same time, novel compositions appear more surprising to me than examples obtained with the composition model. This is particularly the case for the family explorations in Fig.4.19. Improvisations around a given  $z$  seem more diverse, with looser constraints. The additional degrees of freedom allowed by the mixture distribution enable arrangements beyond my compositional habits, while remaining weirdly familiar. Nonetheless, the *stroke repetition* artifact (particularly in Fig.4.19b) and some extravagance in longer strokes (see Fig.4.19c) provoke qualitative interferences, rendering this model somehow deceptive. As already stated, these concerns leave room for interesting future works.

### 4.2 Interpolation

In order to explore the latent space and experience the newly created continuity between strokes and compositions, we need to find a correct way to travel through these dimensions. A path is usually defined as the displacement from a point  $a$  to a point  $b$ . In a vectorial space, we designate such trajectory as an *interpolation*. We are manipulating spaces with a high number of dimensions, so the shortest path between  $a$  and  $b$  may not be optimal. High-dimensionality implies counter-intuitive effects on distances, areas, and volumes. So, this section describes some situations to manipulate with caution, and tries to make them visually understandable. Real practical issues due to high-dimensionality, and the so-called *curse of dimensionality*, will be tackled in the next chapter (see Section.5.1).

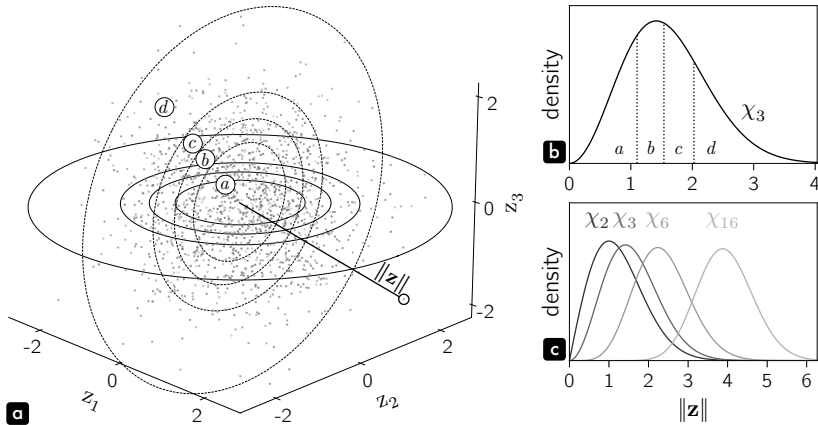


Figure 4.20: Hyperspace density. Panel **a**, shows 1600  $\mathbf{z}$  samples from a 3-d standard normal distribution. Clusters  $a$ ,  $b$ ,  $c$ ,  $d$  are composed of an equal number of points ordered by their norm. Panel **b** plots the  $\chi_3$  distribution of  $\|\mathbf{z}\|$ . Panel **c** displays  $\chi_K$  distributions for multiple  $K$  values.

### Hyperspace density

Before introducing specific characteristics of hyperspaces, we begin with a 3-dimensional space, easier to represent visually. Fig.4.20a shows 1600 dots randomly sampled from a 3-d standard normal distribution, and the point cloud seems denser around  $\mathbf{0}$ . Each component is independent, so it is expected that most of the points are located in the center of the space. However, if we make four clusters ( $a$ ,  $b$ ,  $c$ ,  $d$ ) with an equal number of points ordered by their norm, i.e. the distance from the center, we observe that these 3-d rings are of varying thickness. Clusters  $a$  and  $d$  are wider than  $b$  and  $c$ . It therefore indicates that the space must be denser at some periphery from the center. The distribution of the norm  $\|\mathbf{z}\|$  of a  $K$ -dimensional standard normal distribution follows a  $\chi_K$  distribution. In Fig.4.20b, the probability density function of  $\chi_3$  is plotted with the boundaries of the four clusters. The maximum density, i.e. the mode, is close to the limit between  $b$  and  $c$ . This mode is actually easy to compute by  $\text{mode}(\chi_K) = \sqrt{K-1}$ , with  $K$  the dimensionality of  $\mathbf{z}$ . Then, it is expected that the maximum density shifts outward as  $K$  increases (see Fig.4.20c).

This observation seems paradoxical. On the one hand, the densest location of the space is at the center because in each dimension  $z_i$  the standard normal distribution is concentrated around 0. On the other hand, if we look at  $\|\mathbf{z}\|$ , the densest region appears to be located on a sphere of radius  $\text{mode}(\chi_K)$ . How to reconcile these two realities? Which one is more pertinent to characterize the latent space?

One thing to know about continuous probability density functions is that they

#### 4 Model results and tools

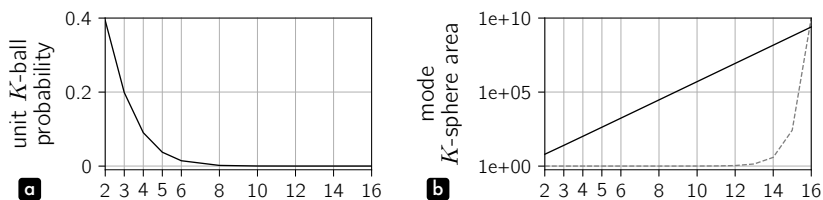


Figure 4.21: Counter-intuitive hyperspace characteristics. Panel **a** shows the probability of  $p(\|\mathbf{z}\| \leq 1.0)$  for different  $K$ , i.e. the probability of the unit centered  $K$ -ball. Panel **b** shows the surface of the  $K$ -sphere of radius  $\text{mode}(\chi_K)$  in log-scale for different  $K$ . The dotted line is an illustration of the same curve in non log-scale.

should not be evaluated in one point. For instance, looking back at Fig.4.20b, the probability that  $p(\|\mathbf{z}\| = 1.0)$  is actually null. The probability is zero, because it is very unlikely that  $\|\mathbf{z}\|$  would be exactly 1.000...000. So, probabilities have to be evaluated in some interval, e.g.  $p(\|\mathbf{z}\| \in [0.99, 1.01])$ . With this in mind, let us compute the probability of  $p(\|\mathbf{z}\| \leq 1.0)$ . Fig.4.21a plots this probability for different  $K$ . Even if this range of value seems to include samples quite far from the center, we remark that with  $k = 3$ , the probability is already of only  $\approx 0.2$ . With  $K = 8$ , the probability that a sample  $\mathbf{z}$  appears in the unit  $K$ -ball is almost null. As a result, even if the center of the space is the most likely location to be sampled, as the dimensionality increases, the local neighboring volume around it becomes exponentially insignificant at the scale of the whole space.

Furthermore, in Section.1.3 introducing the probabilistic space (see particularly Fig.1.10), we have seen that different compositions may globally have the same probability, the same relevance in the compositional space, but for different reasons. Individual dimensions can be alternatively close to the dull average, or explore more unlikely/expressive aspects. Thus,  $\|\mathbf{z}\|$ , in its quality of distance from the center, appears to be a good summary of the global informational content of a composition, no matter the exact expressed regularities. Then, the shape of the distribution of  $\|\mathbf{z}\|$  is telling us that compositional attributes cannot be simultaneously all close to the mean or very far from it (see Fig.4.20b). A composition is therefore, most of the time, a balanced mix of regularity and surprise.

In high dimensions, we therefore believe that the concept of *most likely location* should disappear in favor of *most likely surface*. Once a  $\|\mathbf{z}\|$  is fixed, we can draw a hypersphere (a circle if  $K = 2$  and a sphere if  $K = 3$ ), where compositions on this surface are equiprobable. Since the central volume *shrinks* as  $K$  increases, the *most likely surface* grows exponentially (see dotted line of Fig.4.21b). Visualizing these values in a log-scale (solid line), we notice that the surface goes from  $\approx 10$  to  $10^{10}$  as  $K$  ranges from 2 to 16. The probabilistic space offered by a model in high dimensions is therefore far from being normative. There are many (many) ways to be expressive, while still fulfilling standard statistics.

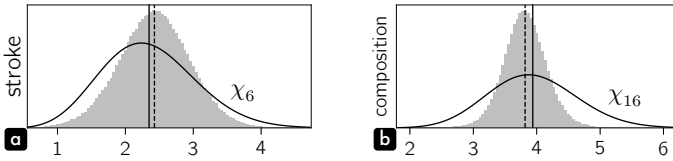


Figure 4.22: Panel **a** (for the stroke model) and panel **b** (for the composition model) show  $\|\mathbf{z}\|$  distributions with gray histograms. Expected  $\chi_6$  (stroke) and  $\chi_{16}$  (composition) are plotted with a black line. Vertical lines compare the estimated mean (dotted line) with the theoretical mean (solid line).

We have checked the actual distribution of stroke and composition latent space individual dimensions in Fig.4.12 and Fig.4.13. In Fig.4.22, we verify if the norm of each  $\mathbf{z}$  fits their respective  $\chi_K$  distribution. For the stroke model, the variance is slightly tighter than the expected density. Nonetheless, both mean values are very close from each other (see Fig.4.22a). Concerning the composition model (Fig.4.22b), mean values are also well aligned, but the peak is much higher than expected. It indicates that compositions are even more squeezed around the mode hypersphere described above. This behavior may be explained by the bimodal distributions of some individual dimensions presented in the previous section. However, despite this deviation from the expected prior, the model remains highly functional.

### Linear interpolation

The objective of any type of interpolation is to define intermediary positions between a starting point  $\mathbf{z}_a$  and an ending point  $\mathbf{z}_b$ . A straight line is for instance the shortest path between two points, no matter the number of dimensions. Beyond its simplicity, it remains a customary choice. Mathematically speaking, linear interpolation, *lerp*, is supported by:

$$\text{lerp}(\mathbf{z}_a, \mathbf{z}_b, u) = (1 - u)\mathbf{z}_a + u\mathbf{z}_b, \quad u \in [0, 1] \quad (4.1)$$

with  $u$  the interpolation parameter.  $u$  is ranging from 0 to 1 and the *lerp* corresponds to  $\mathbf{z}_a$  at  $u = 0$  and  $\mathbf{z}_b$  at  $u = 1$ . In most cases, we choose  $u$  as a ramp of equally spaced values. However, this constant *speed* in the parameter space  $u$  does not guarantee a constant perceptual change of model outputs. This is particularly true in our probabilistic latent space. In Fig.4.23a, we see the straight line constructed between two points in a 3-d space. The trajectory is visualized through the 3 planar projections. In the second panel, we notice that the interpolation is also straight in each dimension of  $\mathbf{z}$ . Intermediary points go through the density contained in between  $\mathbf{z}_a$  and  $\mathbf{z}_b$ . However, concerning the norm, the trajectory is bent toward 0, toward the center, i.e. through a very low density area. In generative models, this transit close to the center results in

#### 4 Model results and tools

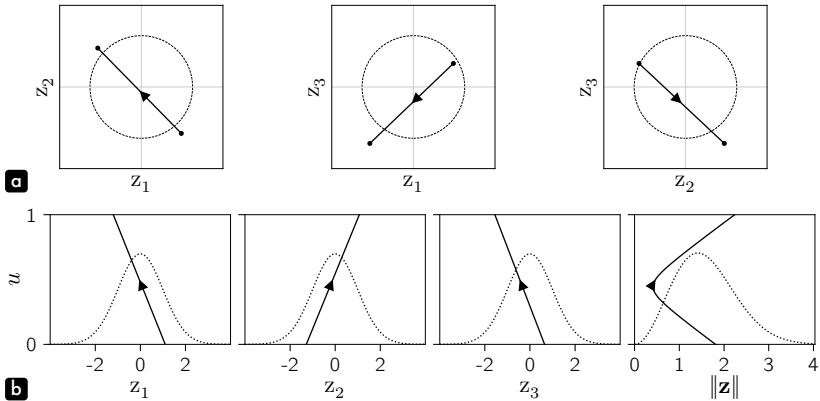


Figure 4.23: Linear interpolation between two points in a 3-d space. In panel **a**, the trajectory is visualized through the 3 planar orthogonal projections. The dotted circles indicate the sphere of radius  $\text{mode}(\chi_3)$ . Panel **b** shows the same trajectory per dimension and for the norm, with  $u$  as the y-axis. Dotted lines indicate respective densities the interpolation passes through.

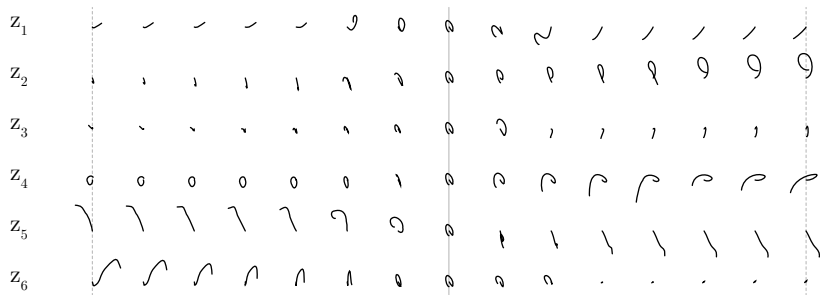


Figure 4.24: Linear interpolations per dimension of the stroke model. Each interpolation is computed from  $-\text{mode}(\chi_6)$  to  $\text{mode}(\chi_6)$  for the concerned dimension, while others are fixed to zero. The gray line highlights the stroke  $\mathbf{z} = \mathbf{0}$ , and dotted lines indicate strokes where  $\|\mathbf{z}\| = \text{mode}(\chi_6)$ .

poorly qualitative intermediary samples. This phenomenon is particularly visible in high-dimensional latent space, and it has been frequently observed empirically.

To illustrate this effect in Fig.4.24 for the stroke model, we compute linear interpolations along each dimension individually, others set to zero. The generated sequences then cover the expressiveness of each dimension from negative to positive values. While extremities are presenting some sort of visual opposition<sup>2</sup> (vertical dotted lines), all dimensions meet around zero with the same stroke (vertical gray

<sup>2</sup>We could describe  $z_1$  and  $z_5$  as controlling the stroke direction as a mix of 4 cardinal points.  $z_6$  seems to encode the length of the strokes, while other dimensions could express different curling attributes.

line). This curly stroke seems too complex to be the most probable one. Strokes from the extremities, with a norm close to  $\text{mode}(\chi_6)$ , look more standard or more likely. Therefore, in order to improve interpolations and trajectories in the latent space, we need to keep a more constant norm, around the mode hypersphere.

### Spherical linear interpolation

Even if the linear interpolation issue has been reported several times in the machine learning community, alternative techniques are not very popular. For instance, the use of spherical linear interpolation, *slerp*, has been lately formalized<sup>3</sup>, and is still not recognized as a best practice. Spherical linear interpolation is though an old and well-known tool in the computer graphics field<sup>4</sup>. It has been employed together with quaternions in 3-d animation to guarantee a constant angular speed during interpolation of rotations. By interpolating along the great arc, i.e. the shortest path on a sphere between two points, it helps to render natural rotations between body poses.

In the case of generative models, this type of interpolation on a sphere is interesting to keep  $\|\mathbf{z}\|$  density as constant as possible along the path. Staying on the surface of the hypersphere of a radius in between  $\|\mathbf{z}_a\|$  and  $\|\mathbf{z}_b\|$  should guarantee perceptually smooth interpolations. Nonetheless, spherical linear interpolation has been initially designed in a unit 3-d space. So, let us consider  $\mathbf{z}$  norm separately. Defining  $\hat{\mathbf{z}}_a = \frac{\mathbf{z}_a}{\|\mathbf{z}_a\|}$  and the angle between two vectors as  $\theta = \cos^{-1}(\hat{\mathbf{z}}_a \cdot \hat{\mathbf{z}}_b)$ , we have:

$$\mathbf{z}(\hat{u}) = \text{slerp}(\mathbf{z}_a, \mathbf{z}_b, u) = \frac{\sin((1-u)\theta)}{\sin(\theta)} \hat{\mathbf{z}}_a + \frac{\sin(u\theta)}{\sin(\theta)} \hat{\mathbf{z}}_b, \quad u \in [0, 1] \quad (4.2)$$

with  $\sin(\theta) \neq 0$  implying that  $\|\mathbf{z}_a\|$  and  $\|\mathbf{z}_b\|$  cannot be collinear.

Even if  $\mathbf{z}_a$  and  $\mathbf{z}_b$  have norms around  $\text{mode}(\chi_K)$ , they are usually different. We could simply operate a linear interpolation, but a better objective is to ensure constant density changes of the norm along the interpolation. We argue that this should theoretically produce a better perceptual continuity of the generated samples. We will experimentally explore this hypothesis in the next chapter. Concretely, the interpolation of the norm must be linear in the cumulative density function space.

$$\|\mathbf{z}(u)\| = \text{CDF}_{\chi_K}^{-1} \left( \text{lerp}(\text{CDF}_{\chi_K}(\|\mathbf{z}_a\|), \text{CDF}_{\chi_K}(\|\mathbf{z}_b\|), u) \right) \quad (4.3)$$

Finally,  $\mathbf{z}(u) = \|\mathbf{z}(u)\| \hat{\mathbf{z}}(u)$ . The result of this procedure is presented in Fig.4.25. In the first panel, we can see that trajectories are curved around the circles. Looking at individual dimensions in the second panel, paths are now bent, and can locally pass through a lower/higher density than the extremities  $\mathbf{z}_a$ ,  $\mathbf{z}_b$  (see

<sup>3</sup>White, 2016.

<sup>4</sup>Shoemake, 1985.

#### 4 Model results and tools

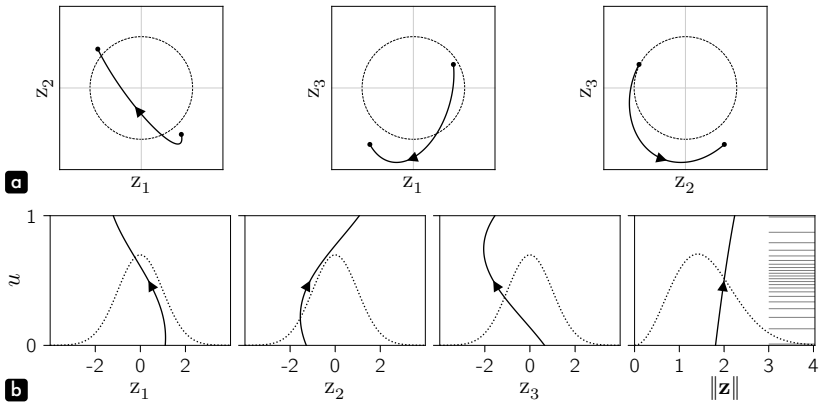


Figure 4.25: Spherical linear interpolation between two points in a 3-d space. In panel **a**, the trajectory is visualized through the 3 planar orthogonal projections. The dotted circles indicate the sphere of radius  $\text{mode}(\chi_3)$ . Panel **b** shows the same trajectory per dimension and for the norm, with  $u$  as the y-axis. Dotted lines indicate respective densities the interpolation is passing through. On the right of the sub-plot dedicated to  $\|z\|$ , the spread of the horizontal lines depicts the non-uniformity of the spacing of the points along the path.

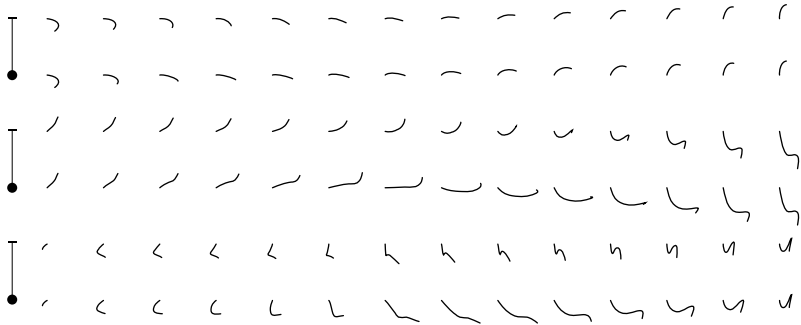


Figure 4.26: *lerp* vs *slerp* stroke interpolations. Three series of interpolation are presented for either a linear interpolation (rows beginning with a dash) or a spherical linear interpolation (rows beginning with a dot).

particularly the path of  $z_3$ ). Nonetheless, and more importantly, the norm of  $z$  is now contained in between  $\|z_a\|$  and  $\|z_b\|$ , with an almost linear trajectory. On the right of this plot, the spread of the horizontal lines depicts the non-uniformity of the spacing of the points along the path, that guarantee a constant change of density.

Besides this theoretical illustration, we can display the differences between *lerp* and *slerp* interpolations on real data. With strokes in Fig.4.26, improvements seems limited. *slerp* sequences (rows indicated by a dot) just present a better preservation

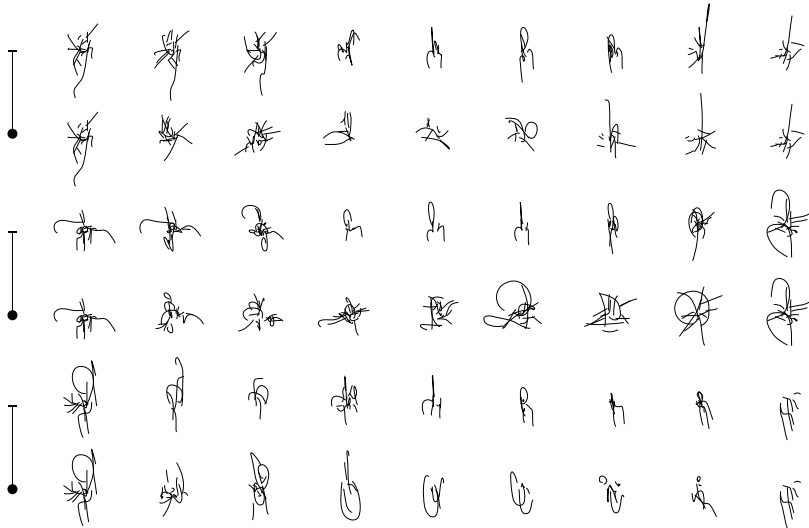


Figure 4.27: *lerp* vs *slerp* composition interpolations. Three series of interpolation are presented for either a linear interpolation (rows beginning with a dash) or a spherical linear interpolation (rows beginning with a dot).

of the length of the strokes in the middle region than *lerp* sequences (rows indicated by a dash). Concerning the composition model in Fig.4.27, interpolations between random samples would require more intermediary positions to display very smooth transitions. Indeed, we will demonstrate in the next chapter that two random samples in high dimensions are almost always orthogonal, i.e. far from each other. The *slerp* sequences are therefore passing through diverse compositions, without an obvious logic. Nevertheless, *slerp* sequences are more coherent in the central region compared to *lerp* sequences, where intermediary compositions almost look alike.

### Quad-interpolation

In order to make the diversity and the complexity of the latent space more tangible, we propose a *quad-interpolation* representation between 4 samples in Fig.4.28 and Fig.4.29. We believe that the generated surfaces of transformation reproduce the more familiar representation of a map, and how we usually describe a space. The underlying procedure is simple. Two *slerp* are computed between the corners of opposite sides of a square. Then, multiple *slerp* are operated in between corresponding intermediary samples of each side. However, depending on the ordering of the four initial points, the resulting surface can correspond to a twisted



#### 4 Model results and tools

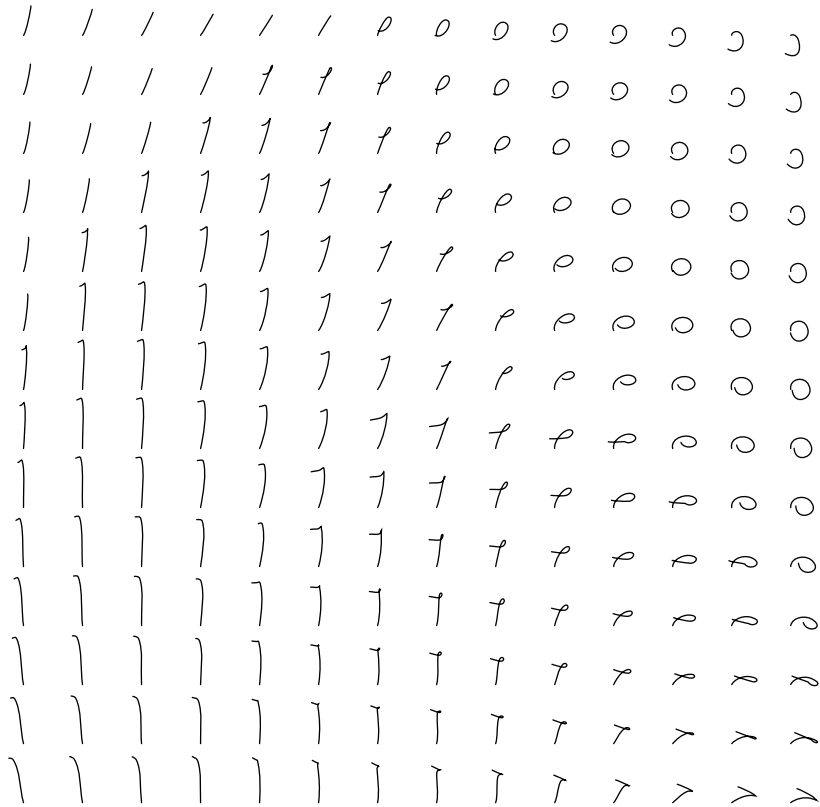
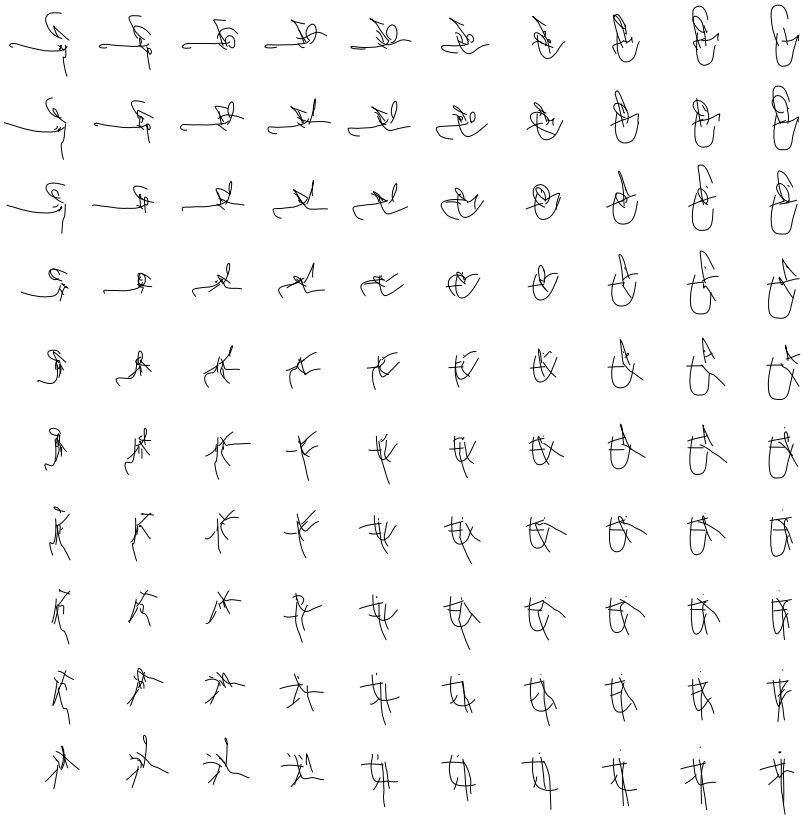


Figure 4.28: *sterp* quad-interpolation of strokes.

area in the latent space, resembling to a bow tie. Such arrangement would produce visually spurious folded quad-interpolations. As a result, we have built a procedure that reorder corners, so that the sum of the four external angles/arcs is minimal.

In the previous section, we have evoked the existence of some discrete frontiers in the stroke latent space due to changes in the number components per cubic Bézier curves. In top left corner of the Fig.4.28, this phenomenon is particularly noticeable. In Fig.4.29, the quad-interpolation representation helps to grasp the smooth nature of the compositional latent space, and to find a logic in the transitions. The selected instance may also be particularly *lucky*, with corners not too far from each other.

Figure 4.29: *slerp* quad-interpolation of compositions.

## 4.3 Measurements

One goal of composition modeling is to permit objective measurements of associated artistic perceptual characteristics. We want to go beyond expert/non-expert qualitative judgments, and further simple spatial metrics. Several image statistics exist, such as the amplitude spectrum slope: they may be informative, yet they are not specifically crafted for artistic or compositional purposes. This inquiry does not sound modest, but we actually consider *measurement* in its simplest sense, i.e. a quantitative value enabling comparison to a norm, or to other measurements. Establishing a meaning for this value, or finding an explanation coherent with a physical scale, is a different issue that can be addressed independently. We also claim that our proposed measurements are objective because the measuring tools

## 4 Model results and tools

are extracted autonomously, with machine learning procedures from the studied material itself. *Objective* is therefore different from being *universal*. Our findings are certainly not applicable to every sort of art, but express at least an objective vision of the chosen dataset, i.e. my personal drawings.

Our models propose several quantitative possibilities, located in the encoded latent space and the probabilistic outputs. They rely on two kinds of measurements: relative positions in multidimensional spaces, and likelihoods of individual or global pictorial events. This scrutiny can finally be applied at different scales, from components of a stroke, strokes in a context, to whole compositions.

Nevertheless, this section is still a showcase of different opportunities given by the different models. The main efforts of this research project were focused on producing theoretical groundings for these models and making them functional. So, even if some aspects of the compositional latent space will be discussed in the next chapter – particularly a successful use of some available measurements to predict human perceptual behaviors – a huge amount of possible researches and investigations of the proposed tools are left for future works. We have to acknowledge that the presented framework is still at its early stage.

### *Latent space measurements*

When a stroke or a composition is encoded by  $Q$ , we obtain two vectors  $\mu_\phi$  and  $\sigma_\phi^2$ . These parameters define a normal distribution for each of the  $K$  dimensions. Three examples ( $a$ ,  $b$  and  $c$ ) are plotted for each model in Fig.4.30a and Fig.4.31a. In both case, the selected strokes and compositions display a large diversity of distributions, in terms of spread and location.

The most interesting aspect of the available measurements is maybe supported by  $\mu_\phi$  and its position among the global latent space distribution. We have set the prior  $p(\mathbf{z})$  to follow a multivariate standard normal, so we can study the likelihood of a given stroke embedding  $\mathbf{z}$  in each dimension. As already explained, a more probable stroke is not better than others. There is no esthetic judgment, or claim of a higher value. This scrutiny is just a way to identify the expressive characteristics of a stroke. The distance from the norm, in a positive or negative numerical manner is precisely a sign of uniqueness. In addition, examples from Fig.4.30a and Fig.4.31a behave as predicted. Each stroke or composition is in the norm for some dimensions and unlikely in others, but none of these inputs are completely close to zero or far from it in every dimension.

This procedure tends to position a stroke or a composition among the whole space, but we can also compare entries one another. We should be able to answer quantitatively: to what extent inputs are similar? and in what aspect they are different? We focus on strokes examples as they are more directly interpretable. For instance, there is a clear ordering of strokes along fourth and sixth dimensions

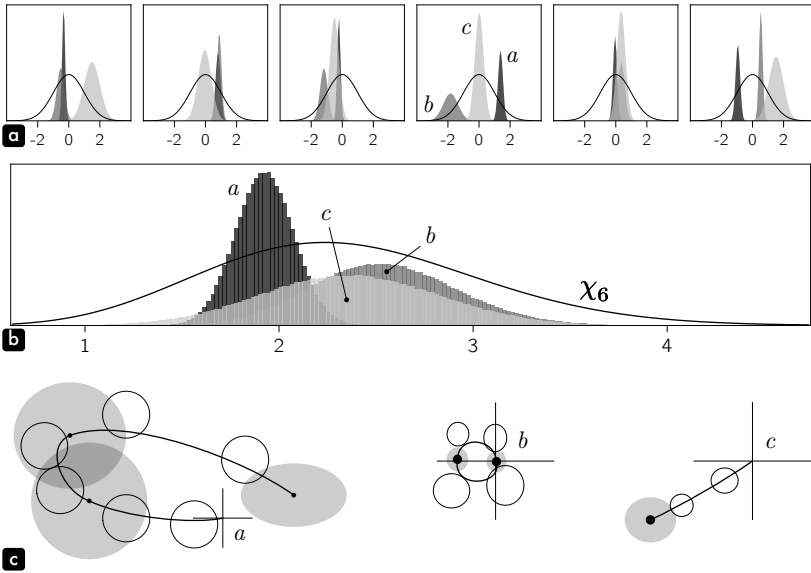


Figure 4.30: Stroke model available measurements through 3 examples *a*, *b* and *c*. Panel **a** shows latent space individual densities encoded by  $Q$  (in gray) over the prior standard normal distributions (black lines). Panel **b** plots  $\|\mathbf{z}\|$  densities of each stroke over the prior  $\chi_6$ . For visualization purposes, vertical scales are adjusted per stroke. Panel **c** displays the spatial prediction of the control points of each component of the cubic Bézier curves (for target points in gray and tangents with a black outline). Delimited areas enclose 50% of each cumulative density. Crosses indicate the origin and the relative scale of each stroke.

(Fig.4.30a). If we look back at Fig.4.24 showing individual dimension characteristics, we notice that  $z_4$  encodes what appears to be two types of curve closure. The straight stroke *c* thus lies in between the other two (Fig.4.30c). We have also qualified  $z_6$  as coherent with stroke length. This dimension presents strokes *b* and *c* closer to each other, and *a* with an opposite polarity (thin crosses in Fig.4.30c indicate the origin and the relative scale of each stroke). On the other hand,  $z_5$  seems irrelevant to discriminate between the 3 strokes.

However, the relative measure of similarity between entries is for now only related to the model recognition space. Even if we hope that there are connections with human perceptual system and inner metrics, there is no evidence at this stage. Compositional latent space relative similarity measurements is precisely what we will investigate in the next chapter. We will define the smoothness of a latent representation as a local perceptual homogeneity of similarity judgments (see Chapter.5 for further discussions).

Mean  $\mu_\phi$  is usually interpreted as the *real* location of an entry in the latent space. It is for instance this embedding that is given by the stroke model to the

#### 4 Model results and tools

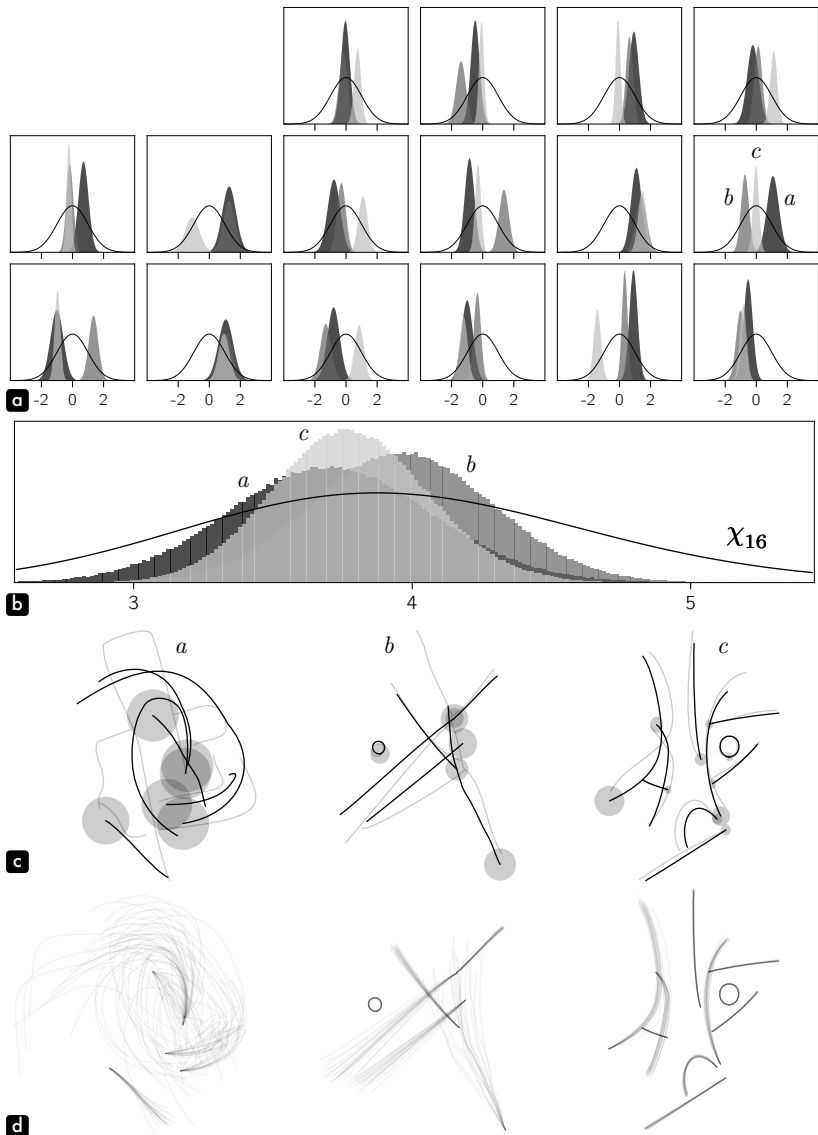


Figure 4.31: Composition model available measurements through 3 examples *a*, *b* and *c*. Panel **a** shows latent space individual densities encoded by  $Q$  (in gray) over the prior standard normal distributions (black lines). Panel **b** plots  $\|z\|$  densities of each composition over the prior  $\chi_{16}$ . For visualization purposes, vertical scales are adjusted per composition. Panel **c** displays the spatial prediction of stroke initial points (gray surfaces). Delimited areas enclose 50% of each cumulative density. Panel **d** illustrates the densities of predicted stroke shapes. The variance is reduced by 4 for legibility.

composition model. On the other hand, the variance  $\sigma_\phi^2$  could be interpreted as model  $Q$  uncertainty on its encoding per dimension. For instance, in stroke examples of Fig.4.30a, we remark that stroke  $c$  globally shows a larger variance than stroke  $a$ . It seems somehow counter-intuitive as shape  $c$  is more common than  $a$  (see Fig.4.30c). We could expect  $Q$  to have more difficulties in identifying stroke  $a$ , but the opposite is happening. In fact, if the nature of stroke  $c$  is more likely, then shape family  $c$  is certainly occupying a larger portion of the latent space. In other words, the encoder  $Q$  may have learned to discriminate more subtle morphological characteristics, and possibly with less confidence.

### $\|\mathbf{z}\|$ norm measurements

There is no closed form of the distribution of the norm of  $\mathbf{z}$  given  $\mu_\phi$  and  $\sigma_\phi^2$ . Nevertheless, it is possible to compute it by simulation. Fig.4.30b and Fig.4.31b present the resulting densities for stroke and composition examples  $a$ ,  $b$  and  $c$ . The spread of the distributions are mostly related to  $\sigma_\phi^2$ , so it could be interpreted as an averaged measurement of the global uncertainty of  $Q$  on its encoding. In Fig.4.30b, the variance of stroke  $a$  is tighter than the one of stroke  $c$ , which is coherent with our former observation. Other than the spread of these distributions, it is difficult to interpret the location of their mode. All compositions in Fig.4.31b are for instance quite close from each other. A smaller norm is perhaps indicating a less confident (messier) output, such as composition  $a$  in Fig.4.31c. However, these observations may be more interesting than our current understanding of it, and require further investigations.

### Visualizing position probability maps

Other sources of measurement are located in model  $P$  outputs. Based on single multivariate normal distributions or mixture distributions, these measurements cover the positions of stroke Bézier control points, stroke initial points, and stroke shape embeddings.

Let us first detail how to visualize probabilities related to position, i.e.  $\delta_t$ ,  $\delta'_t$ ,  $\delta''_t$  for the stroke model, and  $\mathbf{p}_t$  for composition models. We basically evaluate output densities over a square and discrete map of uniformly spread  $N$  samples per dimension. It is in two dimensions, so it is reasonable to construct a grid with a decent precision: we used  $N = 10^3$  producing  $N^2 = 10^6$  evaluated positions. Predicted outputs are normal distributions, and they theoretically spread over an infinite area. Nonetheless, the visualization grid has to be bounded. We know that in average, over the whole dataset, model inputs are standardized. We can therefore expect that the overall output aggregate would also fit a 2-d standard normal distribution. As a result, we chose a covering range per dimension of

#### 4 Model results and tools

$[\text{CDF}_{\mathcal{N}}^{-1}(10^{-8}), \text{CDF}_{\mathcal{N}}^{-1}(1 - 10^{-8})]$ . We can then compute the area  $d^2$  covered by each sample as (we use  $\log$  probabilities for computational accuracy):

$$\begin{aligned} \log d^2 &= \log \left( \frac{\text{CDF}_{\mathcal{N}}^{-1}(1 - 10^{-8}) - \text{CDF}_{\mathcal{N}}^{-1}(10^{-8})}{N - 1} \right)^2 \\ &= 2 \log(-2 \text{CDF}_{\mathcal{N}}^{-1}(10^{-8})) - 2 \log(N - 1) \end{aligned} \quad (4.4)$$

$N - 1$  is used instead of  $N$  because we consider each sample as the center of their tiny square. The actual extents of the complete visualization map is therefore of one supplementary  $d$  in each dimension. Then, the probability of the area associated with a position, e.g.  $\mathbf{p}_t$ , can be approximated by:

$$\log \int_{d^2} p(\mathbf{p}_t) d_{d^2} \approx \log p(p_{t,x}) + \log p(p_{t,y}) + \log d^2 \quad (4.5)$$

Concerning the compositional plane model, outputs are based on a mixture distribution, so that stroke positions  $\mathbf{p}_t$  and shape embeddings  $\mathbf{s}_t$  are independent *only* knowing the mixture index  $m$ . To study  $\mathbf{p}_t$  individually, it must be marginalized as follows:

$$\begin{aligned} \log p(\mathbf{p}_t) &= \log \int_{\mathbf{s}_t} p(\mathbf{p}_t, \mathbf{s}_t) d_{\mathbf{s}_t} = \log \int_{\mathbf{s}_t} \sum_{m=1}^M p(m) p(\mathbf{p}_t | m) p(\mathbf{s}_t | m) d_{\mathbf{s}_t} \\ &= \log \sum_{m=1}^M p(m) p(\mathbf{p}_t | m) \int_{\mathbf{s}_t} p(\mathbf{s}_t | m) d_{\mathbf{s}_t} \\ &= \log \sum_{m=1}^M \exp \left( \log p(m) + \log p(\mathbf{p}_t | m) \right) \end{aligned} \quad (4.6)$$

leading to,

$$\log \int_{d^2} p(\mathbf{p}_t) d_{d^2} \approx \log \sum_{m=1}^M \exp \left( \log p(m) + \log p(\mathbf{p}_t | m) + \log d^2 \right) \quad (4.7)$$

This information is nonetheless only partially informative. It indicates probable areas of next strokes, but no matter these stroke shapes. A more constrained question is: where should be positioned the next stroke, *knowing* its shape? It is basically computed as the following conditional probability.

$$\log p(\mathbf{p}_t | \mathbf{s}_t) = \log \frac{p(\mathbf{p}_t, \mathbf{s}_t)}{p(\mathbf{s}_t)} = \log p(\mathbf{p}_t, \mathbf{s}_t) - \log p(\mathbf{s}_t) \quad (4.8)$$

where  $\log p(\mathbf{s}_t)$  follows a similar development as in Eq.4.6, so that:

$$\begin{aligned} \log \int_{d^2} p(\mathbf{p}_t | \mathbf{s}_t) d_{d^2} &\approx \log \sum_{m=1}^M \exp \left( \log p(m) + \log p(\mathbf{p}_t | m) + \log p(\mathbf{s}_t | m) \right) \\ &\quad + \log d^2 - \log \sum_{m=1}^M \exp \left( \log p(m) + \log p(\mathbf{s}_t | m) \right) \end{aligned} \quad (4.9)$$

In Fig.4.30c and Fig.4.31c, the visualization of these probabilities is operated by a contour map, because it is easier to display with vectorial graphics. We also decided to plot a unique isoline of  $CDF = 0.5$  for legibility. In Fig.4.32a,b and Fig.4.33a,b, more isolines are plotted and represent  $CDF \in [0.25, 0.5, 0.75]$ .

Probability maps of the stroke model (Fig.4.30c), are divided into target control points (gray areas) and tangent control points (black outlines). It appears that longer the stroke is and larger the predicted areas are. This observation is probably related to the lower likelihood of longer strokes (in terms of number of components), but it is also coherent with a strategic behavior for  $P$ , adapting its uncertainty to the scale of its actions. Another striking detail is the circular nature of almost all predictions. Only the last target point of  $a$  presents a spatial directionality, coherent with the main stroke movement. In fact, probability circles can only be squeezed vertically or horizontally, as output covariance matrices have been chosen diagonal with  $\Sigma_{\theta} = \text{diag}(\sigma_{\theta}^2)$ .

Concerning the composition model (Fig.4.31c), we remark that output variances are highly related to the latent uncertainty of the model and the associated global visual accuracy. Strokes are also reproduced from longer to shorter ones, which is reflected in the spread of the predicted areas, decreasing over time (particularly for composition  $c$ ).

For  $p(\mathbf{p}_t)$  of the compositional plane model (see Fig.4.32a and Fig.4.33a), it seems that the density gets more specified over time, with a greater shape complexity. It also changes slowly. Most probable areas are influenced by former strokes, but main spots appear quite constant. This behavior is completely different when a specific stroke shape conditions the prediction. With  $p(\mathbf{p}_t | \mathbf{s}_t)$  in Fig.4.32b and Fig.4.33b, only a subset of  $p(\mathbf{p}_t)$  is kept. Predicted strokes are almost restricted to one location, and probable position radically changes between time steps. However, predicted areas present a large variance, and the model seems to fail at really decreasing its output uncertainty over the training procedure.

### Visualizing stroke shape densities

The visualization of stroke shape densities is more challenging. Compared to spatial positions, stroke embeddings are not directly tangible. They have to be decoded to render visible strokes. Our idea is thus to use stroke opacity as a proxy for density. This way, overlapping strokes could naturally densify on common portions. Secondly, the number of samples to cover a  $K$ -dimensional space with a sufficient accuracy is growing exponentially with  $K$  (if  $N = 10^3$ , the number of evaluated samples is  $10^{18}$ ). A more efficient way to cover the space is therefore to directly pick  $N$  samples from the studied distribution instead of a uniform sampling of the space. As a result, each sample has an equal probability, and can be given



## 4 Model results and tools

the same opacity. In order to make all opacities sum to 1, we set each stroke opacity to  $\frac{1}{N}$ .

Concerning the composition model, we directly have access to  $p(\mathbf{s}_t | \mathbf{y}_{t-1}, \beta_{t,1})$  which is a normal distribution. However, for the compositional plane model, output distribution is a mixture and the sampling operation has to be operated in two steps. To be more explicit, let us recall the definition of the marginal  $p(\mathbf{s}_t) = \sum_{m=1}^M p(m)p(\mathbf{s}_t | m)$ . As we only know how to sample strokes from  $p(\mathbf{s}_t | m)$ , we first have to sample an index  $m$  from  $p(m)$ , that is a categorical distribution. Once an  $m$  is chosen, then we just have to sample one stroke from  $p(\mathbf{s}_t | m)$ . This procedure is repeated  $N$  times.

Nonetheless, the visualization of marginals are only partially informative. As stated previously, such measurement indicates probable next strokes, but no matter the location to draw it. Sampling strokes *knowing* a chosen spatial location is more interesting. Let us rearrange the definition of  $p(\mathbf{s}_t | \mathbf{p}_t)$ .

$$\begin{aligned} p(\mathbf{s}_t | \mathbf{p}_t) &= \frac{p(\mathbf{p}_t, \mathbf{s}_t)}{p(\mathbf{p}_t)} = \frac{\sum_{m=1}^M p(m)p(\mathbf{p}_t | m)p(\mathbf{s}_t | m)}{\sum_{g=1}^M p(g)p(\mathbf{p}_t | g)} \\ &= \sum_{m=1}^M \frac{p(m)p(\mathbf{p}_t | m)}{\sum_{g=1}^M p(g)p(\mathbf{p}_t | g)} p(\mathbf{s}_t | m) \end{aligned} \quad (4.10)$$

We can then consider that  $p(\mathbf{p}_t | m)$  acts as a fixed modifier of the weights of the initial  $p(m)$  categorical distribution, where  $\sum_{g=1}^M p(g)p(\mathbf{p}_t | g)$  just normalizes category probabilities to 1.

Concerning the results for the composition model (Fig.4.31d), the variance associated with stroke shapes is closely related to the location variance (Fig.4.31c). With this visualization, we also observe the cumulative effect of uncertainty along strokes. Long strokes (in linear length and number of components) possibly present dramatic changes in their visual appearance. Nonetheless, outputs are greedily sampled for this model, and they are usually well centered, so that overall results are satisfying.

For  $p(\mathbf{s}_t)$  of the compositional plane model (see Fig.4.32e and Fig.4.33e), predicted stroke densities are really fuzzy and difficult to interpret. Output variance is large, and the stroke spectrum to represent is too diverse for our visualization technique. We should therefore focus on  $p(\mathbf{s}_t | \mathbf{p}_t)$  in Fig.4.32d and Fig.4.33d. Conditioning the prediction on a specific location makes the measurement more readable. Nonetheless, large visual variability of longer strokes is still problematic, prompting us again to address the dataset lack of long stroke diversity in future works.

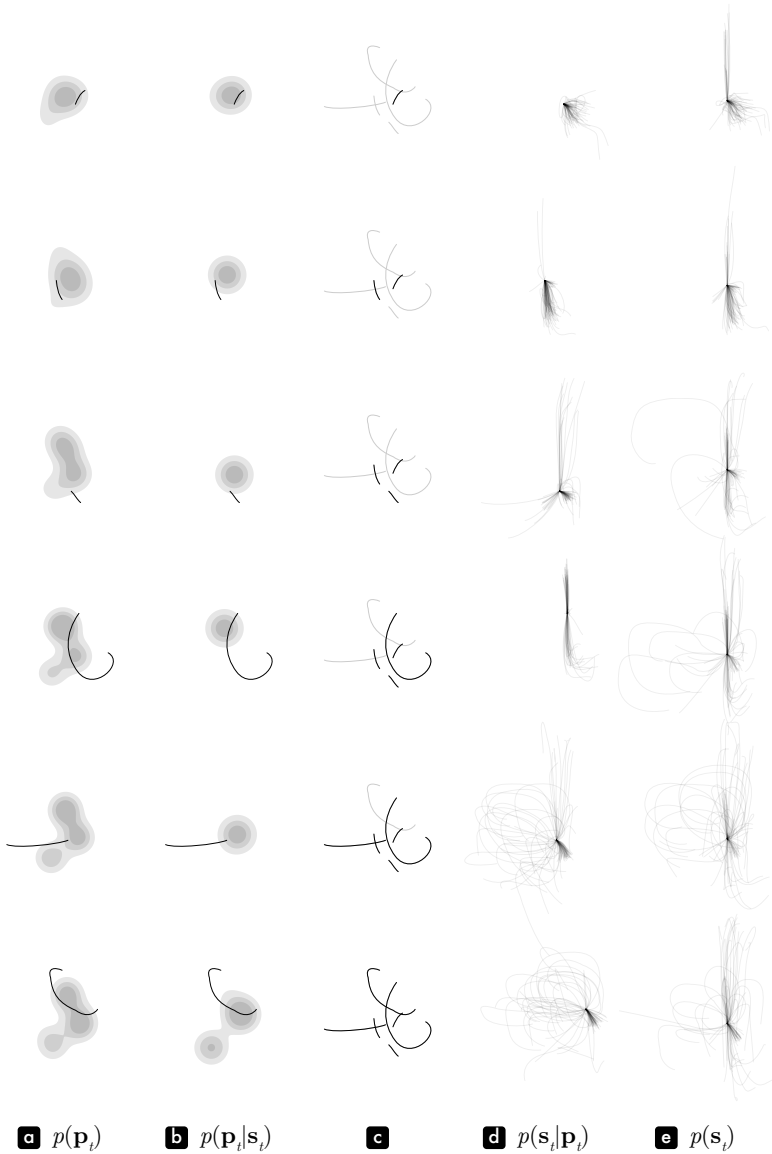


Figure 4.32: Compositional plane model measurements. Panel **c** vertical line shows successive states of a randomly generated composition. Panels **a** and **b** display the spatial prediction  $p(\mathbf{p}_t)$  and  $p(\mathbf{p}_t | \mathbf{s}_t)$  of stroke initial points (gray surfaces). Delimited areas enclose 25%, 50% or 75% of the cumulative density. Panel **d** and **e** illustrate the densities  $p(\mathbf{s}_t)$  and  $p(\mathbf{s}_t | \mathbf{p}_t)$  of predicted stroke shapes. The variance is reduced by 2 for legibility.

#### 4 Model results and tools

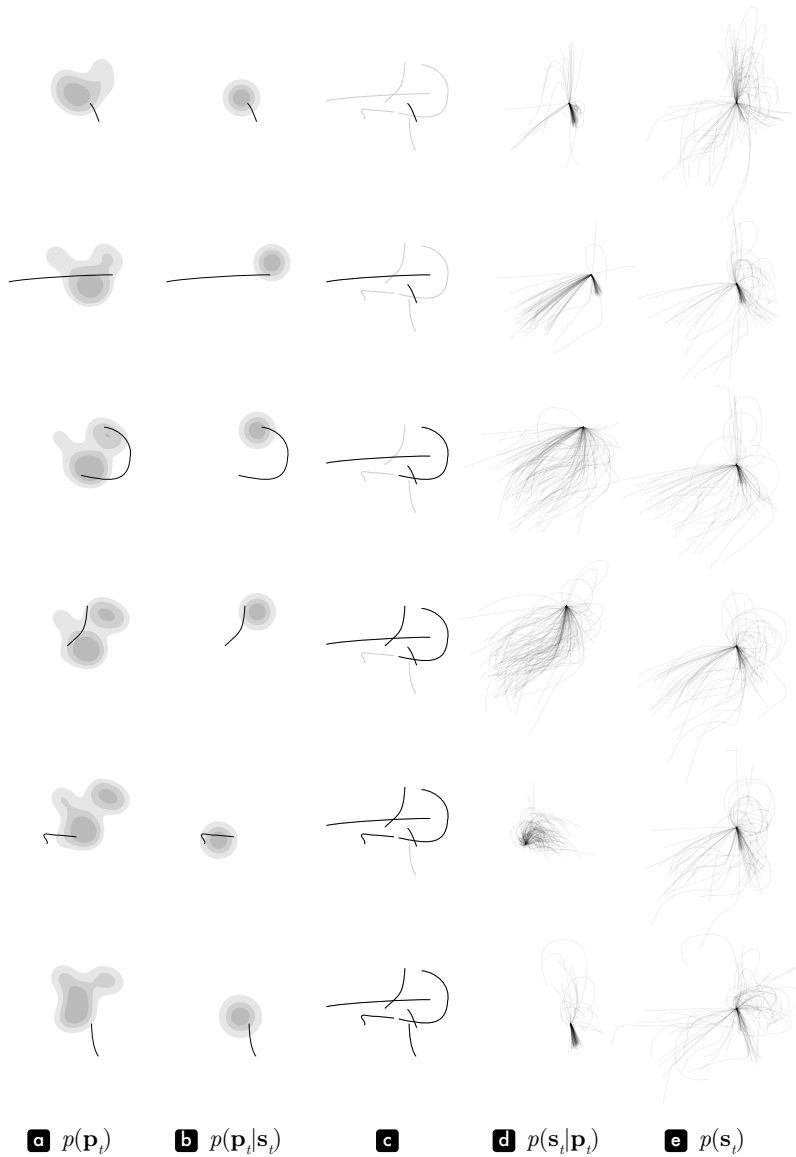


Figure 4.33: Compositional plane model measurements. Panel **c** vertical line shows successive states of a randomly generated composition. Panels **a** and **b** display the spatial prediction  $p(\mathbf{p}_t)$  and  $p(\mathbf{p}_t | \mathbf{s}_t)$  of stroke initial points (gray surfaces). Delimited areas enclose 25%, 50% or 75% of the cumulative density. Panel **d** and **e** illustrate the densities  $p(\mathbf{s}_t)$  and  $p(\mathbf{s}_t | \mathbf{p}_t)$  of predicted stroke shapes. The variance is reduced by 2 for legibility.

## 5 Composition perception

A generative model is a sandbox for experiments. The latent space and the associated decoder act as a *controlled* source of stimuli. The objective is then to study whether regularities captured by the model are in alignment with human perception. For instance, does the latent space present biases similar to those found in natural vision, e.g. preference for horizontal/vertical orientations or symmetries? If so, the following question is what does the model functional mechanisms tell us about pictorial composition in general? By simulation of a creative gesture, we can gain insights on our own internal compositional process. Nonetheless, at the timescale of this thesis project, we have to focus our effort on one simpler, yet meaningful aspect of the latent space, and demonstrate the scientific pertinence of our framework<sup>1</sup>.

As a representative, but preliminary experiment, we have decided to attest of a functional aspect of the model. Despite the monitoring of  $\hat{l}$ , enforcing an even and optimal overlap of  $z$  encodings of real dataset entries (see Section.3.5), we should investigate the perceptual efficiency of our custom training procedure. Basically, we want to address the idea of latent space *smoothness*. Conjointly, we would like to verify an assumption exposed in the previous chapter: that moving on trajectories of constant density in the latent space, e.g. on hyperspheres, permits more perceptually qualitative interpolations (and the dual idea that travelling along the norm causes significant perceptual distortions because of the density variations). In the machine learning community, interpolations in generative models is actually an aspect qualitatively commented very often, but rarely verified quantitatively. Our proposed method could therefore be generalizable outside the frame of our study on composition. Finally, interpolation is an essential tool to artistically experience continuity of the compositional space. Perceptually improving interpolations is thus simultaneously a creative objective.

Initially, we also wanted to study possible *lower frequency* distortions in the latent space. The question was whether there were silent dimensions which do have an effect on compositions from a structural point of view (necessary to describe/reconstruct drawings with a sufficient accuracy), but barely discriminated by viewers. However, such global *coherence* investigation requires an experimental design spanning the whole compositional space. In this search, we harshly faced the *curse of dimensionality*. Even if this objective has been finally discarded for

---

<sup>1</sup>This overall framework is indeed similar to our study on abstract composition orientation judgment (Lelièvre & Neri, 2021). See Appendix.A.1 for details.

## 5 Composition perception

practical reasons, we found it interesting to be described in the first section of this chapter. The second section is then dedicated to the theoretical background of perceptual scaling. We discuss chosen methods and detail our contributions. Finally, experimental results obtained with this psychophysical tool, and associated discussion, end this chapter on composition perception.

### 5.1 Dimensionality issues

In the previous chapter, we have described counter-intuitive phenomena in high-dimensional spaces. While prototyping experiments, we have been confronted to situations raising more practical issues. We consider these problems as an important drawback for an immediate and simple use of deep generative models. This inherent characteristic should be addressed and may be resolved in future work. Nonetheless, instead of jumping directly to operational protocols and their results, we have decided to detail the encountered limitations. We therefore review investigated methods and rely on simulations to demonstrate the current infeasibility of studies on the latent space at a global/complete scale. This exposition is also the occasion to introduce fundamental psychophysical concepts, such as *comparative judgments* and *thresholds*.

#### *Comparative judgments and angular distances*

Whether it is at a local scale, to check the perceptual quality of interpolations in the latent space, or at a global scale, to determine the visual effectiveness of the captured dimensions, we could sum up our experimental goal as quantifying distortions between the space discovered through machine learning, and our perceptual representation of corresponding compositions. The idea is basically to measure the perceptual dilations and compressions regarding the metric provided by the latent space. Our experimental tasks then turn out to rely on a *simple* appreciation of distances, i.e. a qualitative or quantitative comparative judgment between two positions. In particular, we investigate our ability to perceive differences between compositions, rather than our personal interest about compositional values. We find this type of judgment more objective and more easily interpretable for viewers compared to esthetic judgments, as it should involve less knowledge and culture-specific skills of art materials.

To conform to psychophysical concepts, we first need to define our studied *physical* space and a practical metric. Initially elaborated for elementary stimuli magnitudes, such as light intensity or sound pitch through air vibration wavelengths, we must understand *physical* in opposition to the psychological dimension under scrutiny. Therefore, even if our model latent space generates complex stimuli through a highly nonlinear process, it can still be considered as a physical space. However,

even if  $z$  values directly shape stimuli, Euclidean distances (the shortest line in between) are not an optimal metric. In Section.4.2, we have stated that travelling in the latent space, e.g. interpolations, has to be operated on hyperspheres. To go through the denser region of the latent space, this hypersphere even has to be of radius  $\text{mode}(\chi_{16})$ . So, physical distances between samples should be defined as vectorial angles. It naturally implies to be able to make controlled rotations in a high-dimensional space.

Rotation matrices are only well-defined in 2-d planes. In higher dimensions, we usually build *compositions* of rotation matrices operating alternatively on planar sub-spaces. A rotation matrix is an identity matrix except for the two selected dimension in rotation. For example, in a 3-d space, a rotation of angle  $\phi$  in the  $xy$  plane and along  $z$  axis, gives:

$$R_z(\phi) = \begin{vmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{vmatrix}, \quad R_z(90^\circ) \begin{vmatrix} 1 \\ 0 \\ 0 \end{vmatrix} = \begin{vmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} 1 \\ 0 \\ 0 \end{vmatrix} = \begin{vmatrix} 0 \\ 1 \\ 0 \end{vmatrix} \quad (5.1)$$

The problem is that if we want to execute a rotation in an arbitrary plane, we have to first apply a change of basis. Then, when desired rotations are completed, an inverse basis transformation is required to recover the original coordinate system. These steps are not straightforward in  $n$  dimensions, so we dropped the matrix transformation method and preferred an adaptation of the *slerp* equation (see Eq.4.2).

We define  $z_a$  as a reference sample on a hypersphere of radius  $\|z_a\|$ .  $z_a$  also sets the origin of angles  $\phi$ , controlling the applied rotations. Then,  $z_b$  is another sample, defining with  $z_a$ , the hyperplane of the rotation. Of course,  $z_a$  and  $z_b$  cannot be collinear. With  $\hat{z}_a = \frac{z_a}{\|z_a\|}$  and  $\theta = \cos^{-1}(\hat{z}_a \cdot \hat{z}_b)$  being the angle between the two vectors, we have:

$$z(\hat{\phi}) = \frac{\sin(\theta - \phi)}{\sin(\theta)} \hat{z}_a + \frac{\sin(\phi)}{\sin(\theta)} \hat{z}_b \quad (5.2)$$

By default, we set  $z(\phi) = \|z_a\| z(\hat{\phi})$ , but any norm definition could be applied, e.g. the procedure described in Eq.4.3.

Fig.5.1 shows two examples of such procedure along two orthogonal axes. Rotations are applied successively and construct a surface of transformation. In these 2-d maps, each sample is separated from its neighbors by  $2^\circ$ . As a result, the field of changes appears very smooth, so that it is rather difficult to spot local differences. Only travelling across illustration corners emphasizes morphological evolutions. This perceptual discriminative difficulty is precisely connected to the notion of *threshold*.

## 5 Composition perception

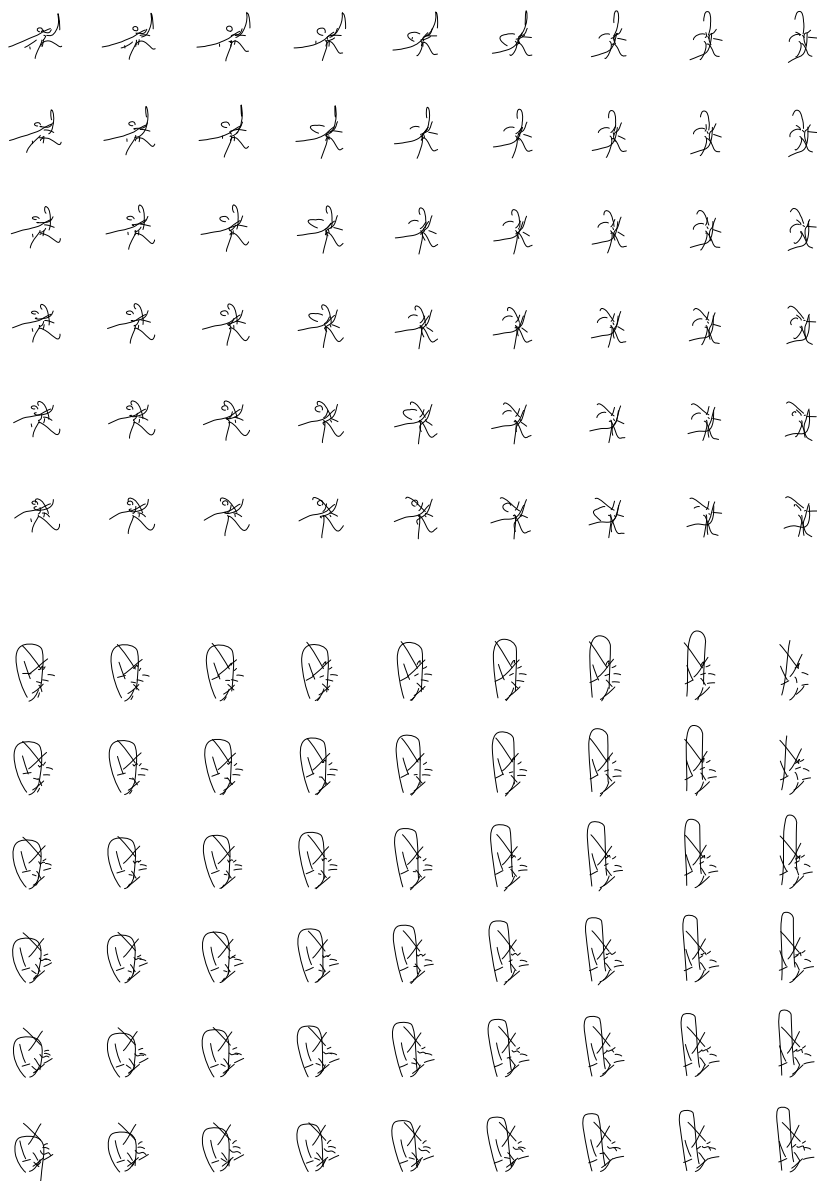


Figure 5.1: Successive rotations in the latent space along two orthogonal axes. For these two examples, each composition is separated from its neighbors by  $2^\circ$ .

## Threshold estimation

In broad lines, we know that we will ask participants to judge distances between compositions. But then, how to choose the appropriate angular distance between inspected compositions? To designed successful perceptual experiments, we should have an estimation of the default human abilities and cognitive limitations. Known as *thresholds*, these boundaries are defined as the sufficient intensity (for absolute threshold of detection tasks), or change in a stimulus characteristic (for difference threshold of discriminative tasks), to cause a significant perceptual response. In the discriminative case, the difference threshold is called the *just noticeable difference* or JND. This notion has been introduced by Weber and Fechner during the 19<sup>th</sup> century<sup>2</sup>. It was initially considered as an elementary psychological unit, and perception was interpreted as a succession of JND increments. However, repeated human judgments of the same stimuli vary time to time. Some random process seems to alter our perception and the discrete conception described above does not support satisfactorily intraindividual variabilities. Formulated under *signal detection theory*, it has been later proposed that “the presence of internal noise, or uncertainty, led to stimuli being represented in the brain not by a single point along a sensory continuum, but as a random sample drawn from a distribution.”<sup>3</sup> It implies that JND, as a unit, or an exact quantity, does not really exist. It has to be considered as an arbitrarily defined probabilistic value of a continuous process.

To illustrate this idea, it may be easier to describe a typical experiment designed to measure thresholds. For the ease of development, we will momentarily leave compositional angular difference aside, and consider a stimuli variable  $s$  with an arbitrary physical metric, for which we have an objective idea of what is *more*. In addition, thresholds are not supposed to be equivalent everywhere on this metric. So, we choose a reference stimuli  $s_0 = 0$ , for which the threshold is estimated.

The following procedure belongs to the 2AFC family (two-alternative forced choice). It means that participants are presented with two stimuli per trial and that, if we ask which stimulus is *more something*, the participant is requested to pick one or the other stimulus: *I don't know* is not an option. Then, two experimental strategies are possible. First, we can ask participants to compare  $s_0$  only once with a large number of different  $s$ . Responses are therefore only binary: 0 if  $s_0$  was chosen as *more*, and 1 otherwise. This is represented by small dots in Fig.5.2b. An alternative is to compare multiple times a reduced number of specific stimuli  $s_a$ . This way, we obtain the probability of each  $s_a$  to be perceived as *more*. It corresponds to empty circles in Fig.5.2b. Based on either data, we can fit a sigmoidal curve called a psychometric function<sup>4</sup>. The JND can then be defined

<sup>2</sup>Fechner, 1860.

<sup>3</sup>Kingdom and Prins, 2010, p. 154.

<sup>4</sup>Fitting is usually done by logistic regression, but it can be computed for several other sigmoidal parametric functions by Generalized Linear Model (GLM) or Maximum Likelihood Estimation (MLE). We will come back on these fitting procedures in the next section.



## 5 Composition perception

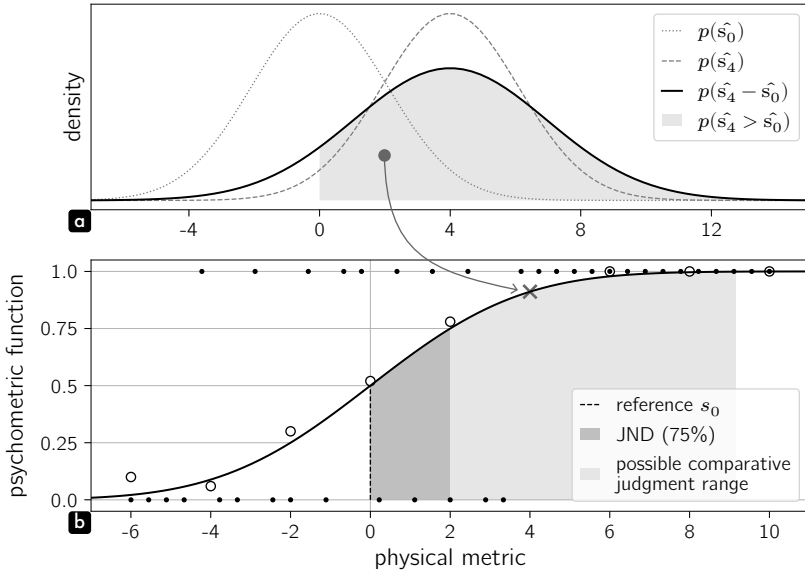


Figure 5.2: With  $s_0 = 0$  chosen as our reference for threshold measurement, panel **b** shows simulated responses of a participant performing a 2AFC task with a single (small dots) or multiple (empty circles) pair repetitions. The black line then depicts the fitted psychometric function. The corresponding difference threshold or 75% JND is plotted in dark gray. Adjacent light gray area materializes the possible comparative judgment range, with an upper limit set to 99.9% of the psychometric function. Finally, panel **a** illustrates the underlying stochastic process explaining the sigmoidal shape of psychometric function.

as a value expressed in physical unit, corresponding to a certain percentage of time that this value is seen as *more*. It is usually arbitrarily set to 75%, and in our example, the JND would be about 2.

Earlier, we have evoked the stochasticity associated with each stimulus, but at the psychometric function level, the underlying perceptual procedure is masked. So, let us explicitly add some noise to a stimulus  $s_a$  to obtain a random variable of the perceived stimulus such as:

$$\hat{s}_a = s_a + n_{s_a} \quad \text{with} \quad n_{s_a} \sim \mathcal{N}(0, \sigma_{s_a}^2) \quad (5.3)$$

Added noise is likely to be the result of many independent internal and external sources. Due to the central limit theorem, we can safely assume that this noise is normally distributed. In Fig.5.2a, we plot the density of two stimuli,  $p(\hat{s}_0)$  and  $p(\hat{s}_4)$  (dotted lines). Then, we want to know the proportion of time that  $\hat{s}_4$  is perceived as *more* than  $\hat{s}_0$ . We remark that distributions are largely overlapping, so this judgment is not binary. It corresponds to  $p(\hat{s}_4 > \hat{s}_0)$ , equivalent to  $p(\hat{s}_4 - \hat{s}_0 > 0)$ . Graphically, it can be represented by the gray area under the curve

of  $p(\hat{s}_4 - \hat{s}_0)$ , which is also normally distributed. In our example, this value is around 0.91. Reporting such value in the panel below for every possible stimulus finally reconstructs the psychometric function (see Fig.5.2b).

The transposition to our compositional angular differences is not straightforward. The first difficulty is that there is no objective *more* or *less* direction of angular changes. An alternative would be to use the *same-different* procedure, but it is subject to various kind of biases<sup>5</sup>. Secondly, with an indirect physical metric related to complex stimuli, it is likely that every composition will lead to slightly different thresholds. Then, we cannot parse the whole dataset to precisely measure every threshold. In addition, thresholds are very dependent on environmental conditions and chosen experimental variables, e.g. the duration of the stimuli presentation. In our case, longer you have access to pairs of composition and easier it is to spot subtle differences (e.g. for immediate neighbors in Fig.5.1). So, despite the existence of psychophysical methods, threshold measuring is too costly for our study. An empirical estimate of our ability to discriminate between compositional samples is sufficient to evaluate the feasibility of experimental designs beforehand. Even with a twofold error, our estimation would remain useful.

Compared pairs of compositions have to be significantly different to trigger a perceptual change. But at the same time, they should present a sufficient amount of common features to permit an informed judgment. We are therefore searching for a minimum and a maximum angle. As the JND for the lower bound, an upper limit can be arbitrarily set to a stimulus producing 99.9% positive responses. In Fig.5.2b, such range is materialized by the light gray area, with an upper bound in the physical space around 9.

We can roughly determine this range with Fig.5.3. Central elements of horizontal triplets stay unchanged, while lateral samples are rotated by the indicated angle  $\phi \in [1^\circ, 2^\circ, 4^\circ, 8^\circ, 16^\circ, 32^\circ]$ . Looking at these triplets *quickly* ( $\sim 2\text{sec}$ ), we could agree on the following operational values [a:(2,16), b:(2,8), c:(8,16), d:(1,4)]. As a result, we can coarsely state that possible judgments range from  $2^\circ$  to  $8^\circ$  with an optimal angular difference around  $4^\circ$ . Even if Fig.5.2 was intended to be illustrative only, values have been chosen in alignment with this estimation.

Finally, with this in mind, it is useful to know the distribution of angles between two random samples. Fig.5.4 shows the unsigned angle densities for different dimensionalities  $K$ . On a two-dimensional plane, all angles are equiprobable. With  $K = 16$ , any couples of random points are likely to be nearly orthogonal ( $90^\circ$ ). So, any naive experiment based on randomly selected samples is likely to fail. Such judgments would be mostly uninformative because of their very supra-threshold nature. Compared compositions would have, most of the time, nothing in common.

<sup>5</sup>A bias is basically a deviation from the expected theoretically balanced behavior by a participant. This bias may be specific to individual strategies, or widely shared among participants. In this latter case, the bias is likely to represent an optimal adaptation to the environmental prior.

## 5 Composition perception

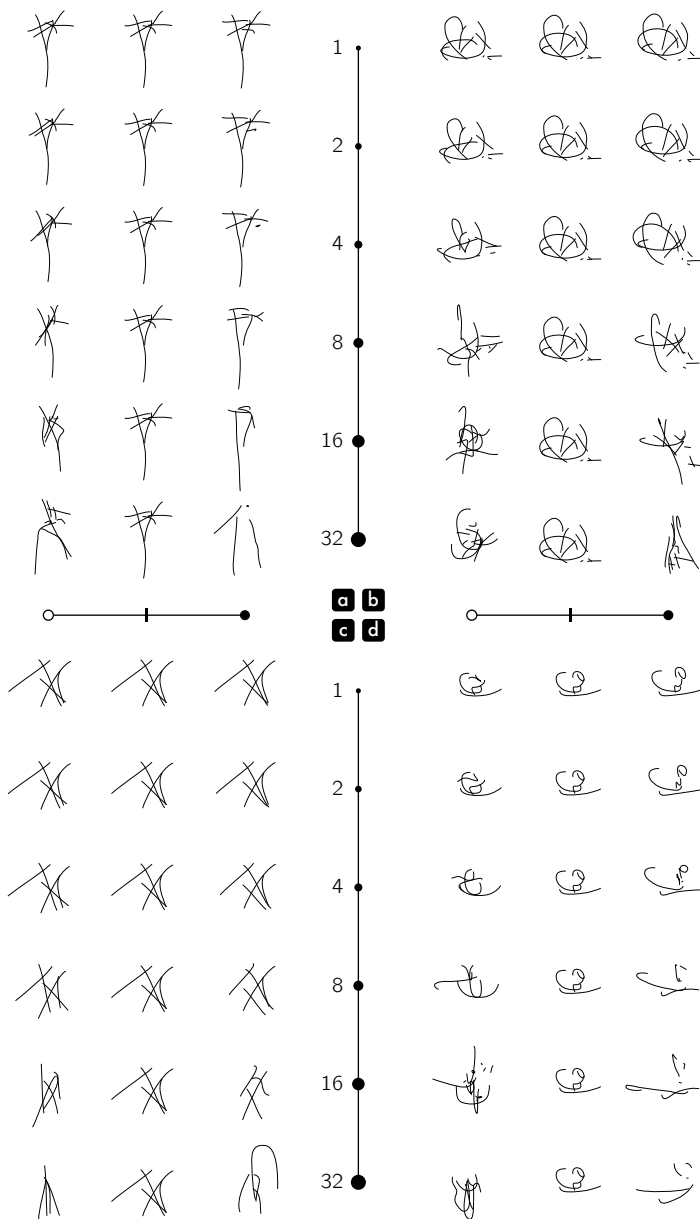


Figure 5.3: Similarity judgment threshold estimation with an angular metric. Each sample is separated from its neighbors by the angle indicated along the vertical line (in degree).

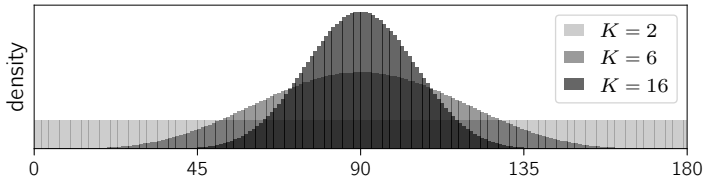


Figure 5.4: Unsigned angle distribution between random samples for different  $K$  dimensionalities (in degree).

### Multidimensional scaling

Multidimensional scaling (MDS) is originally a method to reduce the number of dimension of a dataset for visualization purposes. Compared to Principal Component Analysis<sup>6</sup>, MDS does not search for a linear transformation matrix into a lower dimensional subspace. MDS is designed to directly find new data-points coordinates in a possibly nonlinear manner. The idea is to construct a pairwise distance matrix between every data-points, and to find new coordinates optimally respecting these pairwise distances. In our case, it could be to find a 2-d cartography of our 16-d composition embeddings. However, our latent space is quite compact and homogeneous in all dimensions. It is supposed to be a 16-d standard normal distribution with independent components. So, there is no particular way to find a lower dimensional subspace.

The distance matrix is usually computed with Euclidean distances, but it can be constructed upon any type of metric. If distances are perceptually measured, MDS actually becomes a psychophysical tool. We can ask participants to directly rate all pairs of stimuli on an abstract psychological scale, e.g. categorical such as *very similar*, *similar*, *dissimilar*, *very dissimilar*. So, MDS is a very versatile tool dealing with metric and non-metric distance matrices, and accepting dataset of elements with known or unknown intrinsic dimensions. For instance, in the industry, MDS helps to interpret most significant aspects of objects that influence buying decisions, from ketchup to cars. Concerning our study on compositions, dimensional reduction is not an objective. We are looking for global perceptual distortions of the latent space, and we just want to reconstruct the best geometry that respect the perceptual pairwise distances collected with participants.

Original MDS relies on classical minimization algorithms, but a link has been highlighted with Kernel-PCA<sup>7</sup>. This method is much quicker and presents more robust resolutions. Already available in scikit-learn<sup>8</sup>, we have based our Kernel-

<sup>6</sup>PCA is a procedure searching for new dimensions with maximized variance, reordered by decreasing magnitude. We can then keep the few first dimensions only, which are supposed to be the most informative.

<sup>7</sup>C. K. Williams, 2002.

<sup>8</sup>Pedregosa et al., 2011. Find more information at: <https://scikit-learn.org>

## 5 Composition perception

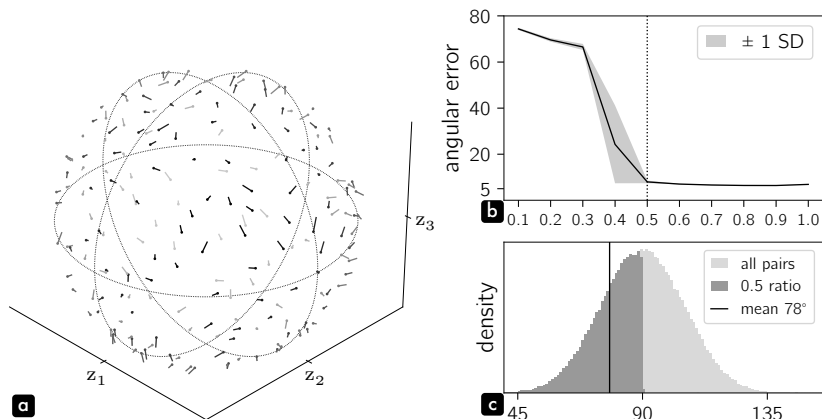


Figure 5.5: Multidimensional scaling. In panel **a**, we plot original points with small dots, and Kernel-MDS reconstruction error with short line segments. This simulation is based on 200 randomly sampled points on a unit sphere in 3-d. Their uniform spreading on the sphere is enforced by an over-sampling and a clustering by  $k$ -means. The distance matrix is clustered to 3 discrete levels. In panel **b**, we plot the reconstruction error of Kernel-MDS on 200 points (16-d, 7 discrete levels distance matrices) for different ratios of evaluated pairwise distances. Missing entries of the distance matrices are estimated by low-rank distance matrix completion. The error is expressed in angles ( $\pm 1$  SD in light gray). In panel **c**, we show the angular distance distribution between all pairs (light gray) and after a 0.5 reduction of further pairs (dark gray). The resulting mean angular distance is  $78^\circ$  (black line).

MDS solver on this algorithm. A simulated result in 3-d is displayed in Fig.5.5a with 200 points randomly sampled on a unit sphere. The distance matrix is computed using Euclidean distances, and clustered into 3 discrete levels. Original points are indicated by a small dot and the Kernel-MDS reconstruction error is materialized by short line segments. Resulting geometry is satisfactory, but even with a dataset of 200 elements only, the distance matrix required to complete the MDS is already of shape  $(200, 200)$ . Lower triangle of this matrix is of course a mirror of the upper triangle, and the diagonal is necessarily filled with zeros, but it still constitutes 19900 pairs. This procedure is therefore not feasible with real participants.

In addition, previous subsection highlighted the necessity to reduce the average angular distance between stimuli of evaluated pairs. As a result, we can imagine to present to participants the closest pairs only. Then, missing matrix entries can be estimated by low-rank distance matrix completion<sup>9</sup>. Fig.5.5b explores the feasibility of this scenario. We plot the reconstruction error of Kernel-MDS on 200 points (16-d, 7 discrete levels distance matrices) for different ratios of evaluated pairwise distances. Error is expressed in angle ( $\pm 1$  SD in light gray). We remark that a completion ratio of 0.5 is the lower limit for an acceptable reconstruction accuracy. In Fig.5.5c, we show the resulting angular distance distribution between

<sup>9</sup>We have ported from Matlab to Python the code associated with Mishra et al., 2011.

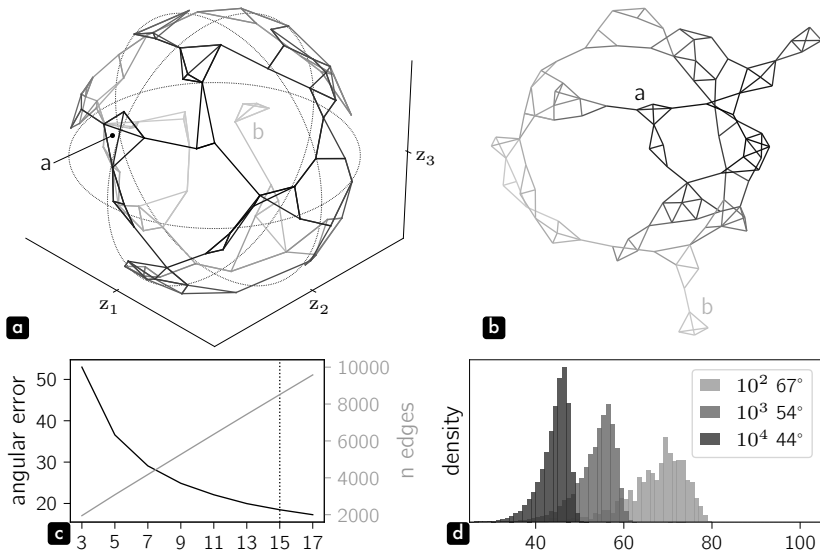


Figure 5.6: Isomap. Panels **a** and **b** presents the graph of 100 3-d points connected to 3 nearest neighbors, respectively in volume and unwrapped in 2-d (regions a and b correspond in both views). Panel **c** plots the reconstruction error, expressed in angle for different  $k$ -neighbors (black line). The simulation is based on 1000 16-d points, and evaluated distances are clustered to 7 discrete levels. We also display the implied number of evaluated edges per  $k$ -neighbors (gray line). Panel **d** presents the distribution of angular distances between stimuli for different number of points with 15 neighbors. The legend shows the mean angular distance for each condition.

all evaluated pairs in dark gray. Despite these efforts, the mean angular distance is still high, around  $78^\circ$ , and far above threshold.

### Isomap

The isomap algorithm is grounded on MDS, and proposes to take advantage of an existing representation when it is available. The main idea is to rely on the computation of the nearest neighbors to reduce the number of required evaluated pairs. These nearest neighbors can be represented as a graph, where each point is at least connected to  $k$ -neighbors. In Fig.5.6a,b, we show the graph of 100 3-d points connected to their 3 nearest neighbors, respectively in volume and unwrapped in 2-d. From this graph, being a subset of all possible pairwise distances, we can compute a complete distance matrix by propagating distances to neighbors, and take advantage of geodesics<sup>10</sup>. Popular algorithms to accomplish this task are Floyd-Warshall algorithm and Dijkstra algorithm. Both options are implemented

<sup>10</sup>A geodesic is the shortest path between two points on a surface, which is in our case on a hypersphere.

## 5 Composition perception

in Scipy, and the second is usually preferred when the number of data-points is large. Nonetheless, a requirement for these algorithms to work is that the nearest neighbor routine ends up in a unique graph. With a very small number of neighbors, multiple clusters are likely to happen, and to make filling algorithms fail. Once the full distance matrix is computed, Kernel-MDS can be used to reconstruct the geometry.

In Fig.5.6c, we plot the reconstruction error, expressed in angle, for different numbers of neighbors (black line). The simulation is based on 1000 16-d points and evaluated distances are clusterized to 7 discrete levels. We remark that even with 15 nearest neighbors, the angular error is still high. In gray, we show the corresponding number of edges to evaluate. With  $k = 15$ , the figure is around 8500 pairs of stimuli. It is already a lot, but much smaller than original 499500 possible pairs. In Fig.5.6d, we then look at the corresponding distribution of angular distance between stimuli. With 1000 points and  $k = 15$ , the mean angle is around  $54^\circ$ . Even if we could scale the number of evaluated stimuli to 10000 (around 85000 edges/pairs), it would only reduce the mean angle to  $44^\circ$ . As a reminder, a comparative judgment between compositions seems possible until  $8^\circ$  to  $16^\circ$  (see Fig.5.3). In conclusion, any attempt to completely characterize a latent space in 16-d is not feasible under the constraint of a realistic number of pairs to be evaluated with participants.

## 5.2 Perceptual scaling

Dimensionality involved in our compositional latent space forces us to modify our experimental strategy. Distortions need to be investigated at a more local scale. As a result, we renounce to capture a perceptual multidimensional geometry of the space, and concentrate on specific trajectories, i.e. univariate interpolations mapped in our high-dimensional space. This scheme actually gets back to a more customary psychophysical domain, known as *perceptual scaling*. In this section, we first go through the theoretical background of the method used in our experiments, i.e. MLDS, and then we detail our contributions to this method, such as the reduction of the required number of evaluated pairs, and the extension to periodic physical spaces.

### *Perceptual scaling*

Perceptual scaling is basically measuring the relationship between a physical metric and its induced psychological response. It is not only studying perceptual characteristics of a particular stimulus like in threshold estimation, but finding a transfer function covering a whole range of stimuli (see Fig.5.7).

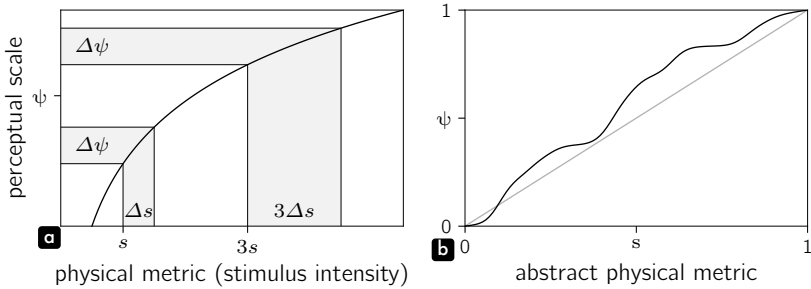


Figure 5.7: Perceptual Scaling. Panel **a** is an illustration of the Fechner’s law with stimuli varying in intensity of a physical medium. A stimuli  $3s$  requires a difference 3 times more important to produce the same perceptual change as a stimulus  $s$ . In panel **b**, we show that perceptual scales can be linearly transform in their physical and psychological domains without loss of significance. They can be bounded in the range  $[0, 1]$  for the ease of manipulation.

A naive question is then why sensory modalities do not simply react linearly to physical magnitudes of stimuli? First, our biological sensors have physical limitations in terms of spectrum and signal intensities. For instance, our vision is restricted to specific electromagnetic wavelengths, and direct sunlight is quickly harmful. Secondly, a linear scaling is not generally a very efficient strategy. We could imagine that our brain has limited computing abilities. In this case, it would be better to selectively allocate its power, e.g. to a wide range of stimuli with little attention, or to a restricted portion of possible stimuli with a higher accuracy. Finally, the environment does not provide an equiprobable physical space. Some colors and sounds are more likely than others, and may interact with our sensory system depending on our physical needs. The brain plasticity is precisely designed to adapt over time, life, and circumstances. As a result, it seems very likely that transfer functions between physical stimuli and our psychological representation are non-trivial. Perceptual scaling is therefore at the core of cognitive sciences and psychophysics in particular.

Let us now formalize our sensory model. First, let the random variable  $s$  be *physical* stimuli over a uni-dimensional space  $\mathcal{S}$ .  $s$  is a random variable because we assume that stimuli are distributed in the world according to some prior  $p(s)$ . Then, we assume that physical stimuli are mapped to a perceptual representation  $\psi$ , itself a random variable with the distribution  $p(\psi)$ . Perceptual scaling is then searching for the transfer function  $\Psi$ , so that:

$$\psi = \Psi(s) \quad (5.4)$$

We also constrained  $\Psi$  to be strictly monotonically increasing. A specific representation  $\psi$  should correspond to one stimulus  $s$  only. In addition, as the first derivative of  $\Psi$  is strictly positive, respective ordering of stimuli and representations is guaranteed.



## 5 Composition perception

In Fig.5.7a, we depict the most well known perceptual scale, as it relies on Fechner's law. This pioneer of psychophysics stated that the *intensity of our sensation* evolves logarithmically with the intensity of the stimuli, so that  $\psi = k \log(s)$ . It suggests that  $\Delta\psi = \frac{k}{s} \Delta s$ . Thus, to obtain an equivalent perceptual difference at  $3s$ , it requires a stimuli difference of  $3\Delta s$ . In short, for low intensities, when the slope is steep, we are more discriminative, and when the perceptual scale is flatter for high intensities, where are less accurate at discriminating stimuli variations. This relation has been experimentally demonstrated for some immediate metrics of sensory modalities, e.g. light intensity for vision, but Fechner's Law is far from universal. In addition, this explicit mapping of a perceptual representation from a stimulus expressed in physical units leads to interpretation confusion. For Thurstone, we have to "deny that it [the psychological continuum] measures sensation intensity or any other quantitative characteristic of sensation. [...] The sense quality is not itself an intensity or magnitude of any sort"<sup>11</sup>. The perceptual scale should rather be interpreted as purely abstract and arbitrary with its own *mental unit*. The perceptual scale can therefore be linearly transformed without any loss of information, since it will keep relative distances. The same apply on the stimuli dimension. We can think of  $s$  as an intermediary variable to a real physical metric. Then, without loss of generality, we can assume the space  $\mathcal{S}$  as bounded in the range  $[0, 1]$ . As a result, we can set  $\Psi(0) = 0$  and  $\Psi(1) = 1$  to constrain the perceptual scale to be also bounded in the range  $[0, 1]$ . This convention is plotted in Fig.5.7b and will greatly simplify further developments.

### Thurstonian scaling

We should now describe methods to construct perceptual scales. As evoked earlier, the initial procedure to build some sort of psychological representation was by integrating successive JND. But Thurstone remarked that:

The just noticeable difference is in every case a stimulus measurement. [...] Hence, it is in reality a physical unit which in some situations can serve indirectly the purposes of mental measurement.<sup>12</sup>

For him, judgment distribution along the psychological continuum is a more interesting unit, as it makes explicit stochastic relations attached to stimuli and our perception.

It is reasonable to assume that two perceptual sense qualities or processes which are close together on the psychological continuum are qualitatively similar and that therefore either one of them may more or less readily be perceived in the same stimulus.<sup>13</sup>

---

<sup>11</sup>Thurstone, 1927b.

<sup>12</sup>Thurstone, 1927a.

<sup>13</sup>Thurstone, 1927b.

Formally, it led Thurstone to introduce a third law of perception, the *law of comparative judgments*, expressed as:

$$\psi_1 - \psi_2 = x_{1,2} \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \quad \text{with} \quad x_{1,2} = \Phi^{-1}(p_{s_1 > s_2}) \quad (5.5)$$

However, this canonical formulation is not easy to handle. Let us show how to derive it from a more theoretical view. First, we assume that our sensory system can only rely on some noisy measurements  $m_a$  of world stimuli  $s_a$ . It can be written as:

$$m_a = \Psi(s_a) + n_{\psi_a} = \psi_a + n_{\psi_a} \quad \text{with} \quad n_{\psi_a} \sim \mathcal{N}(0, \sigma_{\psi_a}^2) \quad (5.6)$$

Again,  $n_{\psi_a}$  is assumed to be normal, as the result of many independent sources of noise. Then, in an experimental design, e.g. 2AFC, we will ask a participant to judge multiple times the difference between two stimuli  $s_i$  and  $s_j$ <sup>14</sup>. In practice, the question is *Which one is more?*, so that we actually measure  $p(m_j > m_i)$  (not  $p(s_j > s_i)$  as the participant only has access to measurements of given stimuli).

$$\begin{aligned} p(m_j > m_i) &= p(\psi_j + n_{\psi_j} > \psi_i + n_{\psi_i}) \\ &= p(n_{\psi_i} - n_{\psi_j} < \psi_j - \psi_i) \end{aligned} \quad (5.7)$$

where  $n_{\psi_i} - n_{\psi_j}$  is a difference of normally distributed variables. The result is a random variable  $n_{\psi_i, \psi_j} \sim \mathcal{N}(0, \sigma_{\psi_i}^2 + \sigma_{\psi_j}^2 - 2\rho_{i,j}\sigma_{\psi_i}\sigma_{\psi_j})$ , itself normally distributed, with  $\rho_{i,j}$  the correlation between  $n_{\psi_i}$  and  $n_{\psi_j}$ . Then:

$$\begin{aligned} p(m_j > m_i) &= p\left(\bar{n} \sqrt{\sigma_{\psi_i}^2 + \sigma_{\psi_j}^2 - 2\rho_{i,j}\sigma_{\psi_i}\sigma_{\psi_j}} < \psi_j - \psi_i\right) \quad \text{with} \quad \bar{n} \sim \mathcal{N}(0, 1) \\ &= p\left(\bar{n} < \frac{\psi_j - \psi_i}{\sqrt{\sigma_{\psi_i}^2 + \sigma_{\psi_j}^2 - 2\rho_{i,j}\sigma_{\psi_i}\sigma_{\psi_j}}}\right) \\ &= \Phi\left(\frac{\psi_j - \psi_i}{\sqrt{\sigma_{\psi_i}^2 + \sigma_{\psi_j}^2 - 2\rho_{i,j}\sigma_{\psi_i}\sigma_{\psi_j}}}\right) \\ \psi_j - \psi_i &= \Phi^{-1}(p(m_j > m_i)) \sqrt{\sigma_{\psi_i}^2 + \sigma_{\psi_j}^2 - 2\rho_{i,j}\sigma_{\psi_i}\sigma_{\psi_j}} \end{aligned} \quad (5.8)$$

where  $\Phi$  is the cumulative distribution function of a standard normal random variable, and  $\Phi^{-1}$  its inverse. Replacing  $j \rightarrow 1$  and  $i \rightarrow 2$  gives a result coherent with the canonical formulation (see Eq.5.5). We have just reversed subscripts order because intuitively we imagine  $s_j > s_i$ , and that it implies a positive distance between  $\psi_i$  and  $\psi_j$ , only if we consider  $\psi_j - \psi_i$ .

<sup>14</sup>It could also be conducted on multiple participants producing only one judgment, and pooled together. These two scenarios constitute cases I and II described in Thurstone, 1927a.

## 5 Composition perception

In practice, we will make further assumptions, as anticipated by Thurstone in his case V. First, we will assume there is no correlation between stochastic aspects of each measurement made by participants, i.e.  $\rho_{i,j} = 0$ . In addition, we will hypothesize that these internal noises occurring at the representation level are constant, no matter the triggering stimulus. As a result,  $\sigma = \sigma_{\psi_i} = \sigma_{\psi_j}$  and we can reformulate Eq.5.6 as:

$$m_a = \Psi(s_a) + n = \psi_a + n \quad \text{with} \quad n \sim \mathcal{N}(0, \sigma^2) \quad (5.9)$$

then Eq.5.8 becomes,

$$\psi_j - \psi_i = \Phi^{-1}(p(m_j > m_i))\sigma\sqrt{2} \quad (5.10)$$

Assuming a constant internal noise is actually an important choice concerning our sensory model because it provides a *goal* for the perceptual scaling operated in our brain, i.e. noise uniformization. In particular, we can think of our measurements  $m_a$  as the processing by  $\Psi$  of stimuli  $\hat{s}_a$ , of which stochasticity is considered in the physical domain (remember Eq.5.3). As  $n_{s_a}$  explores a locality around  $s_a$ , we can apply the Taylor series, so that:

$$m_a = \Psi(\hat{s}_a) = \Psi(s_a + n_{s_a}) = \Psi(s_a) + \Psi'(s_a) n_{s_a} + o(n_{s_a}) \quad (5.11)$$

comparing with Eq.5.9, we have:

$$n \simeq \Psi'(s_a) n_{s_a} \quad \text{and} \quad \sigma \simeq \Psi'(s_a) \sigma_{s_a} \quad (5.12)$$

In other words,  $\Psi$  derivative is tuning our sensitivity w.r.t physical stimuli, and uniformizes the uncertainty of our internal representation. We can actually see this process as an uniformization in the sensory domain of psychometric functions and thresholds (see Fig.5.8).

### Maximum Likelihood Difference Scaling

Before addressing practical computations of perceptual scales from data collected with participants, we introduce a second perceptual scaling framework, the Maximum Likelihood Difference Scaling (MLDS)<sup>15</sup>. Thurstonian scaling (TS) and MLDS are considered in the literature as more different from what they really are, and we will hypothesize some explanations later in this chapter. Nonetheless, the only fundamental difference between TS and MLDS is the type of judgment asked to participants. With TS, an absolute understanding of the *physical* scale is required. This is for instance applicable while studying light intensity, or image degradation due to compression<sup>16</sup>. Participants directly understand *What is more?* So, given a pair of stimuli  $s_i, s_j$  in a 2AFC task, they have an objective reference

<sup>15</sup>Knoblauch and Maloney, 2008; Maloney and Yang, 2003.

<sup>16</sup>Watson and Kreslake, 2001.

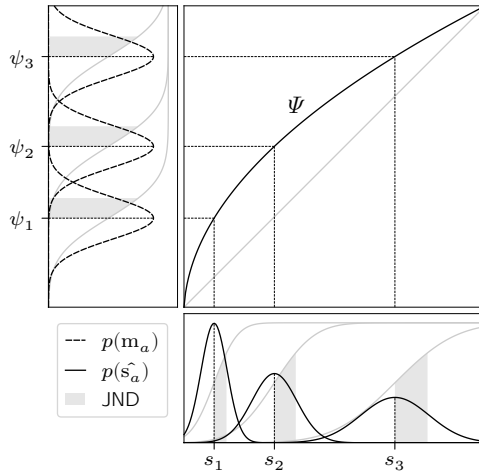


Figure 5.8: Perceptual scale tunes sensitivity associated with stimuli by uniformizing the uncertainty of our internal representation i.e.  $n \approx \Psi'(s_a) n_{s_a}$ . The same apply on psychometric functions and JND, becoming equivalent in the psychological domain.

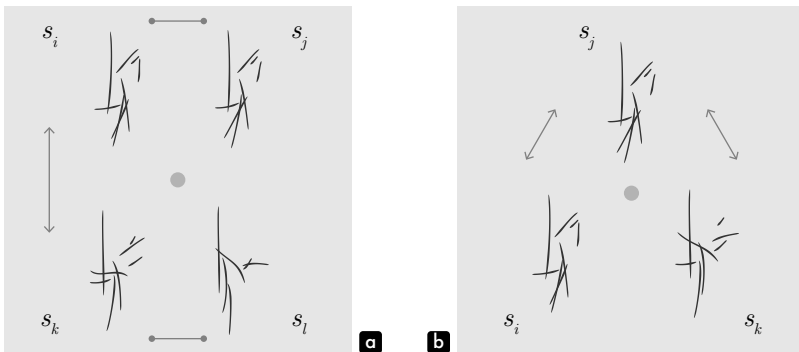


Figure 5.9: MLDS task setups for the quadruplet case in panel **a** and the triplet case in panel **b**.

to base their judgment on. On the contrary, with compositions, there is no such intuitive directional metric. MLDS basically addresses this issue by proposing to judge *differences of differences* of stimuli. Given two pairs of stimuli, the task can be rephrased as: *Which pair shows the largest interval?* (Fig.5.9a). Alternatively, if one stimulus of each pair is shared and taken for reference, we obtain a triplet for which the task is: *Which stimulus is more different to the reference?* (Fig.5.9b).

Let us concentrate on the quadruplet case. Participants are presented with four stimuli  $s_i, s_j, s_k, s_l$ , presented as pair  $(s_i, s_j)$  and pair  $(s_k, s_l)$ . In the process,

## 5 Composition perception

they have to measure distances  $m_{i,j}$  and  $m_{k,l}$ , before comparing the two. The particularity of estimating  $m_{i,j}$  and  $m_{k,l}$  is that the orientation/sign of these distances is ignored (because usually unknown or irrelevant). As a result, we must write:

$$\begin{aligned} m_{a,b} &= |m_b - m_a| \\ &= |\psi_b + n_{\psi_b} - \psi_a - n_{\psi_a}| \end{aligned} \quad (5.13)$$

But this formulation does not guarantee the normality of the distribution of  $m_{a,b}$ . Consequently, the comparison between  $m_{i,j}$  and  $m_{k,l}$  would not be normally distributed either, and the MLDS framework would be impracticable to measure perceptual scales. We illustrate this observation for two neighboring stimuli in Fig.5.10a. We plot the distribution of  $m_{a,a+1}$  for various  $\tau$  (see Definition.1). The distribution of  $m_{a,a+1}$  is compared with a classical comparative judgment as defined in Eq.5.10 (TS case V). We remark that the issue is negligible for  $\tau = 0.5$  and severe with  $\tau = 2$ . In Fig.5.10b, we provide a quantitative measure of this discrepancy by computing the Kullback-Leibler divergence between the two distributions. In short, if the noise ( $\sim \sigma$ ) associated with the perception of individual stimuli  $\psi_a$  and  $\psi_b$  is small compared to the distance between them, the normality assumption can hold. We will therefore assume  $p(n_{\psi_a} - n_{\psi_b} < |\psi_b - \psi_a|)$  to be high, so that  $p(m_{a,b} > 0)$  is also high with:

$$m_{a,b} = |\psi_b - \psi_a| + n_{\psi_b} - n_{\psi_a} \quad (5.14)$$

Then, making similar assumptions as for the TS case V, we have:

$$\begin{aligned} p(m_{k,l} > m_{i,j}) &= p(|\psi_l - \psi_k| + n_{\psi_l} - n_{\psi_k} > |\psi_j - \psi_i| + n_{\psi_j} - n_{\psi_i}) \\ &= p(-n_{\psi_i} + n_{\psi_j} + n_{\psi_k} - n_{\psi_l} < |\psi_l - \psi_k| - |\psi_j - \psi_i|) \\ &= p(2\sigma \bar{n} < |\psi_l - \psi_k| - |\psi_j - \psi_i|) \\ &= \Phi \left( \frac{|\psi_l - \psi_k| - |\psi_j - \psi_i|}{2\sigma} \right) \end{aligned}$$

$$|\psi_l - \psi_k| - |\psi_j - \psi_i| = \Phi^{-1}(p(m_{k,l} > m_{i,j}))2\sigma \quad (5.15)$$

As a reminder, function  $\Psi$  is strictly monotonically increasing, so if  $s_i < s_j$  and  $s_k < s_l$ , then  $\psi_i < \psi_j$  and  $\psi_k < \psi_l$ . To simplify the notation, we can finally reorder stimuli and remove absolute values.

$$\psi_l - \psi_k - \psi_j + \psi_i = \Phi^{-1}(p(m_{k,l} > m_{i,j}))2\sigma \quad (5.16)$$

In the triplet case, we set  $s_k \rightarrow s_j$  and  $s_l \rightarrow s_k$ , leading to:

$$\psi_k - 2\psi_j + \psi_i = \Phi^{-1}(p(m_{j,k} > m_{i,j}))2\sigma \quad (5.17)$$

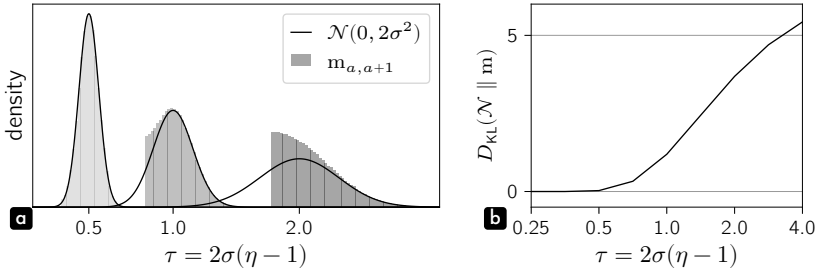


Figure 5.10: MLDS pair measurement normality verification. Gray areas in panel **a** shows  $m_{a, a+1}$  distributions for various  $\tau$  (see Definition.1 for details about  $\tau$ ). Pair measurements are compared with the distribution of a classical comparative judgment (TS case V, black line). Panel **b** plots the  $D_{KL}$  divergence between the two distributions as a function of  $\tau$ .

**Definition 1** We define  $\tau$  as the standard deviation of an MLDS comparative judgment, expressed in step size units between  $\eta$  equally spaced stimuli. In the specific context of a physical range of stimuli rescaled in the interval  $[0, 1]$ , we have  $\tau = \frac{2\sigma}{1/(\eta-1)} = 2\sigma(\eta - 1)$ , where  $2\sigma$  corresponds to the expected spread of an MLDS comparative judgment (see the demonstration provided in this subsection). Nonetheless, we would like to repeat that  $\tau$  and  $\sigma$  are by definition related to the noise associated with our psychological representation only. It does not have any direct meaning in the physical space.  $\sigma$  is related to the noise associated with each stimulus  $\sigma_s$ , and has implications in terms of thresholds or JND, only if  $\Psi$  is fully characterized or assumed linear (see Eq.5.12). The primary idea behind  $\tau$  is to study the behavior of comparative judgments, independently of the chosen  $\eta$  number of stimuli and  $\sigma$  imposed by the task.

### Fitting methods

Let us first rephrase the problem. We have  $\eta$  stimuli, and we want to know their  $\eta$  scaled values  $\psi_a$ , as well as  $\sigma$  of the associated perceptual noise. Although, as proposed earlier, the perceptual scale will be rescaled in the range  $[0, 1]$  in both physical and perceptual domains, leading to  $\psi_0 = 0$  and  $\psi_{\eta-1} = 1$ . As a result, we have  $\eta - 1$  parameters to estimate.

In a second time, we need to determine the number of experimental sets (pairs, triplets, or quadruplets) we can produce with  $\eta$  stimuli. With TS, we just have to find pairs of different stimuli, which gives  $\eta(\eta - 1)$  pairs. However, if we experimentally measure  $p(m_j > m_i)$ , we can compute  $p(m_i > m_j) = 1 - p(m_j > m_i)$ . So, measuring both configurations is not necessary. Then, expected pairs correspond to the well-known combinations without replacement, giving a number of combinations  $T = \binom{\eta}{2}$ . So, if  $\eta \geq 2$ , the problem is solvable.

## 5 Composition perception

Concerning the quadruplet version of MLDS, we have two stimuli ordering constraints,  $s_i < s_j$  and  $s_k < s_l$ . So, overlapping pairs  $s_i < s_k < s_j < s_l$  is theoretically possible. One stimulus of each pair could also be identical, as in the triplet version. In practice, these two conditions may be confusing for participants and introduce uncontrolled bias. Thus, we impose  $s_i < s_j < s_k < s_l$  as in the original paper<sup>17</sup>. It limits the number of possible sets of stimuli, but this is not problematic to make the problem solvable. The second positive aspect is that quadruplets can again be computed as combinations without replacement, giving  $T = \binom{\eta}{4}$  quadruplets. The problem is now solvable if  $\eta \geq 5$ . For the triplet case, we impose  $s_i < s_j < s_k$ , and the number combinations is  $T = \binom{\eta}{3}$ . If  $\eta \geq 4$ , the problem is solvable.

We now focus on the MLDS quadruplet version to detail the two implemented fitting procedures<sup>18</sup>, i.e. Generalized Linear Model (GLM) and Maximum Likelihood Estimation (MLE). Adaptation to TS and the triplet MLDS is then straightforward. We have  $T$  combinations  $\mathbf{x}_t = [i_t, j_t, k_t, l_t]$ , so that we can build a system of  $T$  equations like Eq.5.16. For instance, with  $\eta = 5$ ,  $T = 5$ , we have:

$$\begin{cases} \psi_0 - \psi_1 - \psi_2 + \psi_3 = \Phi^{-1}(p(m_{2,3} > m_{0,1}))2\sigma \\ \psi_0 - \psi_1 - \psi_2 + \psi_4 = \Phi^{-1}(p(m_{2,4} > m_{0,1}))2\sigma \\ \psi_0 - \psi_1 - \psi_3 + \psi_4 = \Phi^{-1}(p(m_{3,4} > m_{0,1}))2\sigma \\ \psi_0 - \psi_2 - \psi_3 + \psi_4 = \Phi^{-1}(p(m_{3,4} > m_{0,2}))2\sigma \\ \psi_1 - \psi_2 - \psi_3 + \psi_4 = \Phi^{-1}(p(m_{3,4} > m_{1,2}))2\sigma \end{cases} \quad (5.18)$$

To simplify the representation, we can rewrite this system with vectors. First, we define the perceptual representation of our selection of stimuli as  $\boldsymbol{\psi} = [\psi_0, \dots, \psi_a, \dots, \psi_{\eta-1}]$ . Secondly, we can build a matrix  $\mathbf{X}$  of size  $(T \times \eta)$ , where each value  $X_{t,a}$  corresponds to the weight of  $\psi_a$  for this particular  $\mathbf{x}_t$  combination. Applied to our example:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 & 1 \\ 1 & 0 & -1 & -1 & 1 \\ 0 & 1 & -1 & -1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{X}\boldsymbol{\psi} = \begin{bmatrix} \psi_0 - \psi_1 - \psi_2 + \psi_3 \\ \psi_0 - \psi_1 - \psi_2 + \psi_4 \\ \psi_0 - \psi_1 - \psi_3 + \psi_4 \\ \psi_0 - \psi_2 - \psi_3 + \psi_4 \\ \psi_1 - \psi_2 - \psi_3 + \psi_4 \end{bmatrix} \quad (5.19)$$

Finally, we combine all participant responses as a vector  $\mathbf{r} = [p(m_{k_0, l_0} > m_{i_0, j_0}), \dots, p(m_{k_{\eta-1}, l_{\eta-1}} > m_{i_{\eta-1}, j_{\eta-1}})]$ , and rearrange so that:

$$\mathbf{X} \frac{\boldsymbol{\psi}}{2\sigma} = \Phi^{-1}(\mathbf{r}) \quad (5.20)$$

<sup>17</sup>Maloney and Yang, 2003.

<sup>18</sup>With Jonathan Vacher, we have developed a Python package dedicated to TS and MLDS. It includes improvements described in this section, and we hope to make it open source as soon as possible.

This system has now the form of a Generalized Linear Model, with  $\Phi^{-1}$  as a link function. Defining unknowns as  $\beta = \frac{\psi}{2\sigma}$ , they can be solved by *least squares*. It consists of iteratively minimizing the following loss function:

$$\mathcal{L}_{GLM} = \sum^T (\mathbf{X}\beta - \Phi^{-1}(\mathbf{r}))^2 \quad (5.21)$$

In practice, we drop the first column of  $\mathbf{X}$  and the first value of  $\beta$ , because  $\beta_0 = \frac{\psi_0}{2\sigma} = 0$  does not have to be solved. Then, once  $\beta$  is known:

$$\sigma = \frac{\psi_{\eta-1}}{2\beta_{\eta-1}} = \frac{1}{2\beta_{\eta-1}} \quad \text{and} \quad \psi = 2\sigma\beta \quad (5.22)$$

Despite its simplicity, regular *least squares* algorithm does not accept additional fitting constraints. Indeed, we know that  $\beta > 0$ , and that  $\Psi$  is strictly monotonically increasing. Therefore, if stimuli are ordered, we should have  $\psi_a < \psi_{a+1}$ , i.e.  $\beta_{a+1} - \beta_a > 0$ . To take advantage of these constraints, we use the SLSQP algorithm<sup>19</sup>, which accepts inequalities.

Another important implementation detail is about  $\Phi^{-1}$ . This function, also called *probit* is going to infinity in 0 and 1 ( $\Phi^{-1}(0) \rightarrow -\infty$  and  $\Phi^{-1}(1) \rightarrow \infty$ ). It causes numerical problem, so it is customary to clip participant responses by a small amount  $\alpha$ . By default, we set  $\alpha = 10^{-16}$ , but we found that this value has a huge impact on the fitting accuracy, especially for  $\sigma$ . We do not have good explanations for this phenomenon, and it should be addressed in future works. Nonetheless, we provide a simulation in Fig.5.11. Simulation details are chosen to be as close as possible from our experimental setup (see Footnote.27). It appears that  $\alpha = 10^{-4}$  gives the lowest error on  $\sigma$  and a good accuracy for  $\psi$ . This  $\alpha$  value is therefore far from a machine precision limitation value. It is closer to a lapse rate, referring in psychophysics to the probability of unintentional responses due to some inattention of the participants, or some noise in their motor actions when pressing response buttons. Then, this lapse rate acts as a ceiling performance for supra-threshold choices<sup>20</sup>.

In Fig.5.11, we also remark that MLE (dotted lines) is better GLM for both optimizations of  $\psi$  and  $\sigma$ . To describe the MLE algorithm, let us recall that to estimate  $p(m_{k_t, l_t} > m_{i_t, j_t})$ , we average multiple binary responses  $b_t$  from participants, so that  $p(m_{k_t, l_t} > m_{i_t, j_t}) = \frac{1}{U} \sum_{u=1}^U b_{t,u}$ , with  $U$  the number of repetitions. Then, we can build a matrix  $\mathbf{B}$  of size  $(T \times U)$  as  $\mathbf{r} = \frac{1}{U} \sum^U \mathbf{B}$ . As a result, the MLE minimization function is basically the negative log-likelihood of a Bernoulli random variable:

<sup>19</sup>Sequential Least Squares Programming (SLSQP), developed by Kraft, 1988 is conveniently available in Scipy.

<sup>20</sup>Even if not elaborated to get rid of the  $\alpha$  issue, an alternative experimental design proposed by Boschman, 2001 asks participants to directly operate a categorical rating, corresponding to  $\Phi^{-1}(p(m_j > m_i))$ . It may be suitable with very specific stimuli, but repeated binary selections is usually easier for participants than having an abstract categorical response scale such as: *much less, less, more, much more*



## 5 Composition perception

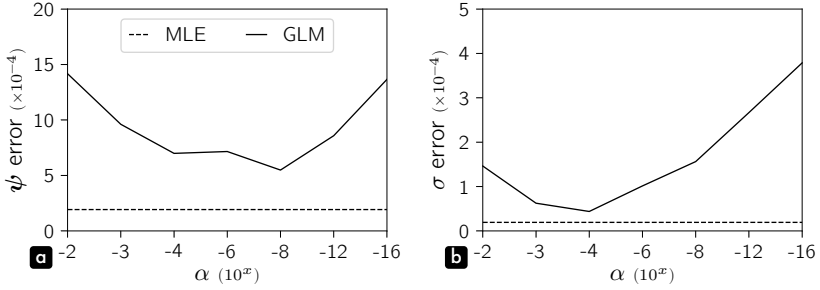


Figure 5.11: Fitting methods comparison for  $\psi$  (panel **a**) and  $\sigma$  (panel **b**). It is based on simulations with  $\tau = 1$  (see Footnote.27 for other simulation details). MLE errors are shown with dotted lines. GLM errors are plotted for different  $\alpha$  values, that clip participant responses before the application of  $\Phi^{-1}$  (solid lines).

$$\begin{aligned}
 \mathcal{L}_{MLE} &= -\frac{1}{U} \sum^T \sum^U \log \Phi(\mathbf{X}\beta)^{\mathbf{B}} (1 - \Phi(\mathbf{X}\beta))^{1-\mathbf{B}} \\
 &= -\frac{1}{U} \sum^T \sum^U \left( \mathbf{B} \log(\Phi(\mathbf{X}\beta)) + (1 - \mathbf{B}) \log(1 - \Phi(\mathbf{X}\beta)) \right) \\
 &= -\sum^T \left( \log(\Phi(\mathbf{X}\beta)) \frac{1}{U} \sum^U \mathbf{B} + \log(1 - \Phi(\mathbf{X}\beta)) \frac{1}{U} \sum^U (1 - \mathbf{B}) \right) \\
 &= -\sum^T (r \log(\Phi(\mathbf{X}\beta)) + (1 - r) \log(1 - \Phi(\mathbf{X}\beta)))
 \end{aligned} \tag{5.23}$$

with  $U$  introduced as a dividing constant to simplify notations. In practice, we also drop the first column of  $\mathbf{X}$  and the first value of  $\beta$ , because  $\beta_0 = 0$ . SLSQP is used as the minimizing algorithm with similar constraints as for GLM. Concerning outputs  $\psi$  and  $\sigma$ , Eq.5.22 is still valid to extract them from  $\beta$ . Finally, MLE requires  $\beta$  to be initialized. So,  $\beta_{init}$  is chosen to be a linear progression in the range  $[0, \eta - 1]$ .<sup>21</sup>

GLM should be theoretically preferred as it provides an easier problem to solve (linear). It is also guaranteed to converge on a single solution, compared to MLE that can fail during minimization. However, failures did not happen on our data, and rarely occurred on simulations. In addition, the  $\alpha$  issue seems to make GLM less accurate. We have therefore chosen MLE in all the following results.

### TF vs MLDS

We have seen that the sensory model and fitting procedures behind TS and MLDS are similar. In the specific case of an objective perceptual ordering of stimuli, both

<sup>21</sup>See Subsection 5.2. Combinations optimization for more details on this choice.

methods could even be used indifferently. However, TS and MLDS have usually been employed in different experimental setups, letting their claimed advantages and limitations wiping out common features with a certain confusion.

MLDS was primarily developed to enable experiments for which an objective perceptual ordering of stimuli is inaccessible to participants. The counterpart is the requirement of a physical metric to pre-order the stimuli, even if this metric is meaningless to the participants. So, MLDS has been mostly employed in experiments conducted on synthetic stimuli, where appropriate sets can be freely generated. They are generally constant, but between sessions, it would be possible to adapt the number of stimuli for a given physical range, and optimally fit observers' performances and thresholds. By opposition, TS has been associated with uncontrolled stimuli. For instance, with pre-existing, natural or *ready-made* stimuli, it may be difficult to measure their corresponding value along the chosen physical metric. For this reason, a study scaling the translucency of ocean waves in paintings used TS instead of MLDS<sup>22</sup>. In this situation, there is no physical metric to order stimuli, but participants do have an idea of *what is more translucent*<sup>23</sup>. With this picture in mind, let us now review MLDS claims against TS, as formulated by its authors.

TS have three major drawbacks. First, the scale depends crucially on the assumed distribution of judgment errors (it is not robust) while MLDS is remarkably robust [...]. Second, stimuli must be spaced closely enough so that the observer's judgments will frequently be inconsistent. This typically means that many closely-spaced stimuli must be scaled, and the number of trials needed to estimate the scale is much greater than in MLDS. The third drawback is the most serious. It is not obvious what the TS measures, at least not without further assumptions about how JND add up to produce perceptual differences. The MLDS scale based on quadruples is immediately interpretable in terms of perceived differences in interval lengths since that is exactly what the observer is judging.<sup>24</sup>

First, MLDS is supposed to be more reliable than TS, because it would be less dependent on the real distribution of the perceptual noise, if deviating from normality. However, both methods are grounded on the same assumptions concerning the sensory model. Ironically, we have shown that MLDS presents a theoretical issue of non-normality (see Fig.5.10), from which TS does not suffer. On the other hand, MLDS is a summation of 4 different sources of noise while TS only 2. By the central limit theorem, it is indicating a higher likelihood of normality in favor of MLDS, but is this significant? We believe, that except the higher number

<sup>22</sup>Wijntjes et al., 2020.

<sup>23</sup>This scenario is addressed in our Python toolbox. We use the fitting procedures described in Subsection.5.2.Fitting methods, and just remove the constraints on  $\beta$  positivity and monotonicity. For MLE,  $\beta_0$  is initialized to  $\mathbf{0}$ , and for GLM, SLSQP is replaced by a simple *least squares*. Finally, despite the unknown ordering of  $\beta$  values, we still consider  $\beta_0 = 0$ , that it is not optimized. We finally just add an extra step before the computation of  $\psi$  and  $\sigma$  from  $\beta$ , i.e.  $\beta$  elements are reordered, and we subtract the minimal first value (which may be negative).

<sup>24</sup>Knoblauch and Maloney, 2008.

## 5 Composition perception

of possible combinations allowed by MLDS, there is no strong justification making it more robust than TS.

The second comment is targeted toward the supposed limitation of TS for around-threshold pairs of stimuli. This observation comes from the fact that in the unordered stimuli scenario (stimuli picked in the *wild*), all possible pairs must be evaluated to obtain a reliable scaling. Then, even most distant stimuli should be in the participant confusion range, but this scenario is really specific. In a more general case of ordered stimuli, only neighboring pairs may be evaluated. Being only locally around-threshold, the complete physical range would then be much larger. So, the MLDS claim to be the only alternative to explore wider physical ranges is unfair. A related interrogation is about the supposed supra-threshold nature of MLDS. It is true that for participants looking at pairs such as  $(s_0, s_5)$  and  $(s_{10}, s_{16})$ ,  $s_0$  may not be confused with  $s_5$ , and  $(s_{10}, s_{16})$  similarly. Within pairs, we do have supra-threshold stimuli differences, but participants are judging difference of difference. In our example this value is still of one step only. On the contrary, comparing pairs  $(s_0, s_5)$  and  $(s_6, s_7)$  would be very easy and therefore useless for reconstructing the perceptual scale<sup>25</sup>. Longer distances require a lot of repetition to be accurate due to the shape of  $\Phi^{-1}$ . Both techniques are therefore around-threshold because perceptual confusion is driving perceptual scaling in both cases. Finally, does MLDS really require fewer trials than TS? Without specific constraints, MLDS combinatorial  $\binom{\eta}{4}$ , is actually mechanically higher than TS with  $\binom{\eta}{2}$ .

The last supposed issue concerning TS is about the interpretation of measured  $\psi$ . This is true, but again for the unordered stimuli scenario only. Not knowing physical positions of stimuli  $s$  prevents from plotting any  $\Psi$  function. For instance, assuming a lower density of  $\psi$  values around  $\psi_0$  than around  $\psi_{\eta-1}$ , may be explained by equally spaced stimuli and a compressive function  $\Psi$ , e.g.  $\Psi = \sqrt{s}$ , or a linear function  $\Psi = s$  with an uneven spread of stimuli, denser around  $s_{\eta-1}$ . Another explanation for this remark may be about TS conducted without *case V* assumptions. But MLDS is only *case V* in practice, so it would be like comparing different sensory models.

### Combinations optimization

The number of experimental combinations is directly influencing the task duration for participants. Randomly subsampling combinations has been proposed<sup>26</sup>, but as we consider MLDS as an around-threshold method, localized discrimination may certainly be a more efficient selection strategy.

---

<sup>25</sup>In the next subsection, we will further discuss this aspect. We believe that pairs intra-distance must be also limited to around-threshold differences.

<sup>26</sup>Maloney and Yang, 2003.

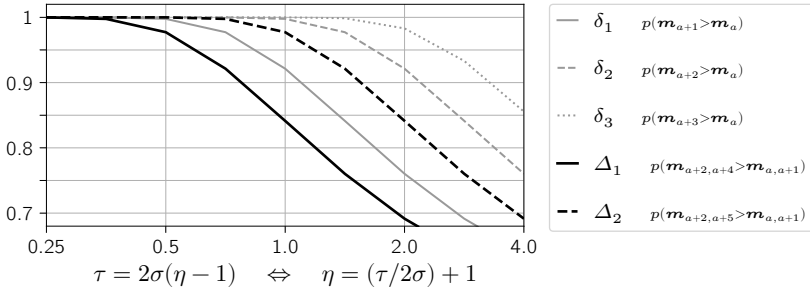


Figure 5.12: Theoretical local discriminative probabilities within pairs (gray lines) and between pairs (black lines) for various  $\tau$  (see Definition.1 for details about  $\tau$ ). Probabilities are shown for stimuli distances, or difference of stimuli distances, respectively expressed in steps by  $\delta_a$  and  $\Delta_a$ . A linear perceptual scale is assumed.

Supra-threshold situations are generally less demanding for participants as we are not probing the limit of their perceptual abilities. On the other hand, supra-threshold tasks are barely informative, and erroneous to some extent. The reason is practical: they cannot be measured with enough accuracy. Fig.5.12 illustrates this idea. We assume  $\eta$  equally spaced stimuli, a linear perceptual scale, and no non-normality issue. Then, we can plot discriminative probabilities within pairs (gray lines) and between pairs (black lines) for various  $\tau$  (see Definition.1 for details about  $\tau$ ). Secondly, these discriminative probabilities are shown for stimuli distances, or difference of stimuli distances, respectively expressed in step numbers by  $\delta_a$  and  $\Delta_a$ . For instance, for an MLDS judgment separated by two steps (black dotted line) with  $\tau = 1$ , the expected probability  $p(m_{a+2,a+5} > m_{a,a+1}) \approx 0.977$ . This would require almost 44 repetitions of the same judgment to be accurate, which is not plausible unless we pool participants.

A solution seems to increase  $\tau$ , but  $\sigma$  is a given variable, so a higher  $\tau$  implies a higher number of stimuli, corresponding to more combinations around-threshold, and a longer task. In addition, we have seen that a higher  $\tau$  also increases the non-normality of the perceptual noise (see Fig.5.10b). A higher  $\tau$  is therefore not the best option. Defining  $\delta$  and  $\Delta$  as the maximum *difference* and *difference of difference* values between combination indices, i.e.  $\delta_a \leq \delta$  and  $\Delta_a \leq \Delta$ , the question is then to find the right  $\tau$  for a given set of combination constraints. In Fig.5.13, we plot the simulated squared errors of  $\psi$  and  $\sigma$  fits for two conditions: ( $\Delta = 1, \delta = 2$ , solid line) and ( $\Delta = 2, \delta = 3$ , dotted line)<sup>27</sup>.

<sup>27</sup> Simulation details: we use the triplet MLDS with the MLE fitting procedure. Squared errors are averaged across 10 fits of sub-conditions parameterized by  $\eta \in [6, 11, 16, 21, 26]$  and  $\Psi(s) = s^\gamma$ , with  $\gamma \in [\frac{1}{4}, \frac{1}{2\sqrt{2}}, \frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2, 2\sqrt{2}, 4]$ . This way, we cover the usual range of stimuli number and several  $\Psi$  shapes. We use a high number of repetitions per combination, i.e. 100, giving a good precision of simulated response. In real experiments, repetitions per participant is at least 10 folds smaller, but we anticipate the pooling of all participants.

## 5 Composition perception

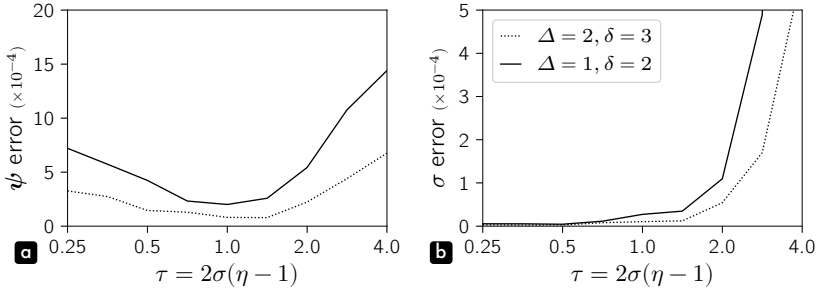


Figure 5.13: Simulation of optimal  $\tau$  (see Definition.1 for details about  $\tau$ ) for two conditions ( $\Delta = 1, \delta = 2$ , solid line) and ( $\Delta = 2, \delta = 3$ , dotted line) (see Footnote.27 for simulation details). We plot squared fitting errors of  $\psi$  in panel a and  $\sigma$  in panel b.

For the first condition, we remark that an acceptable range for  $\tau$  is  $[\frac{1}{\sqrt{2}}, \sqrt{2}]$ , leading to an optimal  $\sigma^* = \frac{1}{2(\eta-1)}$ . Consequently, from Fig.5.12 we notice that every pair will be supra-threshold (for  $\tau = 1$ ,  $p(\delta_1) \approx 0.921$ ,  $p(\delta_2) \approx 0.998$ ). In the precedent section, we arbitrarily defined a *possible comparative judgment range*, from the JND to an upper limit set to 99.9% of the psychometric function. So, pairs with a 2 steps difference are still inside the *confusion* range (with the assumption of a linear perceptual scale). We repeat that when participants evaluate the distance between stimuli, or their degree of similarity, stimuli still have to present common features, a tiny possibility for confusion. For instance, in the second scenario, with  $\tau = 1$ , combinations with pairs separated by 3 steps may be useless, if not harmful. Indeed, the optimal  $\tau$  value is shifted toward the right, around  $\sqrt{2}$  (see Fig.5.13a). From Fig.5.12, it seems to correspond to  $p(\Delta_1)$  at threshold and  $p(\delta_3) < 0.999$ , which is in line with our heuristic of *possible comparative judgment range*.

With  $\eta = 16$ , the number of evaluated combinations without constraint is 560. Thanks to our optimization scheme, we dramatically reduce this figure to 52 and 108 combinations for scenarios 1 and 2 respectively. This is about ten folds lower, while keeping reasonable fitting errors in both cases. We have therefore selected the first alternative ( $\Delta = 1, \delta = 2$ ) to conduct our experiments.

Then, the optimal  $\sigma^* = \frac{1}{2(\eta-1)}$  justifies the initialization of  $\beta$  in MLE, as a linear progression in the range  $[0, \eta - 1]$ . It is basically assuming equally spaced stimuli perceived through a linear perceptual scale, giving  $\psi_{init}$  to be a linear progression in the range  $[0, 1]$ , and  $\beta_{init} = \frac{\psi_{init}}{2\sigma^*} = \psi_{init}(\eta - 1)$ .

Finally, experimenters may decide very different combination constraints depending on the tasks and their needs. So, to be as generic as possible, we should determine acceptable ranges for  $\Delta$  and  $\delta$ , that ensure solvable perceptual scales. The TS case is simple:  $\delta$  is logically bounded in the range  $(1, \eta - 1)$ . For MLDS, we first want to select a  $\Delta$ , and then limit  $\delta$  accordingly. Indeed,  $\delta$  boundaries are just

	$\eta$	$\Delta$	$\delta$
TS	$\geq 2$	—	$(1, \eta - 1)$
MLDS3	$\geq 4$	$(1, \eta - 3)$	$(2, \Delta + 1)$
MLDS4	$\geq 5$	$(1, \eta - 4)$	$(2, \Delta + 1)$

Figure 5.14: TS and MLDS combination constraints to guarantee solvable perceptual scales.

one over  $\Delta$  boundaries<sup>28</sup>. Concerning  $\Delta$ , setting a lower bound to zero would imply too few combinations in the triplet case (MLDS3), and for small  $\eta$  in the quadruplet case (MLDS4). So, we have set  $\Delta$  lower bound to 1. The upper bound is then logically defined. Fig.5.14 summarizes all combination constraints.

### Difference scaling with periodic physical spaces

We have shown that measuring all combinations is not necessary, and that it could even introduce scaling inaccuracy for large supra-threshold judgments, especially with GLM fits. Then, by reducing measurements to local pairs/combinations, we can extend MLDS to periodic physical spaces, such as the circular slices of hypersphere of our compositional latent space. Classical MLDS can already partially address such periodic variables. For instance, with  $\phi$  an angle in the range  $[0, 2\pi[$ , we would be able to construct the perceptual scale, but there would not be any combinations across  $0 \equiv 2\pi$ , e.g.  $[s_i, s_j, s_k, s_l] = [1.8\pi, 1.9\pi, 0, 0.1\pi]$ . The lack of these combinations could then possibly bias the perceptual scale around  $0 \equiv 2\pi$ . As a result, we introduce a new MLDS variant, the Periodic MLDS (PMLDS).

In MLDS,  $\eta$  characterizes a set of stimuli like  $[s_0, \dots, s_{\eta-1}]$ , but in PMLDS, we have  $s_{\eta-1} \equiv s_0$ . Therefore,  $\eta$  is not the number of unique stimuli anymore, being  $\eta - 1$ . However, we will keep the original definition of  $\eta$  for simplicity, as  $s_{\eta-1}$  and  $\psi_{\eta-1}$  are useful to plot the perceptual scale physical and psychological bounds. In addition, even if  $(\psi_{\eta-1} \equiv \psi_0 \pmod{1})$ , implying  $(\beta_{\eta-1} \equiv \beta_0 \pmod{2\sigma})$ ,  $\beta_{\eta-1}$  is still required to estimate  $\sigma = \frac{1}{2\beta_{\eta-1}}$  (see Eq.5.22). This way, the number of unknowns also remains  $\eta - 1$ .

The intended periodicity implies tighter constraints on  $\delta$  and  $\Delta$  to limit combinations to more local pairs. With  $\lambda$  the maximum distance between indices  $j$  and  $k$  for PMLDS4, and  $\lambda = 0$  for PMLDS3, we can set up the following rule:

$$\begin{aligned}
 (\eta - 1) - (2\delta + \lambda) &\geq \delta + 1 \\
 3\delta &\leq \eta - 2 - \lambda \\
 \delta &\leq \lfloor \frac{\eta - 2 - \lambda}{3} \rfloor
 \end{aligned}
 \tag{5.24}$$

<sup>28</sup>In our toolbox, when  $\delta$  is undefined by the user,  $\delta$  is set by default to  $\Delta + 1$ .

## 5 Composition perception

	$\eta$	$\Delta$	$\delta$
PMLDS3	$\geq 8$	$(1, \lfloor \frac{\eta-5}{3} \rfloor)$	$(2, \Delta + 1)$
PMLDS4	$\geq 8 + \lambda$	$(1, \lfloor \frac{\eta-5-\lambda}{3} \rfloor)$	$(2, \Delta + 1)$

Figure 5.15: PMLDS combination constraints.  $\lambda$  is the maximum distance between indices  $j$  and  $k$  for PMLDS4.

Fig.5.15 summarizes all new constraints. The priority is still given to the selection of  $\Delta$ , which has a lower bound of 1. We can then compute the minimal  $\eta$  to have a solvable system with:

$$\begin{aligned} \lfloor \frac{\eta - 5 - \lambda}{3} \rfloor &\geq 1 \\ \eta - 5 - \lambda &\geq 3 \\ \eta &\geq 8 + \lambda \end{aligned} \tag{5.25}$$

We remark that  $\lambda$  excessively drives PMLDS4 constraints. We believe it makes this alternative almost impractical. In addition, for our experiment on compositions, we only use the triplet version, which appears more natural and easier to understand for participants. So, we will now focus on the implementation of PMLDS3 only.

To generate PMLDS3 combinations, the idea is to use the conventional algorithm for combinations without replacement, but with a virtually larger  $\eta$ . Then, we just have to filter combinations where the first index is  $\geq \eta - 1$ , and combinations violating  $\Delta$  and  $\delta$  constraints. Finally, applying the modulo operator with  $\eta - 1$  on indices produces the expected set of combinations.

However, we cannot build the useful matrix  $\mathbf{X}$  similarly as for the MLDS. First, based on combinations indices, we can only construct an  $\mathbf{X}'$  of shape  $(T \times \eta - 1)$ . Matrix multiplication is then only possible with a fitting vector  $\beta' = [\beta_0, \dots, \beta_{\eta-2}]$ , excluding  $\beta_{\eta-1}$ . In addition, valid combinations in the PMLDS3 context, such as  $[s_{\eta-3}, s_{\eta-2}, s_0]$  and  $[s_{\eta-3}, s_0, s_2]$ , would not produce the right difference values by computing  $\mathbf{X}'\beta'$ , because of the ordering violation. As an illustration:

$$\begin{aligned} \beta_{\eta-3} - 2\beta_{\eta-2} + \beta_0 &= (\beta_{\eta-3} - 2\beta_{\eta-2} + \beta_{\eta-1}) - \beta_{\eta-1} \\ \beta_{\eta-3} - 2\beta_0 + \beta_2 &= (\beta_{-2} - 2\beta_0 + \beta_2) + \beta_{\eta-1} \end{aligned} \tag{5.26}$$

So, for the first combination, the resulting difference would lack one  $\beta_{\eta-1}$ , and in the second case, it would have one extra  $\beta_{\eta-1}$ . We should therefore compute a vector  $\mathbf{y}$  correcting the periodicity, and corresponding to coefficients of  $\beta_{\eta-1}$ . Then, a very smart way to implement this correction is to concatenate  $\mathbf{y}$  with  $\mathbf{X}'$ , as the last column of a new matrix  $\mathbf{X}$ . Finally,  $\mathbf{X}\beta$  is working as expected, and it can be used with the fitting procedures of MLDS described in Subsection.5.2.Fitting methods. Algorithm.5.1 details the complete method to generate the experimental combinations and the periodicity correction vector.

**Algorithm 5.1:** PMLDS3 combinations and periodicity correction vector

---

**In:** ·  $\eta$ , the number of unique stimuli + 1 ( $\geq 8$ )  
·  $\Delta$ , combination indices maximum difference of difference  
·  $\delta$ , combination indices maximum difference  
**Out:** · *combs*, list of combination indices  
· *period*, a periodicity correction vector

---

*combs*  $\leftarrow$  compute combinations without replacement with  $\binom{\eta-1+2\delta}{3}$   
*selection*  $\leftarrow$  a binary vector of the length of *combs*  
**for** each *s*, *comb* in *selection*, *combs* **do**  
  |  $s \leftarrow \text{comb}[0] < \eta - 1$   
  |  $s \times \leftarrow |\text{diff}(\text{diff}(\text{comb}))| \leq \Delta$   
  |  $s \times \leftarrow \prod (\text{diff}(\text{comb}) \leq \delta)$   
*combs*  $\leftarrow \text{combs}[\text{selection}]$   
*period*  $\leftarrow$  a vector of the length of *combs*  
**for** each *p*, *comb* in *period*, *combs* **do**  
  |  $p \leftarrow \text{diff}(\text{diff}(\text{comb} \geq \eta - 1))$   
*combs*  $\leftarrow \text{combs} \bmod \eta - 1$   
**return** *combs*, *period*

---

### 5.3 Interpolation – Experimental results

We have now the right framework to explore the latent space *smoothness*, which is a perceptual characteristic accessible locally. We hypothesized that density homogeneity in the latent space is essential to the perceptual quality of interpolations in this representation. This aspect of generative model has been often qualitatively commented, but rarely verified quantitatively. We also remarked that constant densities can only be found on hyperspheres, with a maximum density at  $\text{mode}(\chi_{16})$ . Circular slices of this mode hypersphere will therefore constitute our first condition. The second condition will be orthogonal, i.e. traveling along the norm, and should cause significant perceptual distortions because of the density variations.

#### Theoretical perceptual scale

We can easily hypothesize the theoretical perceptual scale along a circular slice of the mode hypersphere. If there is no distortion, the perceptual scale must be linear. But what happens along the norm? How to compute theoretical perceptual scales in general? First, we have seen that our sensory model is supposed to uniformize the uncertainty of our internal representation (see Subsection.5.2.Thurstonian scaling). Alternatively, we could say that we want our psychological representation to encode the maximum information about stimuli *s*. In Subsection.1.2.Structural and functional complexity, we have quantified information as the entropy of a random variable: the more unexpected an event is, the more informative about the situation it is. So,



## 5 Composition perception

our sensory model aims at maximizing the entropy of  $\psi$ . As  $\psi$  is defined over a bounded space  $[0, 1]$  (see Subsection.5.2.Perceptual scaling), we know that  $\psi$  will have maximum entropy if  $\psi$  is uniformly distributed. It can be formally written as:

$$F_\psi(\psi) = \int_0^\psi p_\psi(u) du = \psi \quad (5.27)$$

with  $F_\psi(\psi)$  the cumulative density function of  $\psi$ . Then, we know that  $\psi = \Psi(s)$ , but to relate  $p_\psi(\psi)$  and  $p_s(s)$ , we have to use the formula of change of variables in a probability density function.

$$p_\psi(\psi) = p_s(\Psi^{-1}(\psi)) \left| \frac{d}{d\psi} \Psi^{-1}(\psi) \right| = p_s(\Psi^{-1}(\psi)) \left| \frac{1}{\Psi'(\Psi^{-1}(\psi))} \right| = \frac{p_s(\Psi^{-1}(\psi))}{\Psi'(\Psi^{-1}(\psi))} \quad (5.28)$$

with  $\Psi$  a strictly monotonically increasing function, making  $\Psi' > 0$ . Next, we combine this result with Eq.5.27, so that:

$$F_\psi(\psi) = \int_0^\psi \frac{p_s(\Psi^{-1}(u))}{\Psi'(\Psi^{-1}(u))} du = \int_0^{\Psi^{-1}(\psi)} p_s(v) dv = F_s(\Psi^{-1}(\psi)) = \psi \quad (5.29)$$

and,

$$\Psi = F_s \Leftrightarrow \Psi' = p(s) \quad (5.30)$$

In other words,  $\Psi$  is the integration of the prior stimuli distribution  $p(s)$ .

What does  $p(s)$  represent in our two experimental conditions (namely *circle* and *norm* conditions, with subscripts *c* and *n* for short)?  $s$  is a scalar driving the definition of some latent variables  $\mathbf{z}_s$ , finally generating compositions  $\mathbf{x}_s$  through the decoder  $P$ . Fig.5.16a plots  $p(s)$  distributions for our two conditions. On the hypersphere, the distribution is uniform, no matter the real trajectory on this surface. With  $s_c$  in the range  $[0, 1]$ , we have therefore  $p(s_c) = 1$ . Concerning the *norm* condition, we know that  $p(\|\mathbf{z}_s\|) = \chi_{16}$ . To span a pertinent range of  $\|\mathbf{z}_s\|$  values, we set  $\|\mathbf{z}_s\| \in [\text{CDF}_{\chi_{16}}^{-1}(0.01), \text{CDF}_{\chi_{16}}^{-1}(0.99)]$  for  $s \in [0, 1]$ . So,  $p(s_n)$  is simply a rescaled version of a  $\chi_{16}$ . Theoretical perceptual scales are then the corresponding rescaled cumulative density functions, as shown in Fig.5.16b.  $\Psi_c$  is linear and  $\Psi_n$  appears to be s-shaped.

### Stimuli generation

In the following subsection, we will consider unrescaled physical and psychological spaces and assume a linear perceptual scale, so that we can use original physical units in both domains. In Subsection.5.1.Threshold estimation, we have estimated the compositional angular JND as  $2^\circ$  for spherical interpolations. For two stimuli separated by one JND, we can write Eq.5.10 as:

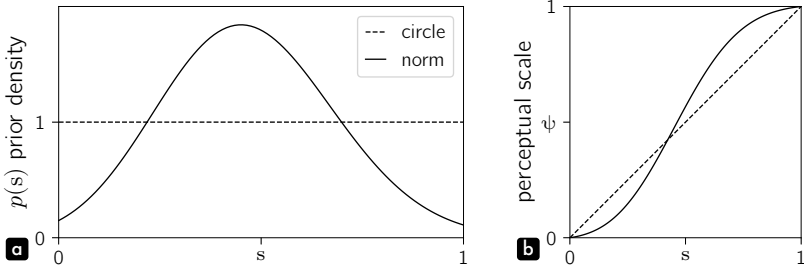


Figure 5.16: Theoretical perceptual scale for *circle* and *norm* conditions (panel **b**). Panel **a** shows  $p(s)$  prior densities.

$$\begin{aligned}
 \psi_{a+1} - \psi_a &= \Phi^{-1}(p(m_{a+1} > m_a))\sigma\sqrt{2} \\
 s_{a+1} - s_a &= \Phi^{-1}(0.75)\sigma\sqrt{2} \\
 \sigma &= \frac{JND}{\Phi^{-1}(0.75)\sqrt{2}} \\
 \sigma &\approx 2.10
 \end{aligned} \tag{5.31}$$

We have also selected MLDS combination constraints  $\Delta = 1, \delta = 2$  to shorten task durations with participants. The optimal ratio between  $2\sigma$  and the step size expressed in angle ( $\phi$ ) is for  $\tau = 1$  (see Subsection.5.2.Combinations optimization).

$$\tau = \frac{2\sigma}{\phi} = 1 \quad \text{and} \quad \phi = 2\sigma \approx 4.19^\circ \tag{5.32}$$

In practice, we rounded this value to  $\phi = 4^\circ$ . The next question is how to compute the number of stimuli  $\eta$  for each condition ( $\eta - 1$  unique stimuli in the *circle* condition). First, concerning the *circle* condition, a full rotation is  $360^\circ$ , and it would basically imply  $\eta - 1 = 90$ , which would be too long for a PMLDS3 experiment (360 combinations with the chosen constraints). We need to reduce this picture, while keeping a  $4^\circ$  angular distance between samples. Actually, circular slices of the mode hypersphere do not have to pass through the center of the latent space to guarantee that the norm of each stimulus is  $\text{mode}(\chi_{16})$ . We can offset the slicing plane in any orthogonal direction to reduce the circumference of the resulting circle until we obtain a chosen  $\eta - 1$  number of unique stimuli separated by  $4^\circ$ . Fig.5.17 illustrates this strategy. The procedure is therefore to first sample three  $\mathbf{z}$  defining a 3-d orthogonal subspace, with basis  $[\bar{\mathbf{z}}_0, \bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2]$  (large empty dots). We generate  $\eta - 1$  equally spaced stimuli by a  $360^\circ$  spherical interpolation according to  $\bar{\mathbf{z}}_1$  and  $\bar{\mathbf{z}}_2$  with radius  $\text{mode}(\chi_{16})$  (small empty dots). Then, we just have to rotate each stimulus of an angle  $\alpha$  toward  $\bar{\mathbf{z}}_0$ , in order to obtain final stimuli (black dots). These stimuli  $\mathbf{z}$  finally generate compositions

## 5 Composition perception

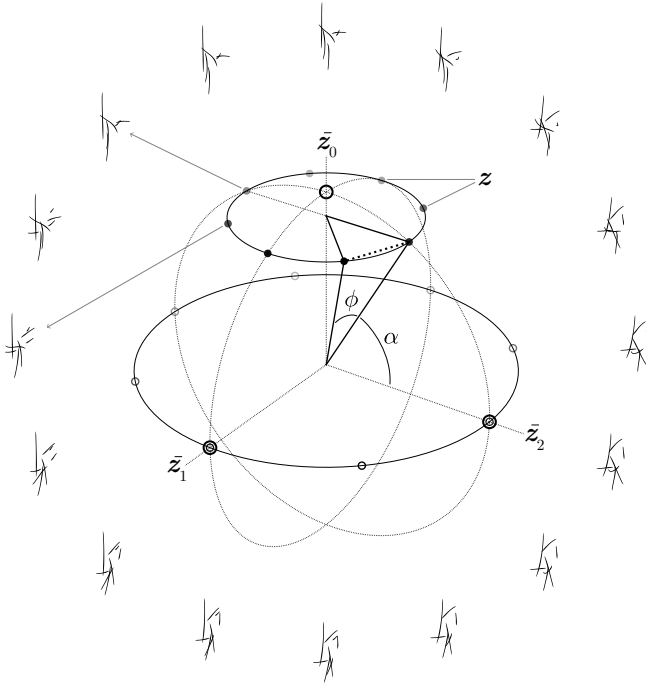


Figure 5.17: Stimuli generation, *circle condition*. In the center, we illustrate the procedure to sample a chosen  $\eta - 1$  number of unique stimuli  $\mathbf{z}$ , of norm  $\|\mathbf{z}\| = \text{mode}(\chi_{16})$ , and uniformly separated by an angle  $\phi$ . In the periphery, the experimental set *circle b* is displayed circularly.

through the decoder  $P^{29}$ . The chord materialized by the thick dotted line is shared by two angles from different circles (thick black lines). A chord length can be computed by  $2r \sin(\frac{\theta}{2})$ , with  $r$  the radius, and  $\theta$  the angle defining the chord.  $\alpha$  angle is therefore expressed as follows ( $\phi$  in radian):

$$2 \sin(\alpha) \text{mode}(\chi_{16}) \sin\left(\frac{\pi}{\eta - 1}\right) = 2 \text{mode}(\chi_{16}) \sin\left(\frac{\phi}{2}\right) \quad (5.33)$$

$$\alpha = \arcsin\left(\sin\left(\frac{\phi}{2}\right) / \sin\left(\frac{\pi}{\eta - 1}\right)\right)$$

We are now able to freely choose  $\eta_c - 1 = 16$ , corresponding to 64 combinations.

Concerning the *norm* condition, distance between stimuli cannot be expressed in angles. However, we can transpose an angle to a chord length at the chosen

<sup>29</sup>Lines of generated compositions are slightly stroked for a better visual result. The method will be detailed in Section.6.2 (Algorithm.6.3), and the chosen stroke profile will be described in Algorithm.6.1.

radius. Then, we can compute the required number of stimuli  $\eta_n$  to span selected  $\|\mathbf{z}_s\|$  range with:

$$\eta_n = \frac{\|\mathbf{z}_s\|_{range}}{2r \sin(\frac{\phi}{2})} + 1 = \frac{\text{CDF}_{\chi_{16}}^{-1}(0.99) - \text{CDF}_{\chi_{16}}^{-1}(0.01)}{2 \text{mode}(\chi_{16}) \sin(\frac{\phi}{2})} + 1 \approx 13.01 \quad (5.34)$$

$\eta_n$  must be an integer, so we chose  $\eta_n = 13$ , producing 40 combinations.

Finally, for each condition, we randomly<sup>30</sup> sampled 3 locations in the latent space. Fig.5.18 shows the resulting 6 conditions. The set, denominated *circle b*, is also presented circularly in Fig.5.17.

### Online experiment details

With the uncertainty conveyed by the Covid-19 pandemic, we decided to move our experiment online. It was possible because our study does not require an extreme control of timings and display hardware. So, data were collected on an online experimental platform, hosted on my website<sup>31</sup>. It has been developed by myself, mainly in Python and JavaScript. These tools require more serious web development skills compared to toolboxes such as jsPsych<sup>32</sup>, but the advantage of such custom platform is to possess a dedicated SQL database. It means that data recording is not operated in the web browser of the user, but instantly posted to the website database. There is therefore less possibility of data loss if the user browser quits unexpectedly for technical reasons, or mishandling of the participant. In addition, data do not have to be sent by mail at the end of the session, they are already centralized, structured and secured. Local copies can be done instantly, and then the analysis is straightforwardly conducted in python. Finally, being able to build your own platform gives you the full control on task implementations.

Most participants were recruited on Prolific<sup>33</sup>. We prefiltered participants to be fluent in English, located in UK/USA, and to have a vision corrected to normal. Before accessing the experimental platform, participants were asked to specify their age, as well as their general knowledge of art material. On the final selection (after attention checks), the age range is from 19 to 60, with a mean at 36 y.o. Reported art knowledge is mostly *little* (level 2 on 4, for 10 of the participants). The second step of the experiment was a calibration of the physical size of displayed elements.

<sup>30</sup>We actually sampled a larger number of sets, and I personally refined the selection with artistic criterion.

<sup>31</sup>Experiments are still available on my website: <https://plelievre.com/experiments>. This online platform is powered by the Python web development framework Django. It enables the creation of dynamical content, e.g. user logging and database reading/recording with PostgreSQL. The visual interface is based on Bootstrap and usual *html*, *css*, *js* web programming languages. Finally, the web app is deployed with Dokku.

<sup>32</sup>jsPsych is a JavaScript framework for creating behavioral experiments. Please find more information at: <https://www.jspsych.org>

<sup>33</sup>Please find details on this platform at <https://www.prolific.co>. Some supplementary participants were directly recruited in France. Nonetheless, they used the same remote platform as Prolific participants.

## 5 Composition perception

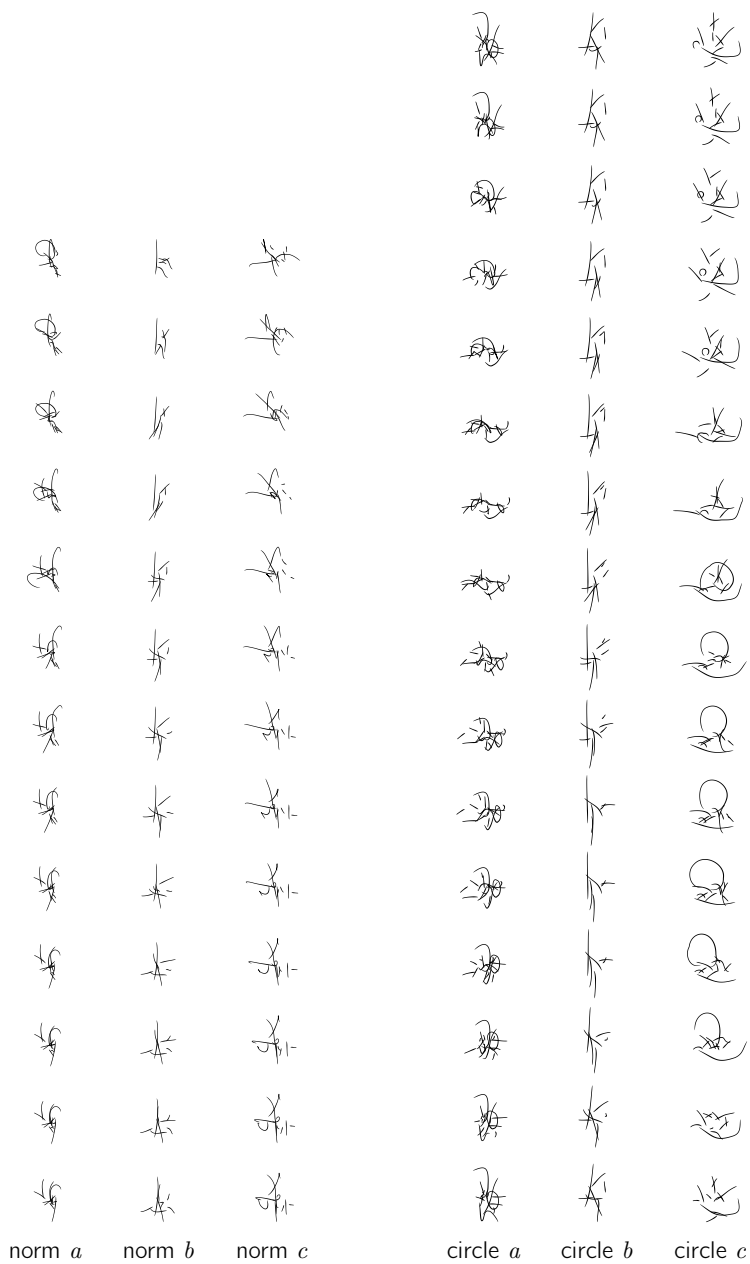


Figure 5.18: Experimental stimuli.

## 5.3 Interpolation – Experimental results

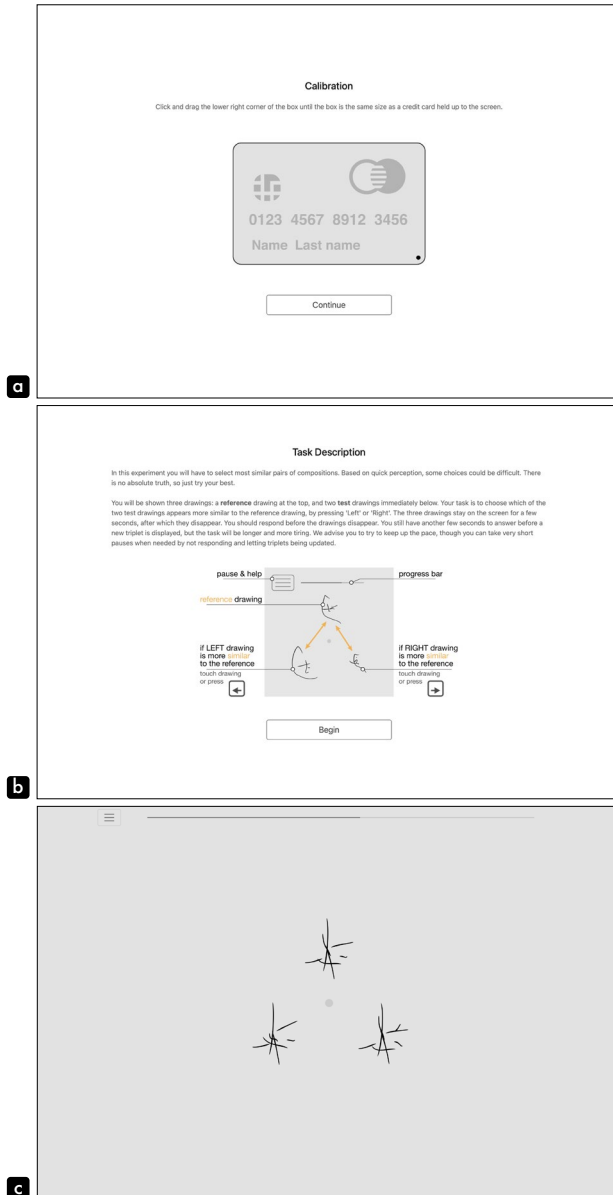


Figure 5.19: Online experiment interface. We present the physical size calibration step (a), the instructions (b), and the main experiment interface (c).

## 5 Composition perception

As shown in Fig.5.19a, we used the *credit card* procedure. Inspired from jsPsych, the idea is to manually adjust a frame to fit the size of a real credit card held on the screen. This way, we can guarantee that stimuli fit a  $4\text{cm}^2$  area. Inside each unit, individual compositions were centered to limit the distraction of drawings positioning noise. However, they were not rescaled. Otherwise, we would have disrupted the direct link between compositions and their  $\mathbf{z}$  representation in the latent space. Next, participants were presented the following instructions:

In this experiment you will have to select most similar pairs of compositions. Based on quick perception, some choices could be difficult. There is no absolute truth, so just try your best.

You will be shown three drawings: a *reference* drawing at the top, and two *test* drawings immediately below. Your task is to choose which of the two test drawings appears more similar to the reference drawing, by pressing *Left* or *Right*. The three drawings stay on the screen for a few seconds, after which they disappear. You should respond before the drawings disappear. You still have another few seconds to answer before a new triplet is displayed, but the task will be longer and more tiring. We advise you to try to keep up the pace, though you can take very short pauses when needed by not responding and letting triplets being updated.

We find it useful to explain/warn participants that there were no *true* responses. During preliminary trials, some participants almost took twice as long as expected. They reported that they were actually waiting for stimuli to reappear, not to *mistake*. Instructions were accompanied by the illustration replicated in Fig.5.19b.

Main task interface is displayed in Fig.5.19c. We used the triplet versions of MLDS and PMLDS, where the reference stimulus  $s_j$  is the stimulus shared by both evaluated pairs  $(s_i, s_j)$  and  $(s_j, s_k)$ . In total, one block corresponds to 312 combinations ( $3 \times (c:64 + n:40)$ ). Each block was then repeated 4 times. This figure sounds small, but with such complex stimuli, our memory seems to be more involved than with elementary stimuli. Preliminary experiments showed too stereotyped responses with more repetitions. In practice, when observers remembered a triplet, or the rule that triggered their initial judgment, they reported preferring to stay coherent with themselves, rather than judging these stimuli again without *a priori*. It was certainly biasing the confusion rate under scrutiny. Beyond the number of repetitions, it is also important to intertwine these repetitions. Of course, combinations and stimuli left/right positions were also fully randomized. Concerning timings, each triplet was at most visible during 2.5sec. Participants could respond before the triplet disappears and up to 2.5sec after. A new triplet was triggered directly after a response or after the timeout, but in this case the current combination was put back in the *to-do* pile. The whole experiment was around 45min long. Participants could make short pauses whenever needed by not responding, and letting triplets being updated, but they were also invited to have breaks between block repetitions. A bar at the top of the interface could finally help them to monitor their progression.

At the beginning of each block, we also added attentional checks, as advised for online experiments. We need to be able to filter out participants poorly involving in the task. So, we introduced 4 triplets with simple vertical and horizontal lines. Answers could be made without ambiguity, but the pace of the task was making some observers surprised by the change in the stimuli nature<sup>34</sup>, and possibly failing attentional checks on a few trials. They had obviously non-random behaviors in alignment with other participants, but still could fail on these simple stimuli. So, we have set a tolerance on the attentional check to be over 75% correct. In our case, there were 16 attentional trials throughout the task, so they could fail up to 3 times. This way, we excluded 6 participants, obviously randomly responding. The following results are thus based on 15 participants.

## Results

In Fig.5.20a-c and Fig.5.21a-c, we first look at inter-participant variability<sup>35</sup>. In these figures, gray lines show individual perceptual scales. They are ordered by increasing  $\psi$  values of the middle stimulus of each series. We remark that the general agreement on locations of scaling distortions is surprisingly good. We mean that the local changes in the perceptual scale are mostly shared even if the magnitude between inflection points may vary. As a result, we obtain a mean observer with a relatively small variance (dark lines on the right, with  $\pm 1$  SD within gray areas). Despite the small number of repetitions (4), a simple averaging already captures the general trend of perceptual scales. Secondly, we observe that the *circle* condition demonstrates a tighter variance than the *norm* condition.

However, a greater number of repetitions per combination is recommended for more accurate perceptual scales. As suggested by Thurstone in its case II of the law of comparative judgments<sup>36</sup>, we decided to pool participants. This way, we believe to capture more objective features, the common ground that drives artistic perception beyond personal judgments. These pooled fits are displayed in Fig.5.20d and Fig.5.21d. Results are very similar to the mean observer. We also estimate the goodness of the fit by bootstrapping pooled probabilities obtained for each combination. The 99% confidence intervals are materialized by gray areas.

The theoretical perceptual scales, defined at the beginning of this section, are reproduced with dotted lines. Initial projections appear quite far from actual perceptual scales. In both conditions, there are significant distortions. Nonetheless, along the norm, the general shape is compressive toward the end, i.e. bow shaped over the linear reference, even if the initial theoretical inflection in the opposite direction is missing. The *circle* condition also presents several distortions, but

<sup>34</sup>As reported by participants on preliminary recordings.

<sup>35</sup>The following results have been presented in a poster at VSS2022 in Florida (Lelièvre & Neri, 2022a) and in a talk at VSAC2022 in Amsterdam (Lelièvre & Neri, 2022b).

<sup>36</sup>Thurstone, 1927a.



## 5 Composition perception

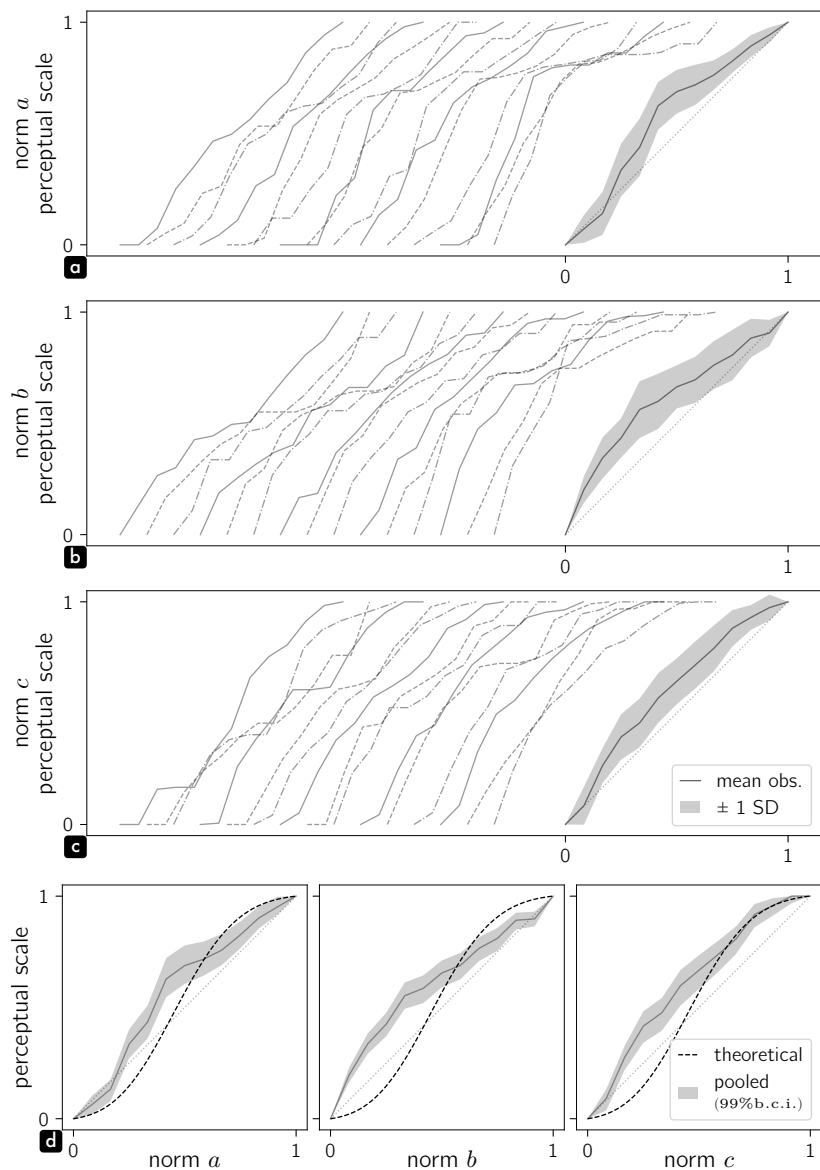


Figure 5.20: Perceptual scaling results for the *norm* condition. Panels **a**, **b** and **c** show inter-participant variability per sub-condition. Individual perceptual scales are plotted in light gray, and the mean observer  $\pm 1$  SD, in dark gray. Panel **d** displays the perceptual scales of all participants pooled together and the 99% bootstrapped confidence intervals of the fits (dark gray lines and areas). Theoretical perceptual scale is materialized by black dotted lines.

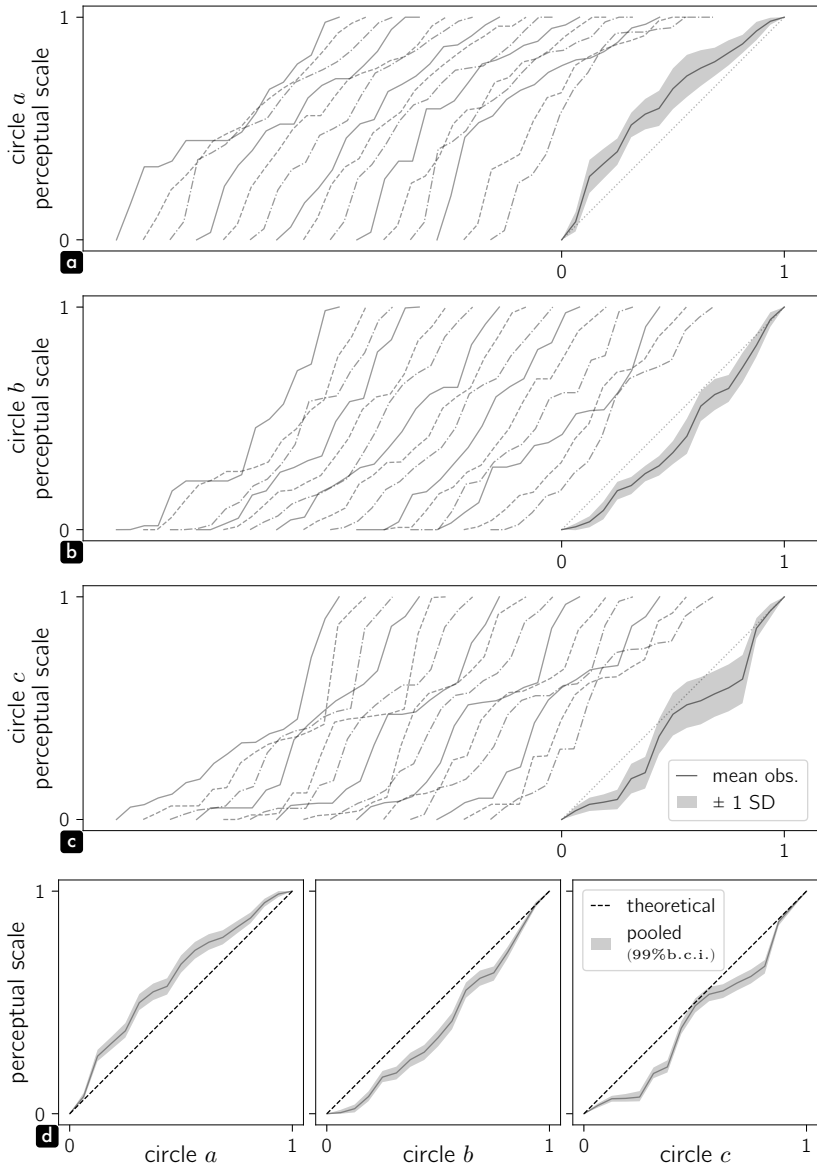


Figure 5.21: Perceptual scaling results for the *circle* condition. Panels **a**, **b** and **c** show inter-participant variability per sub-condition. Individual perceptual scales are plotted in light gray, and the mean observer  $\pm 1$  SD, in dark gray. Panel **d** displays the perceptual scales of all participants pooled together and the 99% bootstrapped confidence intervals of the fits (dark gray lines and areas). Theoretical perceptual scale is materialized by black dotted lines.

## 5 Composition perception

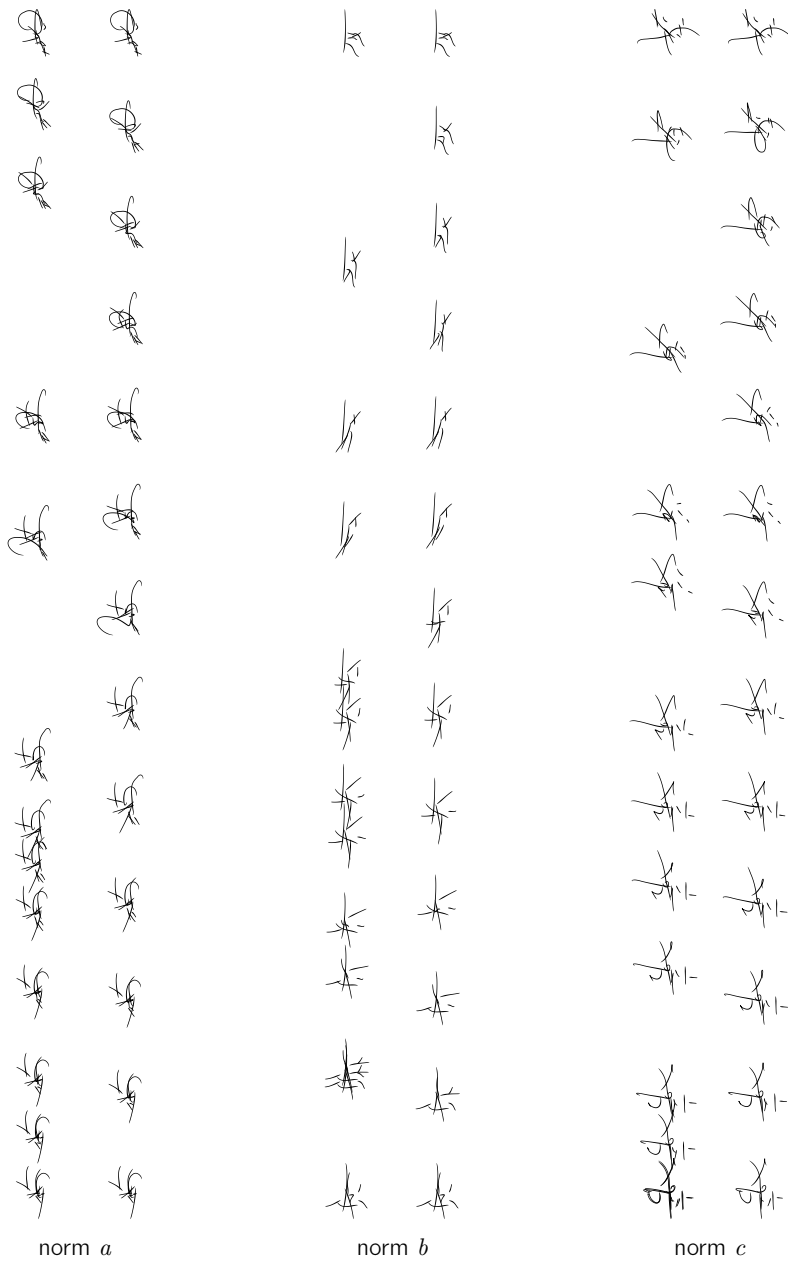


Figure 5.22: Perceptual distortions (left) and resampled inverted scales (right) for the *norm* condition.

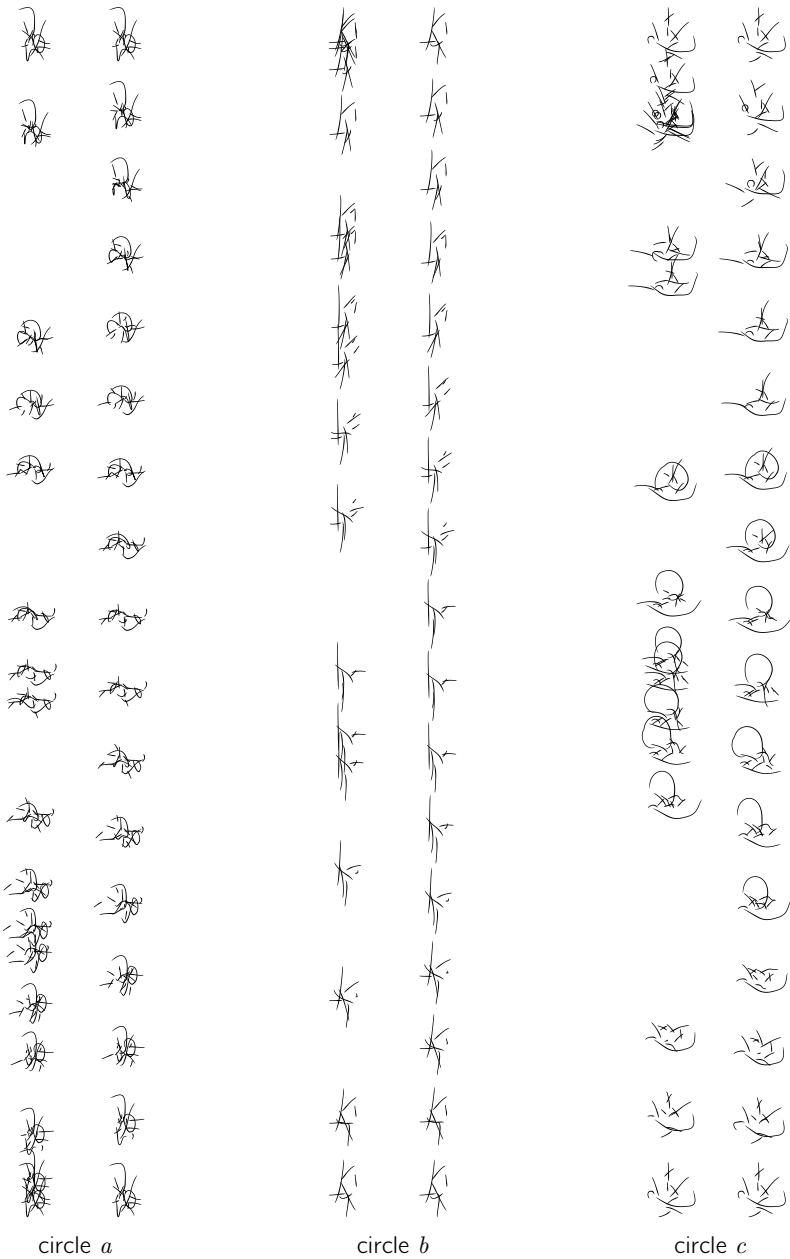


Figure 5.23: Perceptual distortions (left) and resampled inverted scales (right) for the *circle* condition.

## 5 Composition perception

they seem to evolve around the expected linear scale. Actually, plotted functions are periodic and could be represented with different initial points. We could then find similar scales distorted on the opposite side of the dotted line, or optimally centered. In addition, *circle* perceptual scales confirm tighter confidence intervals than the *norm* condition. It seems that perceptual distortions are more easily identifiable in higher density regions of the latent space.

Nevertheless, observed distortions are more intense than expected. Our constraints on the model at training, designed to guarantee the homogeneity of the latent space, and its compliance to the prior, seem to partially fail. Once again, we repeat that the relatively small size of our dataset is unfavorable to achieve this goal. But is this an optimization limitation only, or a more serious discrepancy between model latent representation of compositions and human internal representation? Does the model captures regularities different from those essential to our perception? It is maybe too early to take a strong position. Indeed, we have seen that the model extract several types of information and measurements beyond  $\mathbf{z}$  locations in the latent space (see Chapter.4.3). We should find a way to exploit encoding and decoding uncertainties and this opportunity will be explored in the next subsection.

Left vertical lines of Fig.5.22 and Fig.5.23 show a direct interpretation of distortions measured with perceptual scaling. Stronger derivatives on  $\Psi$  present in Fig.5.20d and Fig.5.21d produce dilation of distances between stimuli, while smaller slopes imply contractions. Based on this simple illustration, it is generally possible to agree on discovered warpings. Compositional similarity judgments seem therefore to rely on quite objective mechanisms.

Measuring distortions also provide the necessary information to correct them. Inverted perceptual scales  $\Psi^{-1}$  can be easily computed. However, with the objective to *smoothen* interpolations, we need to enforce a  $C^1$  continuity around measured stimuli/representation couples  $(\psi, s)$ . To do so, a linear interpolation between values is followed by a Gaussian filtering, i.e. a convolution with a Gaussian kernel. For the *circle* condition, we especially take care of the periodicity to obtain the right derivatives at  $s_0 \equiv s_{\eta-1}$ . Results of this procedure are displayed on the vertical lines on the right of each condition in Fig.5.22 and Fig.5.23. The granularity of the presented resampled sequences is limited for legibility, but corrected interpolations at higher *frame rates* is possible to generate *smooth* animations. Indeed, beyond scientific usages, the MLDS framework will become an important material for artistic explorations (see Chapter.6).

Let us now investigate fitted  $\sigma$  values. In both *norm* and *circle* conditions, step sizes have been set to  $4^\circ$ . But rescaled with different  $\eta$  number of stimuli, fitted  $\sigma$  are not directly comparable anymore. This is one of the reason we introduced  $\tau = 2\sigma(\eta-1)$  in the previous section (see Definition.1). Therefore, Fig.5.24 plots  $\tau$  values for the pooled participant, and the associated 99% bootstrapped confidence intervals for all conditions. Confidence intervals are large, so observed relative

### 5.3 Interpolation – Experimental results

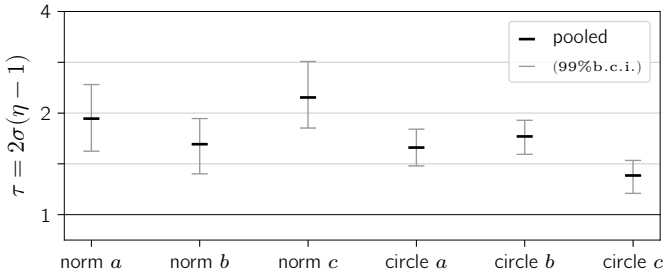


Figure 5.24: Fitted perceptual noise standard deviation of all participants pooled together, and expressed as  $\tau$  (see Definition.1 for details about  $\tau$ ). The 99% bootstrapped confidence intervals of the fits is indicated in gray.

differences may not be significant. However, *norm b*, which is sampled along a norm passing through the center of *circle b*, presents coherent  $\tau$  values. Relative  $\tau$  may be therefore interpreted as compression/dilation from the center of the latent space, and to some extent reflect lower frequency distortions of the latent space, close to what we intended to measure in Subsection.5.1.Multidimensional scaling, and Subsection.5.1.Isomap. At the scale of an interpolation, with a bounded stimuli range, smaller is  $\tau$ , smaller is the confusion, and higher must be the perceptual amplitude of changes. In other words, relative  $\tau$  values give an idea of the rate of changes, the *speed*, of perceived interpolations. Smaller is  $\tau$  and quicker, more intense is the interpolation.

This observation may be related to the notion of NDL (Number of Discriminative Levels) introduced with TS on unordered stimuli<sup>37</sup>. For the authors, the measured perceptual range can be interpreted in terms of number of JND, rendering NDL some sort of absolute psychological unit. However, as we said multiple times,  $\sigma$  is only relevant in the psychological space, unless  $\Psi$  is fully characterized. In the specific case of stimuli with unknown positions in the physical space,  $\Psi$  is necessarily undetermined. At best, we can only assume it linear to transpose JND in the psychological space. For MLDS with rescaled metrics, we could write<sup>38</sup>:

$$NDL = \frac{1}{JND} = \frac{1}{2\sigma\Phi^{-1}(0.75)} \quad (5.35)$$

Nonetheless, the NDL definition is theoretically confusing. When controlled stimuli are available, we think that  $\tau$  is a more straightforward comparative indicator.

A second remark concerns confidences intervals. They are tighter for the *circle* condition, as it was the case for the optimization of  $\Psi$ . We believe it emphasizes the importance of sampling higher density regions of the latent space to produce compositions conveying more reliable and distinctive features.

<sup>37</sup>Wijntjes et al., 2020.

<sup>38</sup>In the original paper, the authors use an 84% JND, leading to  $\Phi^{-1}(0.84) = 1$ .

## 5 Composition perception

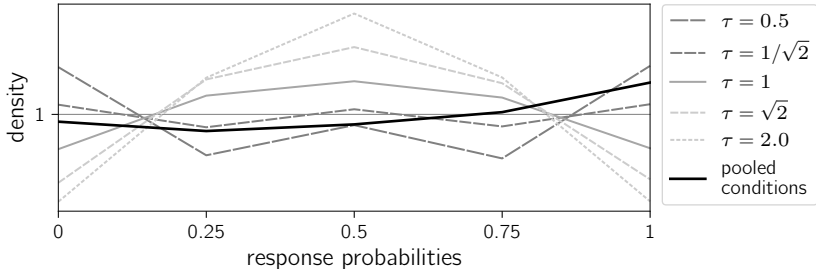


Figure 5.25: The distribution of response probabilities is plotted for all participants and conditions pooled together (black line). We compare this distribution with simulations for different  $\tau$  and a similar number of repetitions, i.e. 4 (gray lines). A linear perceptual scale is assumed for these simulations.

We should also comment the general level of  $\tau$ , which is somewhere between  $\sqrt{2}$  and 2. Let us round it to  $\tau = 2$ . This is over the optimal range  $[\frac{1}{\sqrt{2}}, \sqrt{2}]$  found by simulation (see Fig.5.13a). As result, we should consider adjusting the number of stimuli. We do not change the physical range of stimuli, and  $\sigma$  remains fixed. So, to obtain  $\tau = 1$ , we need a step size twice longer, i.e.  $\eta' = \frac{\eta-1}{2} + 1$ . It would also mean that our estimation of the compositional angular JND of a simple comparative judgment would rather be around  $4^\circ$ . We believe that this figure is over-estimated. If we really had  $\tau = 2$ , the task would be quite challenging, with more responses closer to chance (see Fig.5.12). To verify this supposition, we simulate in Fig.5.25 the distribution of response probabilities for different  $\tau$ , and a limited number of repetitions, i.e. 4, as in the real experiment (gray lines). We assume a linear perceptual scale. Then, we pool response probabilities of all participants and conditions (black line). The result is barely flat<sup>39</sup> and corresponds to a value of  $\tau$  closer to  $\tau = \frac{1}{\sqrt{2}}$ , in the opposite direction from the direct estimation from fitted  $\sigma$ . This second estimation may be taken with caution too, as our experiment deals with complex  $\Psi$  functions, and variable participant behaviors. Looking back at Fig.5.13b, we observe that  $\sigma$  fitting error is higher for higher  $\tau$ . Our hypothesis is thus that the estimation of  $\tau$  suffers from an error amplification as soon as  $\tau > 1$ . In summary, our  $\tau$  is probably slightly above 1, but much smaller than 2. The step size optimization, and the associated verification procedures, are still open questions which would require further investigations and simulations to evolve from heuristics to proper methods. A good start for future works would be to address simpler and well known stimuli, already validated with traditional methods such as 2AFC for psychometric functions extraction.

<sup>39</sup>The distribution is actually slightly higher toward 1. This can be easily explained by the average nature of the fitted perceptual scaling functions  $\Psi$ . Most of them are bowed over the linear function, meaning that right pairs  $(s_j, s_k)$  are usually more similar than left pairs  $(s_i, s_j)$ , biasing  $p(\mathbf{m}_{j,k} < \mathbf{m}_{i,j})$  toward 1. (Note that our convention, *more similar*, is the opposite to the canonical MLDS definition in Eq.5.17.)

### Perceptual scale prediction from Fisher information

We have seen how to measure and correct distortions in the latent space. This procedure is simple and well suited for an artistic use, but it is still too long and expensive to be applicable at a larger scale, i.e. to characterize the whole latent space. In addition, at each model iteration, the experimental work would have to be done again. This is actually a reason why we collected 15 participants only. This figure is sufficient, knowing that it constitutes a proof of concept, a verification circumstantial to a specific model state. As a result, we should find some ways to predict perceptual distortions from information contained in the model itself. We could then automatically improve interpolations, and even use this prediction tool at an earlier stage during training, in the form of a regularizer. The following work is therefore essential, promising, but also really preliminary. Initial insights on the method should be credited to our colleague Jonathan Vacher. We would like to warmly thank him for sharing its mathematical development on Fisher information for sensory models<sup>40</sup>.

Fisher information is a method to quantify the amount of information that a random variable conveys about a parameter upon which its distribution depends. Transposed to our sensory system, Fisher information may be interesting to estimate how much information a measurement  $m$  carries about its triggering stimulus  $s$ . Formally, Fisher information  $\mathcal{J}$  is expressed as the variance of the score of a distribution  $p(m|s)$ :

$$\mathcal{J}(s) = \mathbb{E}_{m \sim p(m|s)} \left[ \left( \frac{\partial}{\partial s} \log(p(m|s)) \right)^2 \right] \quad (5.36)$$

Remembering that  $\psi$  is deterministically defined by  $\psi = \Psi(s)$ , we can write:

$$\begin{aligned} p(m|s) &= p(m|\Psi(s)) \\ \log(p(m|s)) &= \log(p(m|\Psi(s))) \\ \frac{\partial}{\partial s} \log(p(m|s)) &= \Psi'(s) \frac{\partial}{\partial s} \log(p(m|\Psi(s))) \\ \mathbb{E}_{m \sim p(m|s)} \left[ \left( \frac{\partial}{\partial s} \log(p(m|s)) \right)^2 \right] &= \Psi'(s)^2 \mathbb{E}_{m \sim p(m|s)} \left[ \left( \frac{\partial}{\partial s} \log(p(m|\Psi(s))) \right)^2 \right] \\ \mathcal{J}(s) &= \Psi'(s)^2 \mathcal{J}(\Psi(s)) \end{aligned} \quad (5.37)$$

In Subsection.5.2.Thurstonian scaling, we have seen that assuming Thurstonian case V makes  $\Psi$  derivatives tuning our sensitivity to physical stimuli in order to uniformize the uncertainty of our internal representation. In Section.1.2, we have defined information as the amount of uncertainty raised by the observation of a particular

<sup>40</sup>Publications on this theoretical work, associated proofs and experimental demonstrations are ongoing.



## 5 Composition perception

state of a random variable. In other words, our sensory model aims at keeping a constant amount of information about stimuli in our representation, no matter the stimulus, i.e. making  $\mathcal{J}(\Psi(s))$  constant for all  $s$ . This condition is fulfilled, if and only if:

$$\Psi(s) \equiv \int_0^s \sqrt{\mathcal{J}(u)} du \quad (5.38)$$

Perceptual scale can therefore be predicted from the Fisher information computed on stimuli. The relation is only an equivalence up to a linear transformation, because of the integration and the rescaling of  $\Psi$  in the range  $[0, 1]$ , set by convention.

In order to predict perceptual scales from our composition model, we need to find some substitutes for  $p(\mathbf{m}|s)$ . A stimulus  $s$  primarily drives the definition of a latent variable  $\mathbf{z}_s$ , deterministically decoded as a composition  $\mathbf{x}_s$ .  $\mathbf{m}$  is the distribution of the psychological representation given  $s$ , so it corresponds almost straightforwardly to the encoding distribution by  $Q$  of  $\mathbf{z}$  knowing  $\mathbf{x}_s$ , i.e.  $q(\mathbf{z}|\mathbf{x}_s)$ , which can be rewritten as  $q(\mathbf{z}|\mathbf{z}_s)$ . We are actually using the model with inverted encoder and decoder, as a *decoder-encoder*. In summary, we select a stimulus  $s$  producing a latent variable  $\mathbf{z}_s$ , decoded into a  $\mathbf{x}_s$  and finally re-encoded in the latent space as a random variable with some distribution  $q(\mathbf{z}|\mathbf{z}_s)$ , being a multivariate normal distribution. We call this version, the *encoder* version.

A *decoder* version is also possible. We have just said that decoder  $P$  is usually used deterministically to generate compositions  $\mathbf{x}_s$  from any  $\mathbf{z}_s$ . However, in Section.4.3, we have shown that the model captures uncertainty of stroke positions and shapes at each time step (see specifically Fig.4.31c,d). Uncertainty in  $P$  is therefore the uncertainty to produce a physical representation from an internal variable. From this perspective, it seems fundamentally in opposition with  $p(\mathbf{m}|s)$ . Indeed, we believe that if  $P$  is uncertain in producing an  $\mathbf{x}$  from a  $\mathbf{z}$ , it is because  $\mathbf{z}$  is usually poorly determined by the encoder  $Q$ . Then,  $p(\mathbf{x}|\mathbf{z}_s)$  is probably echoing the targeted distribution  $p(\mathbf{m}|s)$ . On practical concerns,  $p(\mathbf{x}|\mathbf{z}_s)$  is a very complex combination of different components. Decoder output is a sequence of  $t$  instants of the form  $\mathbf{x}_t = [\mathbf{p}_t, \mathbf{s}_t, \beta_t]$  (see Section.3.3 for details). So, in order to compute its associated Fisher information, we need to simplify  $p(\mathbf{x}|\mathbf{z}_s)$ . The main idea is to concatenate distribution parameters (i.e.  $\boldsymbol{\mu}$  and  $\boldsymbol{\sigma}$ ) of strokes initial position  $\mathbf{p}_t$  and shape  $\mathbf{s}_t$ , since they are both multivariate normal random variables. In a second time, we concatenate parameters along  $t$  to obtain a unique, but highly dimensional, multivariate normal random variables. Nonetheless, we should consider pen-down strokes only, i.e. when  $\beta_t = 1$ . But  $p(\mathbf{x}|\mathbf{z}_s)$  must be differentiable in  $s$ . Removing pen-up entries would produce final random variables of different dimensionalities for each interpolation step, making them impractical to compute the Fisher information. We could nullify distribution parameters when  $\beta_t = 0$ , but differentiation would be erroneous when a stroke appears or disappears in the series of compositions. To address this issue, we have decided to only include strokes that are always visible across the whole interpolation, and discard the others.

We have now two alternatives to compute the Fisher information from the model. Luckily, in both case, we have distributions of the form  $\mathcal{N}(\boldsymbol{\mu}(s), \boldsymbol{\Sigma}(s))$ , of which the Fisher information has a closed form given by:

$$\mathcal{J}(s) = \frac{\partial \boldsymbol{\mu}(s)^\top}{\partial s} \boldsymbol{\Sigma}^{-1}(s) \frac{\partial \boldsymbol{\mu}(s)}{\partial s} + \frac{1}{2} \text{Tr} \left( \boldsymbol{\Sigma}^{-1}(s) \frac{\partial \boldsymbol{\Sigma}(s)}{\partial s} \boldsymbol{\Sigma}^{-1}(s) \frac{\partial \boldsymbol{\Sigma}(s)}{\partial s} \right) \quad (5.39)$$

In our case, all components of the studied multivariate normal random variable are supposed independent, so that  $\boldsymbol{\Sigma}(s) = \text{diag}(\boldsymbol{\sigma}^2(s))$ . Then:

$$\begin{aligned} \mathcal{J}(s) &= \sum_{k=1}^K \left( \frac{\partial \boldsymbol{\mu}_k(s)}{\partial s} \boldsymbol{\sigma}_k^2(s)^{-1} \frac{\partial \boldsymbol{\mu}_k(s)}{\partial s} + \frac{1}{2} \boldsymbol{\sigma}_k^2(s)^{-1} \frac{\partial \boldsymbol{\sigma}_k^2(s)}{\partial s} \boldsymbol{\sigma}_k^2(s)^{-1} \frac{\partial \boldsymbol{\sigma}_k^2(s)}{\partial s} \right) \\ &= \sum_{k=1}^K \left( \left( \frac{\partial \boldsymbol{\mu}_k(s)}{\partial s} \right)^2 \frac{1}{\boldsymbol{\sigma}_k^2(s)} + \left( \frac{\partial \boldsymbol{\sigma}_k^2(s)}{\partial s} \right)^2 \frac{1}{2\boldsymbol{\sigma}_k^4(s)} \right) \end{aligned} \quad (5.40)$$

For practical reasons, especially tasks duration with participants, we have restricted perceptual scaling experiments to a subset of stimuli per interpolation. However, it is completely possible to *infinitely* sample interpolation trajectories. In Fig.5.26, we therefore present perceptual scales predicted from Fisher information with a 32-fold stimuli granularity. The *encoder* version is plotted with dotted lines and the *decoder* version with solid lines.

The first remark is that the *decoder* version produces surprisingly good results. It is not capturing all ground truth distortions, but the general trend is present, as well as non-trivial local inflections. On the contrary, the *encoder* version is more irregular in matching pooled participants. We only find an improvement compared to the *decoder* version for *circle c*. Otherwise, there are serious spurious artifacts. Our hypothesis is that there is a re-mapping issue.  $\mathbf{z}_s$ , deterministically decoded to  $\mathbf{x}_s$ , should be mainly re-encoded around  $\mathbf{z}_s$ , but it is not guaranteed. It should be an inherent characteristic of the model, but this feedback loop is not enforced at training, and seems to suffer from serious flaws. For instance, we think that the large jump around 0.2 in *norm b* is typically a re-encoding issue, possibly due to the apparition of a small stroke (see drawings 4 and 5 from the top of *norm b* in Fig.5.18). Implementation of a re-encoding regularizer could be investigated in future works. Naturally, this problem is absent from the *decoder* version, making it certainly more reliable. Nonetheless, there is room for improvements, especially in the computation of the Fisher information on series of compositions with changing stroke numbers. Discrepancies between predictions and ground truths is certainly contained in the discarded tailing small strokes. In addition, necessary derivative along  $s$  are for now implemented by a simple neighboring differentiation. We could investigate more accurate computations. In summary, perceptual scale prediction is possible from the Fisher information based on decoder output parameters. Secondly, the development of this procedure as a regularizer, should be placed among our top priorities for future works.

## 5 Composition perception

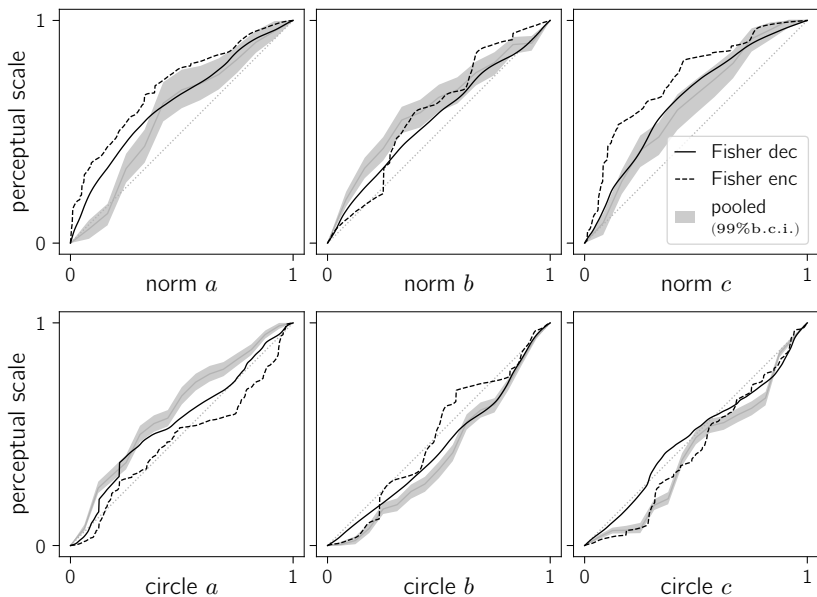


Figure 5.26: Perceptual scale prediction from Fisher information. The *encoder* version, using encoder output parameters, is plotted with a dotted line. The *decoder* version is presented with a solid line. These predictions are compared with pooled participants in gray.

Finally, we can formulate an intermediary answer to whether perceptual scale distortions are real human-model representational discrepancies, or model artifacts. The presented prediction ability of the model based on secondary information seems to show that the model captured important compositional regularities, at least coarsely in alignment with human perception. However, only the next iteration of the model, with an improved latent space thanks to a Fisher regularizer may reveal finer grain representational differences, and establish a more definitive answer. The model and the experimental framework are still in their early stage. They show some drawbacks, but overall, the presented results make this research definitely encouraging and promising to study complex stimuli such as artistic composition.

## 6 Ink and paper

This last chapter is finally dedicated to the artistic investigation of compositional models. This exploration aims at developing visual propositions to convey the expressiveness of the compositional space and communicate the inner dynamics of generated graphical elements. A first difficulty relates to the transposition of captured hidden dimensions. I have already indicated the possibility to make them spatial and temporal, but this chapter is above all the story of a return to the material space, to the humble ink and paper. The temporal ability of digital screens is therefore discarded, and dynamics have to be suggested by other means. In addition, this creative moment is of a completely different nature from the compositional practice of the initial drawings of my dataset. I cannot simply print or redraw by hand compositions generated by the model. I prefer to explore more intensive drawing potentials, only possible with a mechanical pen-plotter. The chapter is thus the convergence of several types of considerations to be addressed simultaneously.

However, at the time of writing this manuscript, this practical research is not finalized yet. Investigations described in previous chapters have monopolized my time, and certainly too much. The following sections are therefore a work in progress: not guaranteed to cohere into satisfactory artworks regarding the objective of communicating my compositional paradigm. Sol LeWitt is somehow comforting when he writes:

The idea itself, even if not made visual, is as much a work of art as any finished product.<sup>1</sup>

Nevertheless, I do not really belong to conceptual art. I am obviously driven by simpler perceptual objectives, i.e. pictorial composition. Similarly, the long and rich history of computer art, together with a proximity of tools and/or visual outcomes can be misleading about my approach. Like in Chapter.1, it will not be the place for a historical review of art made with computers, but a *minima*, I should try to situate my intentions. With the presentation of technical bricks and creation principles, it could finally compensate the lack of hindsight to provide an extended discussion about my upcoming artworks.

---

<sup>1</sup>LeWitt, 1967.

### 6.1 Returning to the material space

Even if I exhibit artworks made of ink on paper, a computer is generating the drawings anyway, or is at least controlling a mechanical pen-plotter. My work will therefore be undoubtedly associated with the wide family of computer art. No matter the exact nature of what is shown, it will be compared to former approaches and visual propositions. Computer art has already a long history, and showing something completely new is challenging. In this section, I thus try to clarify my creative intentions by highlighting steps made aside existing trends, such as generative art or AI art. In a second time, I detail the reasons for using a pen-plotter and fundamentally for returning to the material space.

#### *Somewhere in computer art*

Situating one's intention in former practices is not a question specific to computer art. A computer can ultimately be considered as a tool like any other. Similarly to oil painting used across a wide range of style and epochs, computers do not imply a unique type of artistic preoccupation. The following review is not intended as an exhaustive span of the computer art history, so definitions of the different movements will be partial, as well as the list of associated protagonists.

Generative art is maybe the broader and oldest category of computer arts. Beyond the richness of their works, pioneers of the sixties such as Vera Molnár, Manfred Mohr or Harold Cohen should be appreciated for the very demonstration that computers could be used in non-utilitarian contexts. Access to technologies and development environments were particularly unfavorable. The democratization of computational units and the accessibility of open-source tools make now generative art more popular than ever.

Under generative art, I consider any approach relying on parametric algorithms, where allowed degrees of freedom are disturbed by an additional randomness to generate the diversity of the final artworks. In other words, precisely controlled procedures are given the ability to produce uncontrolled results, with the help of controlled types and amounts of noise.

If we can describe the procedures for expressing an art concept, then we can code those procedures and work with them arithmetically. The description of an art concept is essentially an outline of the decision system that governs the art-making procedures.<sup>2</sup>

For me, the first intrinsic limitation of the approach, as stated by Roman Verostko, is a tendency to limit your art expectations to procedure that you can actually formalize and code. For instance, external randomness will not be able to replicate

---

<sup>2</sup>Verostko, 1990.

the unfathomable regularities you could have expressed freely otherwise, without a computer. Thus, generative art outlines an artist closer to a mechanical designer, whose organic creativity is located in the execution uncertainty. On the contrary, I prefer to see artists as simulators of an internal model, hidden to them, and that an external source of noise on pre-defined rules can only barely replace the intrinsic inner stochasticity. To overcome this aspect, generative artists usually produce large amounts of outputs, and subjectively choose specific outcomes. It therefore adds perceptual mechanisms, enriching the original coded/intended procedure.

Generative art is also sometimes theorized as an epigenesis. We have already evoked this idea in Section.1.2. Computer programs are then considered as simulations of a genetic determinism, where added noises acts as environmental contributions to the development of the artwork *organism*. However, I still believe that implemented randomness can only give the illusion of a structuring effect, particularly regarding *order from order* and *order from disorder* principles of life organization<sup>3</sup>. Even if my models, and their associated compositional space and compositional plane, are indeed probabilistic, the nature of the stochasticity is extremely different. Generative modeling must not be mistaken with *primitive* aspects of generative art. They are somehow antithetical. In generative art, the noise alters a unique structure, and in generative modeling, the noise initiates a diversity of different structures, measured and learned beforehand. In generative modeling, the diversity is already in the model, not in the noise.

However, generative modeling is now at the core of many *Artificial Intelligence* powered generative art. The elaboration of the visual vocabulary can be relegated to the selection of a dataset, and the use of pre-existing deep learning architectures, which are for now essentially based on the GAN architecture<sup>4</sup>. Among leading artists of this trend, we find Mario Klingemann and Robbie Barrat. Despite undeniable talents to tweak and alter visual results, produced artworks present a high level of similarity, and especially of *glitches*. In fact, these artifacts constitutes a unique style, but to which the chosen datasets do not really contribute. In the art history, new techniques have always stimulated artists to build new means of expression. Nevertheless, there is also a risk for intrinsic characteristics of a technique to be *mis*-taken for new forms. Sol LeWitt would say:

Some artists confuse new materials with new ideas.<sup>5</sup>

Novelty is then quickly outdated. I am not saying that current GAN artists fall in this category, but the use of loosely defined datasets, is somehow blurring the

---

<sup>3</sup>For more details, see Atlan, 1972/2006; Schrödinger, 1944/2013

<sup>4</sup>Generative Adversarial Networks (GANs) have been introduced by I. Goodfellow, 2016; I. J. Goodfellow et al., 2014. It consists of a generator network trying to produce images similar to inputs from a dataset, and a discriminator network trying to determine if images presented to it, are *true* or not, i.e. from the generator or the dataset. Both networks are trained alternatively, so that each network task becomes more difficult as the adversarial game progresses.

<sup>5</sup>LeWitt, 1967.

fundamental plastic approach. In the specific case of cited artists, it is not really an issue, as central questions of their work are the place of the machine and the role of the artist. Gregory Chatonsky pertinently points out that AI should stand for *Artificial Imagination*<sup>6</sup> instead of Artificial Intelligence. He means that artists consciously delegate their imagination to machines, which become a neutral box of surprises. Neural networks provide an infinite flux of images, among which artists select and organize a few instances to articulate a particular concept. The artist's role then switches from a creator to a curator. Thanks to his/her sharp sense of selection, the artist works in the manner of a photographer in front of nature, capturing key instants of life. However, this scheme is of little interest to me. My approach to art and composition is through an active creation of the form. The compositional paradigm actually makes me more a biologist studying his own body.

Latest deep generative neural networks, such as incredible DALL-E and MidJourney<sup>7</sup>, are now able to create images from a text description with an extremely fascinating precision in terms of content and style. The question of artistic authorship of produced artworks is then more relevant than ever. This subject drives active discussions among artists and researchers: *Can computer create Art?*<sup>8</sup> *Can Artificial Intelligence Make Art without Artists?*<sup>9</sup> However, do we simply know what is art in the first place? The definition dramatically evolved and extended since Marcel Duchamp. For Sofian Audry & Ippolito:

*Can machines be artists? is the wrong question. We should instead be asking, what roles does machine-made art leave for artists—imagined or real, flesh or silicon—and the viewers who imagine them.*<sup>10</sup>

What arises in the mind of the viewer, and what he/she projects on an artist's work is thus what really matters. The difficulty is then to convey the desired representation (see details at the end of this section). I cherish my code and models, because I have built them step by step. They are a complex object I became familiar with. However, if it disappears, my capacity to draw and compose remains unchanged. Throughout the manuscript, even if I may have objectified the model as an entity, it is a misuse of language. As far as I am concerned, AI is a statistical tool, and nothing more. I am thus closer to older generation of computer artist, such as emblematic Harold Cohen and his art-making robot AARON. For him, robots are interesting as collaborators, but they are not artists. In addition, I have never been convinced, compositionally speaking, by artworks from machines designed to be autonomous artists, e.g. Leonel Moura's artist-robots.

---

<sup>6</sup>Chatonsky et al., 2017.

<sup>7</sup>More information at: <https://openai.com/dall-e-2/> and <https://www.midjourney.com>

<sup>8</sup>Hertzmann, 2018.

<sup>9</sup>Audry and Ippolito, 2019.

<sup>10</sup>Audry and Ippolito, 2019. For a more complete review of questions raised by the use of AI in art, please refer to the comprehensive book from one of the authors, *Art in the Age of Machine Learning* (Audry, 2021).

In between generative art and AI powered creations, artificial life is also interested in a certain form of autonomy. For artists like Alain Lioret, Chu-Yin Chen or the pioneer Yoichiro Kawaguchi, the idea is to consider pictorial elements as living beings. Artworks are then generated by the simulation of these entities' interactions:

**These creatures do not paint themselves on a canvas. Instead, they use their bodies to compose new paintings, in ballets of autonomous movements.<sup>11</sup>**

This idea resonates with the composition seen as the organization of a system, i.e. as conditional constraints between sub-elements on the canvas (see Section.1.2). Then, if *graphical beings* are able to *compose*, it is also related to the phenomenon of emergence, which I would like to explore in future works. However, I do not understand the fundamental necessity to have actual organic forms mimicking life to address the *organic* creation of pictorial forms. In addition, if I am interested in the dynamic of graphical elements, it is of hidden compositional dynamics, not of approximated bio-inspired functional interactions.

Another branch of AI art explores a concept closer to my own practice. The idea is to build a system imitating the work of an existing artist, while exhibiting the process. I can evoke the collaboration between Robbie Barrat and Ronan Barrot, where the first trained a GAN to mimic the few hundreds oil-painted skull still-lives of the second. Resulting pixel-maps are very convincing thanks to the great homogeneity of the dataset. In *Mind the Machine*, Sarah Schwettmann also developed a model to imitate the work of the Shantell Martin. This time, an RNN and a pen-plotter were reproducing her drawing *style*. I think that *style*, rather than composition, is the right wording. The machine is actually initialized with a structuring line, imposing the large arrangement of smaller elements. Also, the artist adapted her graphical language with highly stereotyped details to help the model. The result is interesting for 300 input drawings only, but found regularities were already obvious for the artist, and spectators as well. I personally think that such endeavor is more instructive if something hidden is revealed in the process. Otherwise, the machine becomes accessory.

A derivative use of AI emerges as a support for co-creation. I particularly think about Sougwen Chung, who paints with an AI pre-trained on her own work. However, from the final artworks, which explore fuzzy intricate arrangements of lines, it is difficult to evaluate how much the robot arm is able to perform further than an external source of noise, as a creative constraint. Her approach is relevant in questioning human/machine creative relationship, but I am dubious about this kind of proposal concerning compositional aspects. I believe that the artist's compositional abilities can compensate any *strange* action of the robot. Composition is intrinsically resilient enough when reaching this level of visual complexity.

---

<sup>11</sup>Lioret, 2005.



## 6 Ink and paper

Similarly, Chu Hsiang-hsien liked to start a painting by randomly putting ink on silk; then he was fading it halfway and was finding the basic figures of a landscape.<sup>12</sup>

From compositional perspectives, this type of AI is still a manner to overcome the first stroke issue, when the empty canvas presents an overwhelming equiprobable number of alternatives. To this end, I hope that my compositional paradigm could push forward compositional questions in computer art, and find a place somewhere in this fertile family.

### *Here with ink on paper*

They are several reasons leading to my wish to project back my ideas on paper. Let me begin with perhaps the most subjective one. My creative journey begins on paper. It may sound silly, but it is where emotions and magic originally happen. Despite having the chance to see my regularities as a hyper-compositional object, from some sort of third person viewpoint, I believe that the benefit of this hindsight must return on paper. Even theoretical ideas must profit to their initial medium. If final artworks were not produced on paper, it would be like relegating original drawings to a byproduct of a more *advanced* creative process. Returning to ink and paper is therefore a humble tribute to the matter.

A similar shape retains its measure, but changes of quality depending on the material, tool and hand. It is not like the same text drawn on different papers, because paper is only the support of the text: in a drawing, it is an element of life, it is at the heart. A form without its medium is not a form, and the medium is a form itself.<sup>13</sup>

What Focillon expresses perfectly is that phenomena occurring with a medium, only happen with this medium. For instance, during the dataset digitization, we argued that interpreting a line as a series of contrasts in the middle of a matrix of pixels could not be relevant regarding the artistic gesture having generated it (see Section.2.2). This remark still holds at the reproduction level. Digital screens, made of pixel arrays, do not support the continuity of strokes, born from unique tensions. It also means that digitally printed images are not an option either. More importantly, computer screens are freed from the constraint of fixity, and opens up to immediate dynamism. I think that, as basic and neutral a surface of paper can be, it is actually a central property of final compositions, where choices and scales are definitive. On screens, images are temporary, modifiable, interchangeable. On

---

<sup>12</sup>From Tang Hou, Yuan dynasty, and reported by Cheng, 1989, p. 63: “De même Chu Hsiang-hsien aimait à commencer un tableau en mettant au hasard de l’encre sur la soie ; ensuite il l’effaçait à moitié et trouvait les figures de base d’un paysage.”

<sup>13</sup>Focillon, 1934, p. 19: “La même forme conserve sa mesure, mais change de qualité selon la matière, l’outil et la main. Elle n’est pas le même texte tiré sur des papiers différents, car le papier n’est que le support du texte : dans un dessin, il est élément de vie, il est au cœur. Une forme sans son support n’est pas forme, et le support est forme lui-même.”

paper, it is done, once for all. The idea is fixed in the matter. A risk is taken. It is a manner to stop frenzy image flux, and to slow down the consummation.

In the same vein, with the risk of appearing a bit conservative, I would like to re-invest the idea of a long-lasting art. Computer becomes obsolescent in no time. For instance, we already have difficulties to display early video art and some interactive installations. I prefer rudimentary simple and free of maintenance final artworks. It may sound contradictory with the whole technological apparatus set up for this PhD project, and also against my fascination for coding, but fundamentally, I do not believe in computer permanence. If I could (and deserve to) engrave compositions in stones, so that everybody could rub a copy on paper centuries later, like some unique calligraphies of Chinese masters, it would be the ultimate of this logic. In a way, an art which is too inscribed in contemporary tools and preoccupations, will not be resilient to the fluctuations of time. Malevich writes in his manifesto:

With the most primitive of means [...] the artist creates something which the most ingenious and efficient technology will never be able to create.<sup>14</sup>

Nonetheless, as long as final artworks are materialized with ink and paper, a mechanical pen-plotting is not illegitimate. The first reason of this choice is also subjectively biased. As I said in the Introduction, I do not consider the drawings constituting my dataset as *real artworks*. It is as if they were too intimate to be given/sold, or even directly exhibited. I like the idea of hiding my secret garden, and only sharing transpositions of it. Consequently, the return on paper must be a different creative moment from the dataset elaboration. For instance, I cannot draw myself or intervene directly in the drawing process. My *hand* must only appear through the hyper-compositional object, and I do not want to fake the expressiveness captured by the model. That is why, returning to ink and paper, became inseparable from mechanical reproductions with a pen-plotter.

This position is perhaps partially contradicting a previous idea, set up for an exhibition which happened the first year of this PhD, and thus before the completion of the compositional models (see Appendix.A.2). I wanted to initiate a creative loop exploring alternatives of a chosen composition. The initial seed was mechanically plotted several times on paper, and I continued/alterd these copies with the same original pen. Digitized again, this new family of drawings were rearranged and plotted back on paper. In Fig.A.6, this creative loop has been executed twice. So, even if I intervened in the process by drawing myself, it was not directly the case of the final artwork.

The second important reason to use a pen-plotter is the inherent multitude associated with the hyper-composition concept. Any attempt to extract visual propositions from the models is confronted to continuity, diversity, and thus infinity. In addition, captured graphical element dynamics are located in transitions, in

---

<sup>14</sup>Malevich, 1927/2003, p. 78.

## 6 Ink and paper

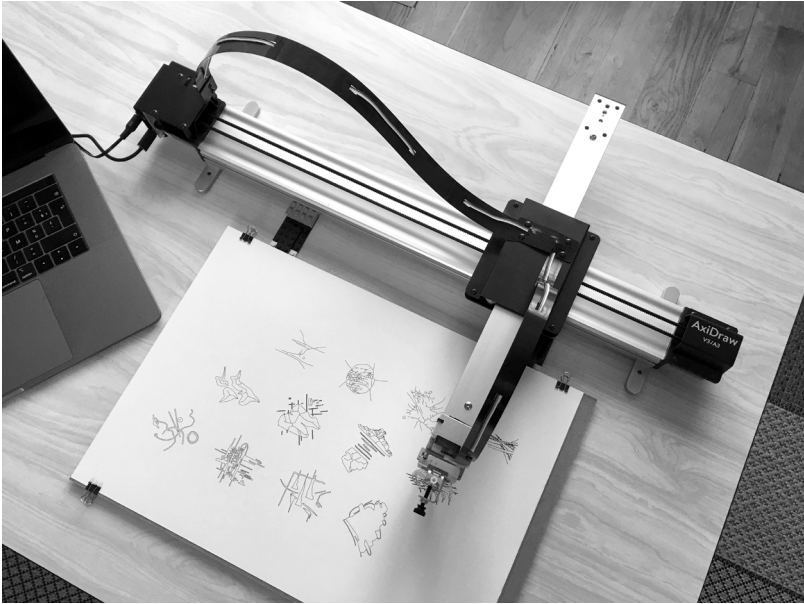


Figure 6.1: Pen-plotter.

interpolations and series. Only a machine can have such intensive drawing ability. A pen-plotter is specifically designed to tirelessly fulfill its duty. I specifically used an *Axidraw* pen-plotter<sup>15</sup> (see Fig.6.1). It is a popular choice among *plotter* artists for its precision and robustness. The unique ballet of the drawing head on the canvas and associated characteristic sound modulations make each plotting of artworks a mix of anxiety and fascination.

However, just as figuration complexifies the reception of compositions, machines also blur the plastic proposal and the focus on compositional aspects. During my first exhibition (see Appendix.A.2), I noticed that the presence of the pen-plotter in the exhibition space, was provoking a certain confusion and invisibilization of the personal creative process. If AI is suggested during the presentation of the work, in people's view the moving machine immediately embodies the creative mind and hand. It is related to the question of AI authorship evoked earlier, but it was often reformulated as a social issue: "Will AI replace artists? Replace my job?" Therefore, *autonomous* machines seem to grab too much attention, and it is difficult to desacralize this fascination. This observation really made me wonder about the necessity to show the pen plotter besides artworks. It is an efficient manner to question the role of the machine and/or to highlight a creative approach in the making, but the central compositional question is not located in the machine, nor in the process itself. Despite a theoretical abstraction of what is composition with intermediary computational tools, I believe that composition only visually expand in the final object. This interesting issue is still unresolved, and for now, I am too afraid of the possible spectators' misconceptions. In a way, it is coherent with my attachment to the basic materiality of artworks, as it eliminates part of viewers' technical questionings.

## 6.2 Stroking the line

On the artist side, using a pen plotter is however not straightforward. Depending on the objectives, technical issues arise, and particularly about surfaces, i.e. thick lines. For instance, to reproduce original drawings, it is required to fill outlines and compensate for stroke widths. Concerning generated compositions, output lines from the model are theoretically non-dimensional along their width. Stroking lines is thus a first challenge to tackle, as well as a first step towards the dynamic of graphical elements.

### *Scaling original compositions*

For my first exhibition, I wanted to reproduce a few hundred drawings sampled from my dataset (see *Accumulation* in Fig.A.3). The exploration of a family of

<sup>15</sup>Please find more details at: <https://axidraw.com>

## 6 Ink and paper

compositions, derived from the same original instance, also requested to be able to plot back on paper any scanned/vectorized drawings (see Fig.A.6). However, both conditions were reproductions of units at a different scale. In Section.2.2, we have seen that the vectorization procedure outputs an outlined and a skeletonized version of each composition. In Fig.6.2, we explore these two options. We first plot the outline version at three scales ( $\times 2$ ,  $\times 1$ ,  $\times 0.5$ ) with a roller pen of thickness 0.2mm. Drawings are displayed in actual size in Fig.6.2a-c. They are then digitally filled and *rescaled* to  $\times 1$  in Fig.6.2f-h. Fig.6.2d is the skeleton version plotted at  $\times 0.5$  and digitally *rescaled* to  $\times 1$  in Fig.6.2i. Finally, Fig.6.2e is the *original* vectorial drawing, serving as a reference. Despite a very precise roller pen, the ball center exactly passes on *true* boundaries, and thus broaden graphical elements of 0.1mm. It seems negligible, but it changes perceptual masses of elements, visual distances, and the overall compositional feeling (see particularly Fig.6.2h). At double scale, the result is acceptable, but still noticeably thicker than the reference (Fig.6.2e,f). At a smaller scale, plotting the skeleton appears to be a better alternative (Fig.6.2d,i). This trick has been used in *Accumulation* (Fig.A.3), but a proper correction procedure will be proposed in the next subsection.

Changing the scale of a composition thus raises a fundamental compositional question. In both artworks evoked above, the change of scale did not exceed a few folds, which is reasonable if corrected. However, this is not the case of a third artwork called *Individuality*, where a composition was plotted alone on a A3 canvas (see Fig.A.4). In this context, each stroke width must be redefined, and the overall work requires additional details to be effective. In Matisse's words:

The artist who wants to transfer a composition from a canvas to a larger one must, in order to preserve its expression, conceive it again, modify it in its appearances, and not simply lay it down a grid.<sup>16</sup>

### Offsetting lines and filling surfaces

In order to correct for the extra thickness provoked by the plotting operation, it is necessary to shrink outlines beforehand. A similar issue happens in the field of computer-aided design and manufacturing, where the correcting operation is known as line offsetting. Such procedure seems trivial; a point on a line just has to be translated along the normal vector to this line. In Fig.6.3a, the dotted line is the offset version of the solid line, thanks to the gray normal vectors. However, this naive algorithm produces a spurious loop/area, highlighted in gray. More generally, it generates several geometrical issues in concave regions. In Fig.6.3b, we show real artifacts happening while offsetting a skeleton line on both sides.

---

<sup>16</sup>"Lay it down a grid" is a traditional and manual scaling procedure. Matisse, 2014, p. 43: "L'artiste qui veut reporter une composition d'une toile vers un toile plus grande doit, pour en conserver l'expression, la concevoir à nouveau, la modifier dans ses apparences, et non pas simplement la mettre au carreau."

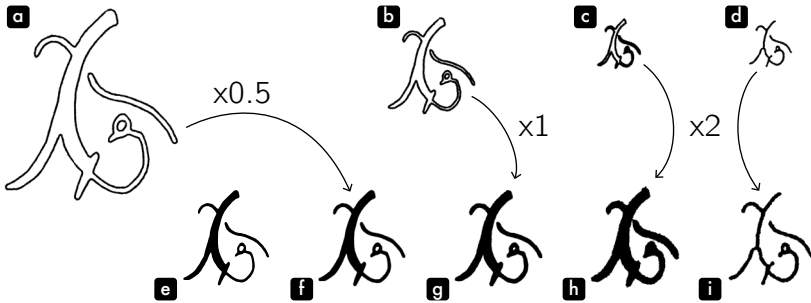


Figure 6.2: Reproducing a drawing at different scales. In panels **a**, **b**, **c**, the outline of a drawing is plotted at three scales ( $\times 2$ ,  $\times 1$ ,  $\times 0.5$ ) with a roller pen of thickness 0.2mm (drawings are displayed in actual size). In panels **f**, **g**, **h**, these plots are digitally filled and *rescaled* to  $\times 1$ . Panel **d** is the skeleton version of the same drawing plotted at  $\times 0.5$ , and digitally *rescaled* to  $\times 1$  in panel **i**. Panel **e** is the *original* vectorial drawing, serving as a reference.

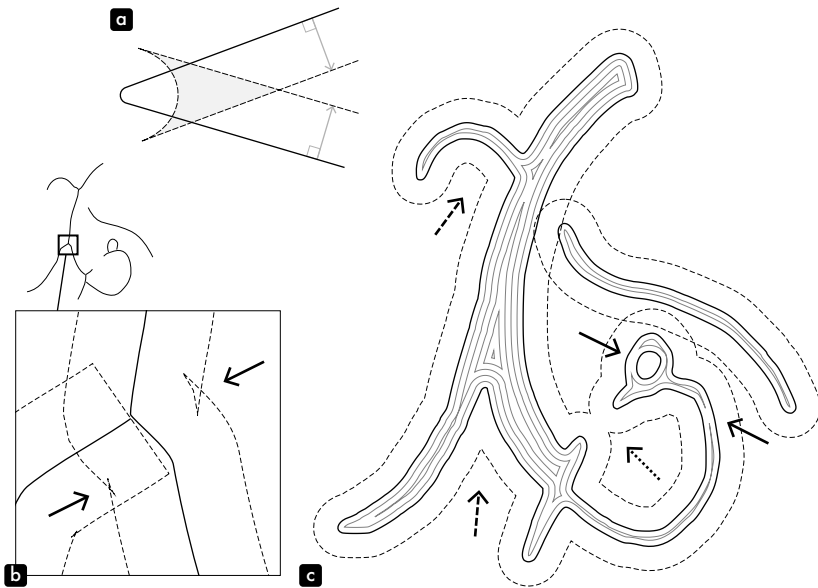


Figure 6.3: Offsetting lines and filling surfaces. Panel **a** shows the issue of a naive offsetting procedure in concave regions. The dotted line is the offset version of the solid line, gray arrows are normal vectors to the line, and the gray area is the resulting spurious artifact. Panel **b** shows real glitches happening while offsetting a skeleton line on both sides. Panel **c** finally displays a successful offsetting operated with the chosen open source library (dotted lines). Gray inner lines are a recursive negative line offsetting, designed to fill closed surfaces.

## 6 Ink and paper

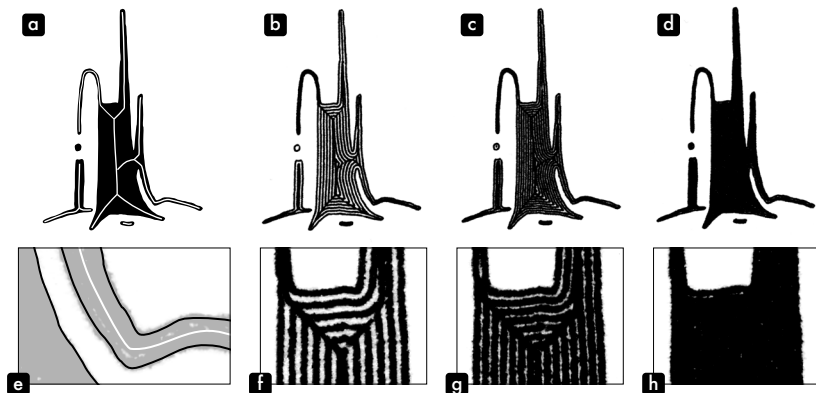


Figure 6.4: Pen thickness correction and different filling intervals. Panel **a** is the *original* closed surface and its inner skeleton (white lines). Panel **b**, **c**, **d** are plotted results in actual size for three filling intervals (0.4mm, 0.3mm, 0.2mm), and panels **f**, **g**, **h** show corresponding close-ups. Panel **e** finally displays the comparison between the *original* outline (black) and the actual plotted version (grayed out).

As a result, the industry drove more advanced scientific developments<sup>17</sup>, and we chose to use the open source library called Clipper<sup>18</sup>. Dotted lines in Fig.6.3c, show the resulting successful positive offsetting of concave areas (largely dotted arrows). In addition, we remark that separated ends of the lower right curve correctly becomes connected (thinly dotted arrow). Concerning negative offsetting (first gray inner line), particularly when trying to correct for the plotting thickening issue, it is advised to plot the skeleton centerline as holes may appear in thin regions (see solid arrows). Consequently, filling a surface is simply applying the negative offsetting procedure recursively (all gray inner lines).

In Fig.6.4, we experiment the whole procedure with the pen-plotter, i.e. outline shrinking of half a pen thickness (0.1mm), followed by surface filling. Fig.6.4a shows the *original* closed surface and its inner skeleton (white line). Fig.6.4b-d are plotted results in actual size for three filling intervals (0.4mm, 0.3mm, 0.2mm). From Fig.6.4f-h close-ups, we notice that only the 0.2mm interval, corresponding to the pen thickness, gives the expected result. It also confirms that the skeleton line is important in thin regions. Finally, the close-up from Fig.6.4e verifies the thickness correction by comparing the *original* outline in black with the grayed out actual plotted version. The result is really satisfactory.

An important technical detail of the Clipper library is that it only processes polylines,

<sup>17</sup>X. Chen and McMains, 2005.

<sup>18</sup>We actually use pycclipper, a Python binding of the C version of Clipper. This library does polyline offsetting and more generally boolean operations on polylines. It is mainly an implementation of Vatti, 1992. More information on the project can be found at: <http://www.angusj.com/clipper2>

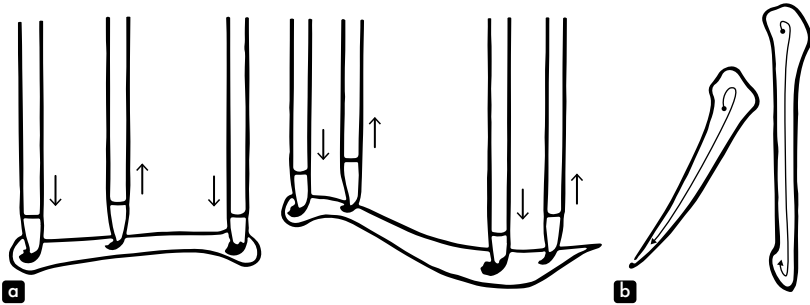


Figure 6.5: Calligraphic strokes<sup>19</sup>. Panel **a** is an illustration of brush vertical movements inducing varying stroke thicknesses. Panel **b** shows the inner trajectories required to render specific stroke styles: small loops at extremities are particularly important to achieve clean stroke ends.

i.e. sequences of linear line segments. This is actually the case of the pen-plotter too. There is therefore the necessity to transform all parametric curves, such as cubic Béziers, to polylines. To do so, we evaluate each curve components at 100 points (see the curve in Fig.6.6a for an illustration with 5 points per component). A second aspect to take into account is concerning compound polylines, i.e. forms with inner holes in larger surfaces. For instance, an expansion for the outer boundary corresponds to a shrinkage of the inner ones. Finally, a particularity of our implementation is to operate all computations at the global scale, by applying all pending transformation matrices on the polyline beforehand, so that the offsetting width is expressed in global units. Of course, inverse transformations are applied when values are written back in the final svg file structure.

### Stroke profiles

In Section.2.2, we have already evoked the idea of a third dimension of lines. This additional degree of freedom and associated perceptual phenomena, originate in the practice of calligraphy with brushes and fountain pens. In Fig.6.5a, we see how varying vertical pressures affect stroke expressiveness along the path. Stroke profiles are thus an inherent aspect of line dynamics. However, it does not really matter that this third dimension is physically produced with a brush and real body movements. Any variation of thickness along the path produces potential tensions. For Kandinsky:

This means of expression creates a certain vibration of the elements in the case of an acute dryness of the main elements of a composition. It brings a

<sup>19</sup>Illustrations adapted from Yee, 1974, pp. 145, 147



## 6 Ink and paper

softening of the rigid atmosphere of the whole, but used to an exaggerated extent, it leads to an almost repulsive preciousness.<sup>20</sup>

The idea is therefore not to emphasize too fake dynamics, but at least to recover a natural stroking of lines, since we had to discard this information during the dataset formatting (see Subsection.2.2.Surfaces to lines). We define a stroke profile as a function producing a value in the range  $[0, 1]$  for each point of an input polyline (0: no thickness, 1: maximum thickness). We will explain in the next subsection how to use profile functions for visualization and plotting, but first, we need to expose a prerequisite on polylines. Fig.6.6a shows a curve and a straight line. The horizontal line is already a *polyline*, but as explained earlier, Bézier curves are usually transposed by evaluating a fixed number of point per component (here 5), uniformly spaced in the parameter space  $u$  (see Eq.2.4). Applying a profile function on these lines, a thickness could only be added at each point. So, despite a fancy profile, the straight line could only become a trapezoid, and the curve would suffer from an uneven stroking definition. As a result, we should rather uniformly sample polylines in length space along their path (see Fig.6.6b). In practice, we define a resolution value in physical units, specifying the distance between polylines points, empirically set to 0.01mm.

Let us now concentrate on the design of the profiles. Among several attempts, on a trial and error basis, I decided to select the two most successful ones. I particularly discarded a lead incorporating randomness in stroke alterations, because it was a type of stochasticity, I have expressed concerns about in the previous section. Stroke profile I is first designed to be simple with a small asymmetry. In Fig.6.6c, we notice a slightly larger onset breaking the central regularity. Extremities are also nicely rounded. This profile function is a simple mix of sine/cosine functions without free parameters, nor geometry specific adaptation. The precise computation can be found in Algorithm.6.1. In Fig.6.7, we apply this stroke profile on generated compositions, first shown in Fig.4.16. I think, that the stroking effect brings subtle variations and an effective presence to lines, while preserving a good homogeneity. That is why, this neutral approach has been used in stimuli generation for the perceptual experiment reported in Section.5.3 (see particularly Fig.5.18).

Stroke profile II explores a more expressive proposal. I remarked that straight lines are usually executed quicker than curvy ones. It is like a physical tension had to be inscribed in curves to echo a directional uncertainty, a mental creative tension. Therefore, I had the idea to use the line curvature to positively drives the stroke thickness. Some results are shown in Fig.6.6d and on generated compositions in Fig.6.8. I think that the result is bringing dynamics and depth to each composition. Despite its genericity, the rule seems well-defined in many non-trivial situations. So,

---

<sup>20</sup>Kandinsky, 1926/1991, p. 109: "Ce moyen d'expression crée une certaine vibration des éléments dans le cas d'une sécheresse flagrante des éléments principaux d'une composition. Il apporte un assouplissement de l'atmosphère rigide de l'ensemble, mais, employé exagérément, il mène à une préciosité presque rebutante."

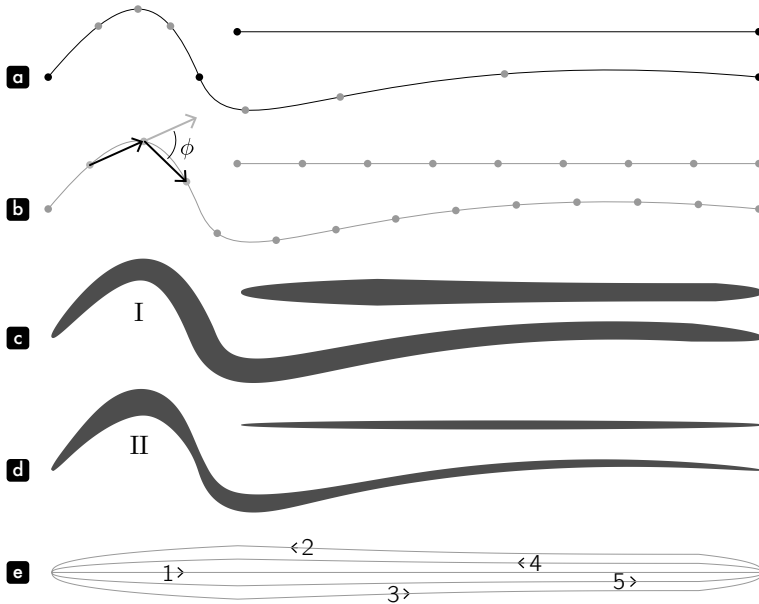


Figure 6.6: Stroke profiles. Panel **a**, show how Bézier curves are evaluated as polyline with a fixed number of point per component; here 5, uniformly spaced in the parameter space  $u$  (see Eq.2.4). Panel **b** illustrates how polylines should rather be uniformly sampled along their path with a fixed resolution. Black arrows are tangent to the curve and  $\phi$  is the angle in between, required to compute the stroke profile II (see Algorithm.6.2). Panels **c** and **d** display stroke profiles I and II applied on a straight line and a curve. Panels **e** finally shows the plotting order of outlines and interlines to render stroke profiles with a roller pen.

---

### Algorithm 6.1: Stroke profile I

---

**Function strokeProfileI(polyline):**  
 $x \leftarrow \text{linearSpace}(0, 1, \text{len}(\text{polyline}))$   
**return**  $\min(\sin(x \times \pi)^{0.4}, 0.6 + 0.4 \cos(x \times \frac{\pi}{2})^4)$

---



---

### Algorithm 6.2: Stroke profile II

---

**Function strokeProfileII(polyline, kernelSize, amplification, gamma, ratio):**  
 $\text{polyline} \leftarrow \text{gaussianFilter}(\text{polyline}, \text{kernelSize})$   $\triangleleft$  nearest-value boundaries  
 $\text{tangents} \leftarrow$  local difference of  $\text{polyline}$  points  
 $\phi \leftarrow$  angle between local  $\text{tangents}$   $\triangleleft$  see Fig.6.6b  
 $y \leftarrow (\text{clamp}(\|\phi\| \times \text{amplification}, 0, 1))^\gamma$   
 $y \leftarrow \text{gaussianFilter}(y, 0.1 \times \text{kernelSize})$   $\triangleleft$  boundaries extended with 0  
 $y \leftarrow \text{concatenate}([0], y, [0])$   
 $x \leftarrow \text{linearSpace}(0, 1, \text{len}(\text{polyline}))$   
**return**  $1 - ((1 - \text{ratio} \times \sin(x \times \pi)^{0.4}) \times (1 - y))$

---

6 Ink and paper

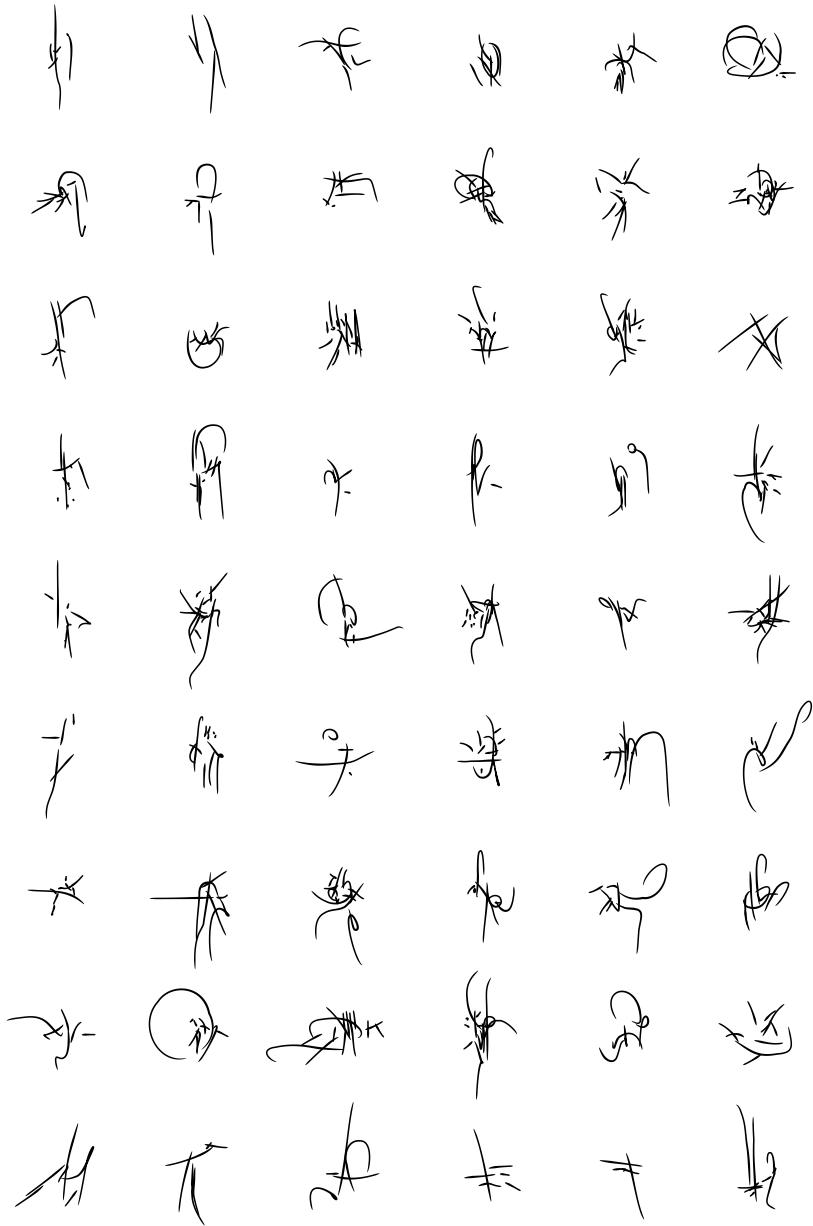


Figure 6.7: Stroke profile I on generated compositions.

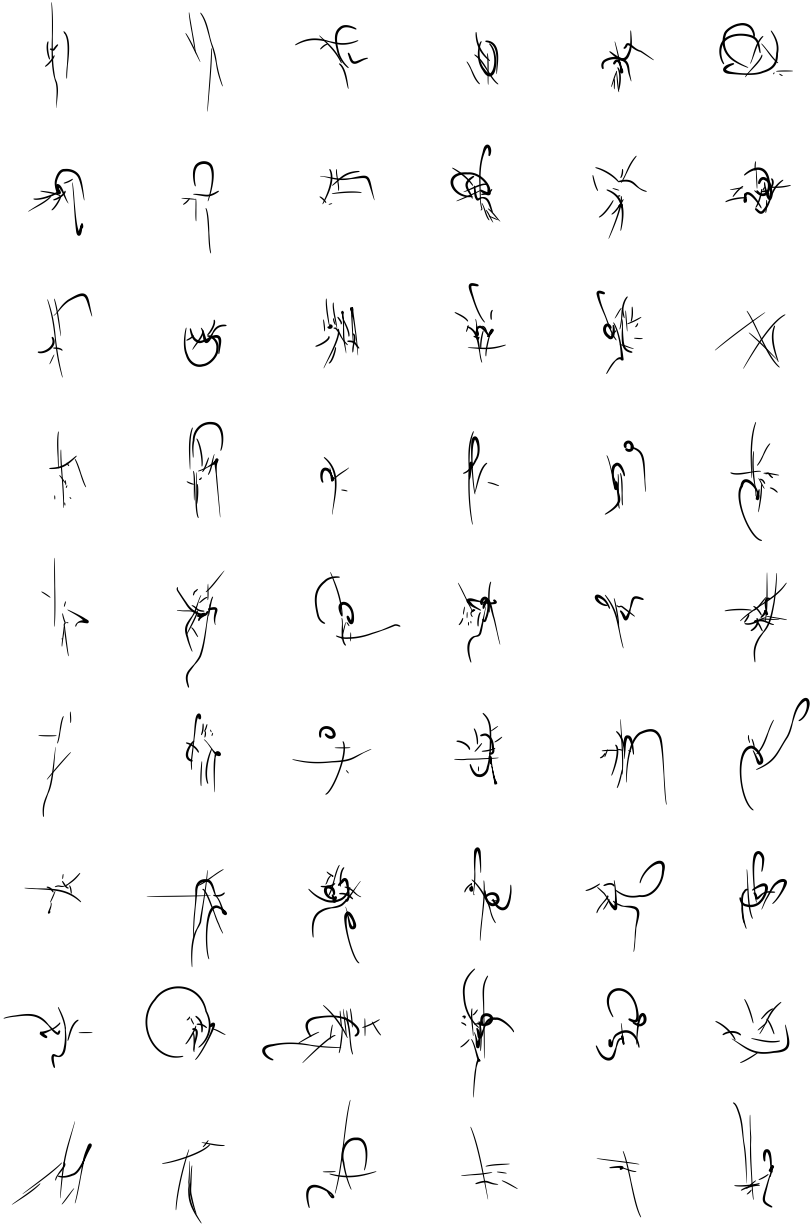


Figure 6.8: Stroke profile II on generated compositions.

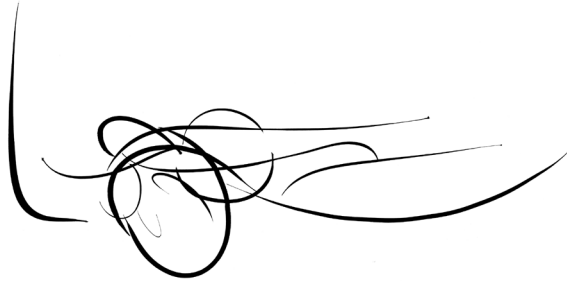


Figure 6.9: Brush plots with the stroke profile II.



Figure 6.10: Constant depth brush plot (3 levels).

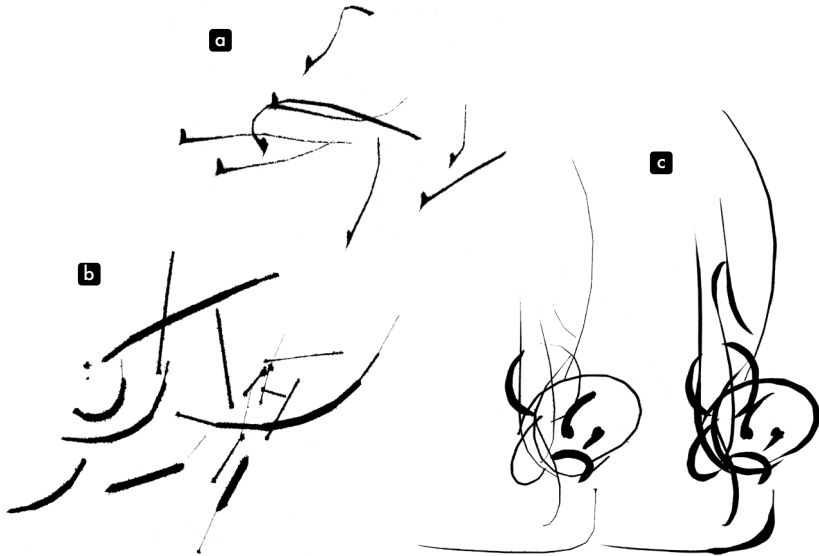


Figure 6.11: Trial and error with brush plots.

it represents a satisfactory replacement to the discarded *true* original information. Nonetheless, implementation and parameterization of this profile function are not easy. Algorithm.6.2 summarizes all required computations.

### Plotting strokes

In order to render stroke profiles on paper, we designed two different methods. A first scenario addresses large scale compositions with brushes, and a second is dedicated to small scale drawings with roller pens. Ideally, using a brush should be the preferred method for all scales because of its natural expressiveness, but there

**Algorithm 6.3:** Line stroking

---

**In:** · *polyline*, input polyline at physical scale, e.g. mm  
 · *stkWidth*, maximum stroke width, i.e. where *profile* = 1  
 · *interline*, interline width to fill the stroke, if *interline* is negative the original centerline is discarded, and the algorithm outputs a polygon, (None)  
 · *profile*, stroke profile function (**strokeProfileI**)  
 · *params*, parameters of *profile* (for **strokeProfileII**: 100, 100, 0.5, 0.3)  
 · *resolution*, resolution to uniformly resample *polyline*, (0.01)  
 · *rdpTolerance*, path simplification tolerance for the RDP algorithm, (0.001) [see description in Subsection.2.2.Parametric curves]

**Out:** · *polylines*, list of polylines in an efficient order and orientation for plotting

---

*polyline* ← uniform resampling along the path of *polyline* at *resolution*

▽ compute outline polylines

*tangents* ← compute normalized local difference of *polyline* points

*normals* ← rotate *tangents* by 90°

*polyline*<sub>1</sub> ← *polyline* + *stkWidth* × *normals*

*polyline*<sub>2</sub> ← *polyline* − *stkWidth* × *normals*

▽ apply stroke profile

*weight* ← *profile*(*polyline*, *params*)

*polyline*<sub>1</sub> ← *weight* × *polyline*<sub>1</sub> + (1 − *weight*) × *polyline*

*polyline*<sub>2</sub> ← *weight* × *polyline*<sub>2</sub> + (1 − *weight*) × *polyline*

▽ compute interline polylines

*polylines* ← [*polyline*, *polyline*<sub>1</sub>, *polyline*<sub>2</sub>]

**if** *interline* is not None **and** 0 < *interline* < *stkWidth* **then**

*n* ← ⌊*stkWidth*/*interline*⌋

**for** each *i* in range(1, *n*) **do**

        [ *polylines* +←  $\frac{i}{n} \times \text{polyline}_1 + (1 - \frac{i}{n}) \times \text{polyline}$

        [ *polylines* +←  $\frac{i}{n} \times \text{polyline}_2 + (1 - \frac{i}{n}) \times \text{polyline}$

**else if** *interline* is not None **and** *interline* < 0 **then**

    [ *polylines* ← [*polyline*<sub>1</sub>, *polyline*<sub>2</sub>]

▽ simplify and orient polylines

**for** each *i* in range(0, len(*polylines*)) **do**

    [ *polylines*[*i*] ← RDP simplification of *polylines*[*i*] with *rdpTolerance*

**if** *i* is odd **then**

        [ *polylines*[*i*] ← reverse points order of *polylines*[*i*]

**return** *polylines*

---

are several technical issues. First, the pen-plotter does not have the default ability to change the pen *height* (depth) over the paper in the middle of a path. Before tweaking the official API of the plotter, we tried to limit the depth to few constant levels and a tilted brush. In Fig.6.10, we see that the slanted brush produces some thickness variations, but that was not fully satisfactory. In addition, the height precision of the servomotor is not accurate enough with a tilted brush to produce many different thicknesses. A vertical positioning of the brush is thus required, but it generates spurious stains on stroke onsets (see Fig.6.11a). This is why Chinese calligraphers make little loops at stroke extremities (see Fig.6.5b). This procedure

is not easy to implement, so we preferred to operate the initial contact with the paper in movement. This implied to modify the core API of the plotter to enable the application of a stroke profile on the fly. In Fig.6.11b,c, we show that the complete calibration of the system was not easy. Nonetheless, Fig.6.9 presents successful results with the stroke profile II.

This first scenario is optimal for large compositions, since rendering thin strokes is challenging with a brush. The line quality produced with the very top of the brush is too inconsistent. Therefore, for small compositions, or when a high precision is required, we developed a second method to mimic stroke profiles with roller pens. The approach is actually close to the previously described recursive surface filling method. The complete procedure is reported in Algorithm.6.3. Basically, we offset a polyline by the intended stroke width on both sides. Then, we linearly interpolate each side between the new boundary and the original polyline, weighted by the stroke profile. The subtlety of this approach is to require an identical number of points between interpolated polylines. The Clipper library is not intended to do so, and especially does not, in order to produce accurate results. Thus, we use the naive approach described earlier, despite its known artifacts in concave regions (remember Fig.6.3a,b). In practice, these glitches happen inside strokes and are not visible. Fig.6.7 and Fig.6.8 actually used this method without noticeable issue. Nonetheless, it could be improved in future works.

### 6.3 Diversity, continuity, and dynamics

This last section introduces the main creation principles, that will contribute to the creation of the final artworks of this project. The goal is to find visual ideas to convey key concepts of the hyper-compositional object, such as diversity, continuity of hidden dimensions, and graphical element dynamics. However, the didactic characteristics of these productions must not supersede elementary compositional, and basically pleasing, attributes.

#### *Representing diversity and continuity*

Both diversity and continuity concepts are related to infinity. Diversity stands for the infinite expressive differences in the generative space, while continuity implies coherent and subtle compositional transitions in the infinitesimal neighborhood of a chosen sample. We could rephrase both ideas as large and small scale infinities. Nonetheless, those infinities are not materially feasible; only incomplete views of the hyper-compositional object are realizable. So, the reasonable accumulation of samples, original or generated, is our proxy for unlimitedness. To this end, the representational question is mostly confined to the choice of the appropriate disposition of composition ensembles. Individual positions on the canvas may be



## 6 Ink and paper

derived from their *true* coordinates in the compositional space, or constrained on any arbitrary grid.

Using original coordinates is an interesting idea, but we must obtain 2-d mappings of original 16-d coordinates. For instance, PCA decomposition cannot be employed since the model enforces a unit variance in each dimension. We can arbitrarily select two dimensions, and project every sample on the chosen plane, but the compositional logic of the chosen dimensions is unlikely to be obvious. In Sub-section.5.1.Threshold estimation, we have seen that the perception of compositional transitions is noticeable for angular distances from 2° to 8°. Then, resulting artworks of randomly picked compositions would not appear more organized than a pure random arrangement on the canvas. More complex unwrapping procedures from 16-d to 2-d, such as MDS and Isomap exist (see Fig.5.6a,b), but the general compositional logic of sub-elements would still seem arbitrary<sup>21</sup>.

The disposition of compositions along regular arrays is then an obvious, but efficient way to reveal the model diversity. Besides scientific papers on generative models, grids have actually been used for a long time in generative art. It is somehow the default choice to show random variations of any generative procedure. Our type of diversity is different, i.e. occurrences of a learned hyper-compositional object, but it can be investigated with similar principles. We thus explore two types of grids; regular and honeycomb with horizontal or vertical orientations (see Fig.6.12a-c). They are the most neutral forms of organization, as no sub-element is favored. Honeycomb structure is used in nature and material design for its equal distribution of tensions, so by extension, it offers an interesting distributed attention on every unit. We employed the regular grid to display raw generated compositions in Fig.4.16 and Fig.4.18, as well as with stroke profile II in Fig.6.8. On the other hand, *Accumulation* artwork in Fig.A.3 adopts the vertical honeycomb grid. The enforced equidistance of neighbors makes the overall structure more pleasing, while not superseding individual compositions. It quietly induces the feeling of a viewpoint on the life of microorganisms. Nonetheless, the choice of the grid remains highly arbitrary. It mostly depends on the nature of the content, and the number of elements to represent.

Unlike diversity, continuity emerge from an accumulation, where the local connectivity of samples respect their relative distance in the latent space. In order to reveal visual regularities, the chosen arrangement on the canvas must be barely homothetic to the sampling grid. For instance, the local diversity of generated compositions around a *target* in Fig.4.17 does not produce the same effect of continuity as with two-dimensional interpolations, like in Fig.4.29 or Fig.5.1. In both cases, families of samples are presented, but the preservation of the original

---

<sup>21</sup>The only exception is maybe an artwork made for the VSAC exhibition, where I used a very naive CNN (a model relying on pixel maps) and PCA to obtain 2-d coordinates (see Fig.A.7). The result is satisfying, but does not reveal a cartography more elaborated than an obvious distribution of drawings in function of the stroke thickness and the horizontal/vertical dominance. In addition, to prevent a confusing superposition, overlapping compositions had to be randomly pruned.

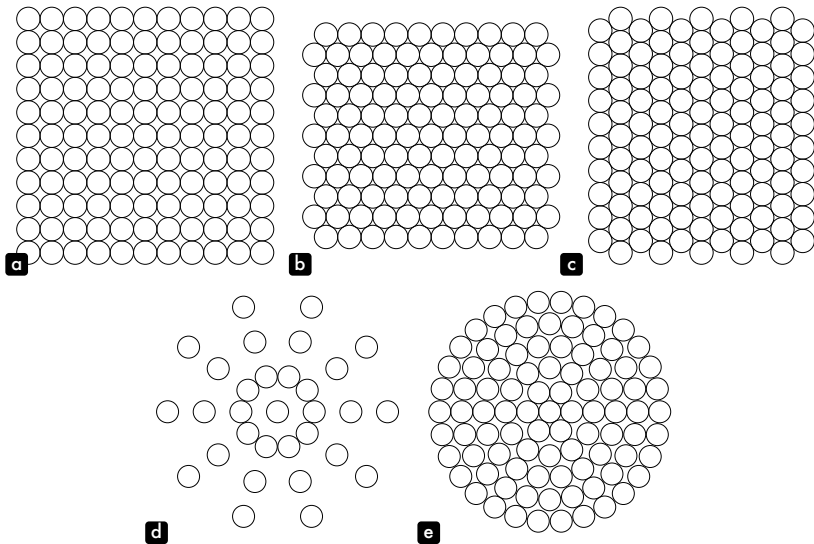


Figure 6.12: Grids. Panel **a**, regular. Panel **b** and **c**, honeycomb with horizontal and vertical orientations respectively. Panel **d**, circular. Panel **e**, circular homogeneous.

sampling topography of the latter transforms the perceptual visual effect from diversity to continuity. This phenomenon is actually already effective with simple line interpolations (see Fig.5.18). If the sampling is circular, then a circle is an appropriate alternative (see Fig.5.17). Following this concept, we explore in Fig.6.13 the idea of a *composition wheel*. It is inspired by the traditional color wheel, representing the continuity of color hues around a circle, and converging to gray in the center. This design smartly highlights complementary combinations of color. In the compositional case, such grid shows transitions in different directions around a well identified *target* composition. All radial lines do not show the same number of elements, but we prefer this resulting spatial homogeneity compared to a vanilla circular grid (see Fig.6.12d,e). Finally, we could imagine more elaborated sampling grids, such as intersections of orthogonal 3-d planes projected into 2-d, but we believe it would only introduce confusion. The complexification of supporting grids automatically diminishes individual compositional effects. An appropriate grid is therefore necessarily discreet and simple.

### Witnessing dynamics

In Subsection.1.1.Cinematic of forms, we have described how the concept of weight given to graphical elements with a fixed delimitation was constructing a *cinematic*

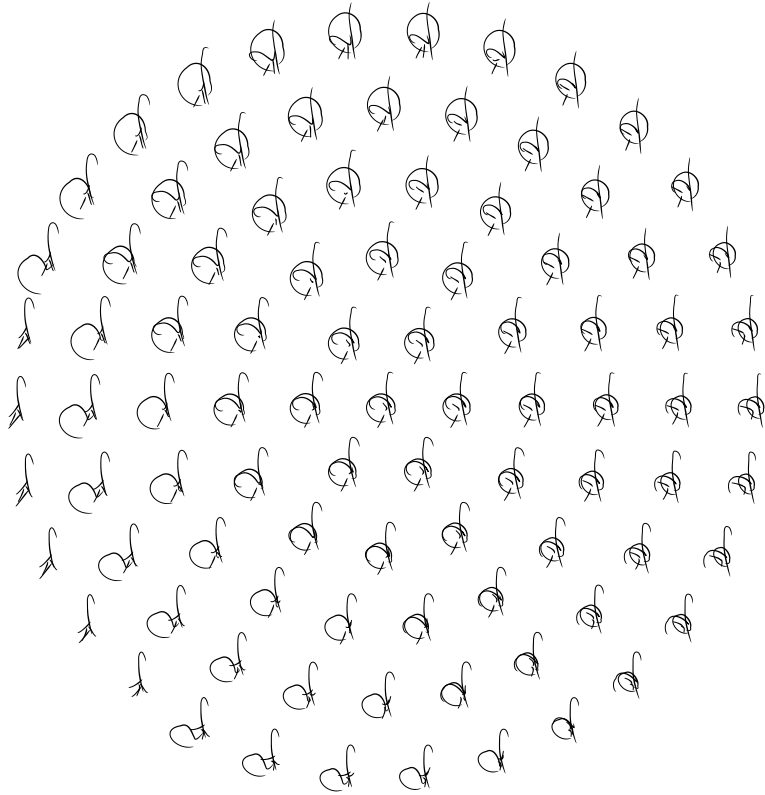


Figure 6.13: Generated compositions over a homogeneous circular grid. Concentric rings correspond to successive circular slices of the mode hypersphere, which respect the necessary number of sample per ring and the chosen angular distance between samples ( $2^\circ$ ).

of forms. Then, how a *dynamic* of forms would singularize from this idea? The essential difference is that the concept of dynamic allows compositions to go beyond positional and scaling arrangements of absolute forms. It enables the modification of the graphical elements themselves, as a function of tensions resulting from mutual interactions. The purpose of making viewers witnessing dynamics is thus to let them have a glimpse on the pictorial formation, on the morphogenetic aspect of the composition producing *de*-formations.

An apparent way of rendering dynamics is thus to directly transpose morphological changes along the time, as an animation. For instance, in Fig.5.22 and Fig.5.23, we inverted the perceptual scale, that was experimentally measured with participants, and proposed perceptually smooth interpolations. But, the necessary resampling can actually be operated at any granularity, until obtaining animations with a sufficient temporal definition to give the illusion of continuity<sup>22</sup>. Resulting animations seem very organic and visually satisfying. They provide an efficient tool to directly study compositional forces. In future works, I could even try to build real-time and interactive applications exploring the latent space. However, it would remain a personal creative tool rather than artworks addressed to spectators. In addition, even if seeing real dynamics of shapes is enjoyable, it hides the essential. Dynamical aspects must remain potential to trigger tensions. With real movements, tensions are satisfied, and thus resolved. In other words, compositional dynamics primarily live on paper.

Italian Futurism directly tackled the representation of movements by integrating on the same canvas the past, present, and future of the depicted subjects. In a way, they extended the idea of Cubism trying to multiply the points of view on the same object. In spite of searching for objectivity through the simultaneity of viewing angles, Futurists wanted to express the force of movements, and their fascination for speed and acceleration. However, notions of time and viewpoint are artificial in the context of the compositional space. If we explore a type of simultaneity on the canvas with our model, it must be by highlighting compositional uncertainties, and accumulating possible alternatives of a neighborhood in the generative space, or exploring transitions from one another.

In Fig.6.14, we superimpose samples from a locality around a chosen position in the compositional space. Little moves along different hidden dimensions reveal most uncertain graphical elements. Consequently, using thin lines, results produce a sketching effect, as if the model was searching for the right stroke to draw its objective. Some important elements are reinforced, while others seem difficult to position or to give the right curvature. The same principle is applied to the compositional plane model (see Fig.6.15). This time, we show the uncertainty located

---

<sup>22</sup>Technically speaking, I have developed a piece of software generating animated svg files powered by embedded JavaScript, and functioning as an old-school flip book. Individual frames are horizontally shifted in a view box at a chosen frame rate, e.g. 25 i/s. Animations produced from the 6 conditions of the experiment described in Section.5.3 are available on my website on the page dedicated to my poster at VSS2022: <https://plelievre.com/projects/vss-2022#anim>

## 6 Ink and paper

in outputs of the model, produced from a unique location in the compositional space. Resulting drawings are more fuzzy and crisp at the same time. Sampled alternatives integrate more different propositions, for which the model is however individually more confident. It gives a feeling of hurry, like the necessity to rush to capture a visual idea, precariously floating in the mind.

In the previous section, strokes thickness have been postulated to be the result of motor accelerations and decelerations, specifically highlighting curvy regions. This choice was motivated by the practical experience, but it is somehow arbitrary. In addition, the uncertainty principle presented above provokes vibrations of lines with diverse strength. It is basically as if information concerning stroke thicknesses were already contained in the model stochasticity. Indeterminacy about a specific line can therefore be considered as a proxy of inherent dynamics, and induces varying widths along strokes. In Fig.6.16 and Fig.6.17, we explore this idea by creating a surface in between corresponding lines of two local samples in the compositional space. This principle is theoretically close to the former proposition, but the visual result is dramatically more expressive. It highlights uncertainty of individual elements with a neat graphical language. Resulting compositions seem more fluid, as more empty space brings deeper contrast, and bolder gesture impressions. A possible issue arises where boundary lines of strokes cross along the path, and look like twisted ribbons (see the bottom stroke of the lower left composition in Fig.6.16). This graphical feature is not really problematic, and actually evokes stroking effects produced with large fountain pens.

Filling the surface in between two close samples was a manner to clear the fuzzy accumulation of uncertainty into a more identifiable intention. However, under the black surface of these two boundaries, we do not know the inner transition from one another. In Fig.6.18, Fig.6.19 and Fig.6.20, we operate a spherical interpolation in between two local samples, and plot intermediary lines. Transitional samples are supposed constant in latent distances, so compressions and dilations within each graphical element visually add an expressive density to stroke thicknesses. Implied grayscale gradations also provide the illusion of different ink dilutions, and the feeling a delicate superstition of layers. Resulting effects may as well be perceived as depth cues, giving a shallow relief to strokes, or the feeling of soft enfolding surfaces.

Finally, in Fig.6.21 and Fig.6.22, we explore a very different visual principle. First, we select interpolation sequences perceptually corrected for continuity thanks to the MLDS experiment described in Section.5.3. We then sample each path with a very high granularity, and record the trace along horizontal or vertical constant shifts. This way, original compositions are completely disrupted, but a calm and natural volume appears, evoking the passage of vanished forms, where only dynamics remain.

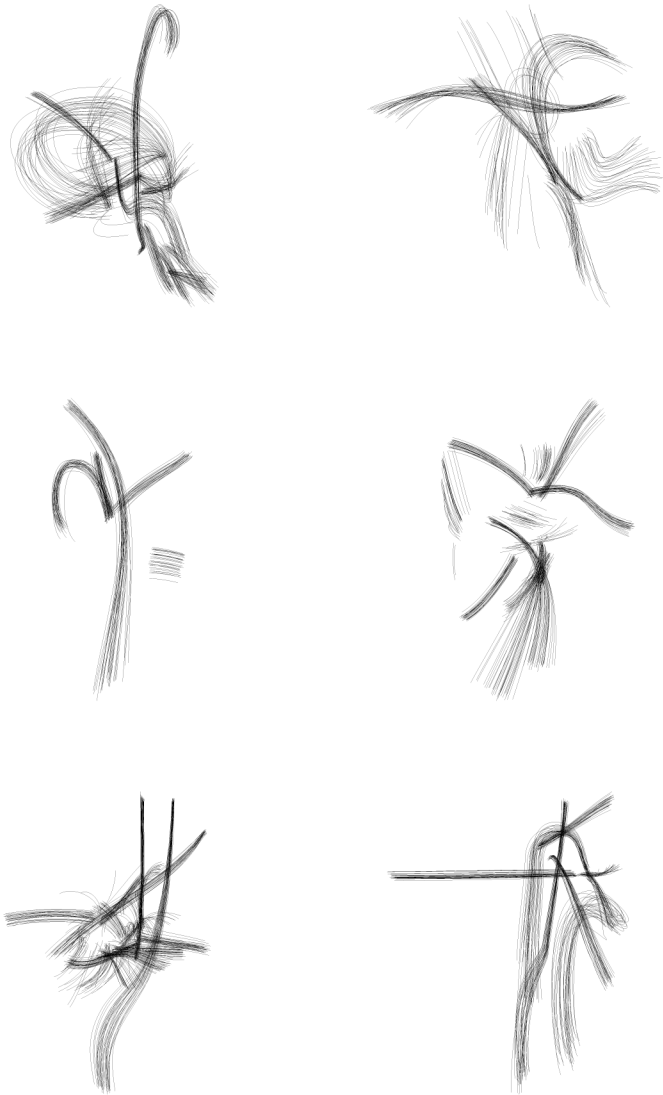


Figure 6.14: Dynamics – Uncertainty – 1.

## 6 Ink and paper

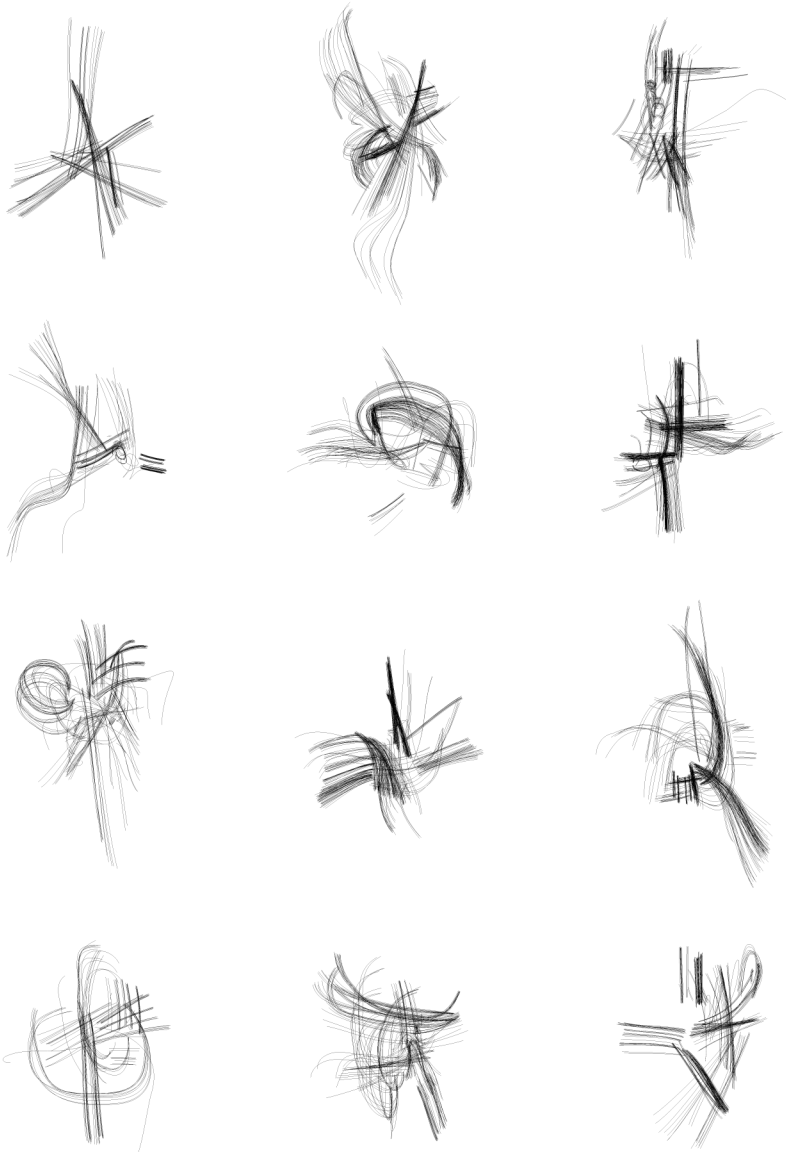


Figure 6.15: Dynamics – Uncertainty – 2 (compositional plane model).



Figure 6.16: Dynamics – Thickness – 1.





Figure 6.17: Dynamics – Thickness – 2.

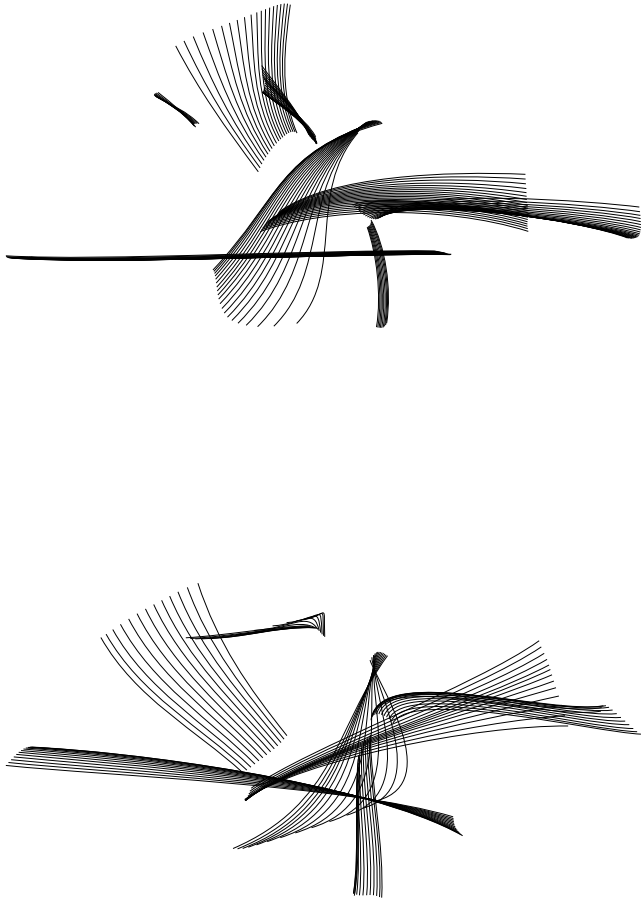


Figure 6.18: Dynamics – Transition – 1. The two propositions explore the same compositional locality.

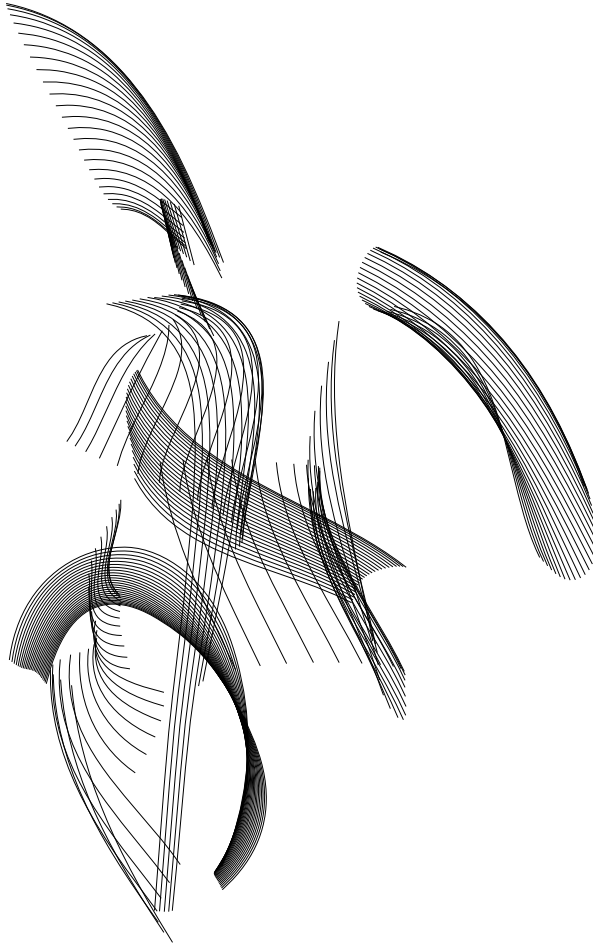


Figure 6.19: Dynamics – Transition – 2.



Figure 6.20: Dynamics – Transition – 3.

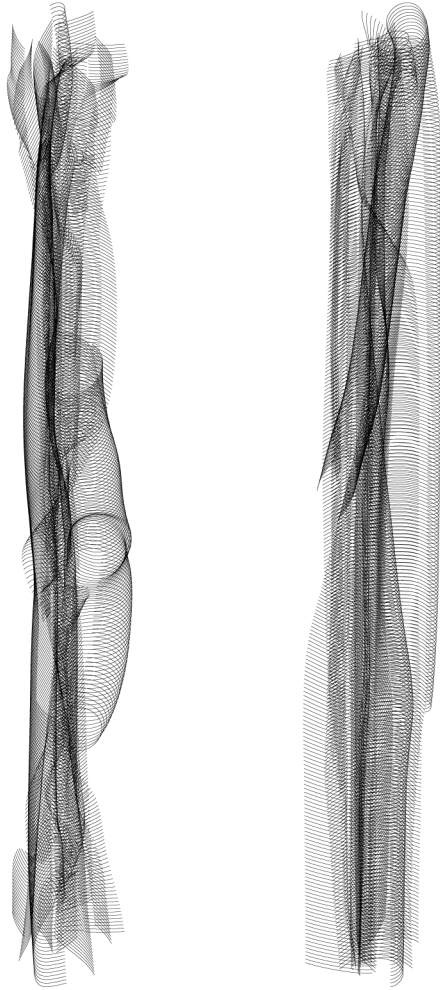


Figure 6.21: Dynamics – Trace – 1.



Figure 6.22: Dynamics – Trace – 2.



## Conclusion

It is somewhat difficult to formulate conclusions for a research agenda that is still at its very beginning. In the preceding chapters, we only had the opportunity to introduce a first iteration of the difficult process involved in laying out each necessary brick of the construction that represents a preliminary result. There were so many unavoidable steps along the path towards securing a functional tool, that those monopolized all my efforts. It could have been expected to contribute more practical findings on the perception of composition, and more advanced discussions on relevant artistic questions, but I am already proud of what I have been able to accomplish so far. The whole framework has been validated, and its utility demonstrated, up to the point of conducting a real experiment with human participants, together with the realization of artworks back on paper. The realization of these ambitious goals has exceeded my initial expectations. In addition, as stated in the Introduction, the projective modeling approach we have favored in this thesis must be constantly reevaluated, ready for the next iteration as soon the previous one has come to an end.

During the course of my PhD program, I naturally engaged in considerations about my legitimacy and competence with relation to the different fields I encountered, and I raised doubts about my ability to reach the end goal. For instance, when my first modeling attempts failed, I did not know whether to attribute this outcome to conceptual issues, technical problems, or to the possibility that my compositions may not carry sufficient regularity. This is possibly the most difficult aspect of working on your own artistic material. In addition, the relatively small size of the dataset required seemingly endless additional layers of complexity, and accentuated the aforementioned uncertainty about the source of my modeling failures. Writing the manuscript also took me longer than expected, but this was an inevitable consequence of recording all implementation details and diverse contributions in order to construct a coherent global picture of my research. However, even though this manuscript is almost 300 pages long, it remains shorter to browse than the associated programming code. This hidden part of the iceberg has probably exceeded 30000 lines. The amount of time spent behind a computer was perhaps the second major downside of the overall process. Other than that, this adventure has been enriching at every level: whether it was about new connections made at laboratories and conferences, learned scientific skills, gained code development consistency, or a matured artistic vision.



## Conclusion

In this conclusion, we will try to summarize our contributions spanning theoretical ideas, scientific findings, computational tools, pieces of software and artistic propositions. In the preceding chapters, we have also put forward several leads for future work. We will restate them here, and try to articulate them with longer-term objectives, delineating a bigger picture of what a possible research program may look like in the future.

## Contributions

Chronologically, our first contribution was in the perceptual domain with a scientific paper on the perceived orientation of abstract paintings (see Appendix.A.1). It was a smaller scale project, conceived as a preliminary proof-of-concept for the main framework of this thesis, i.e. deep learning modeling combined with psychophysical investigations in humans. Our deep learning model, characterized by custom classifiers inserted after each convolutional block of a pre-trained VGG<sup>1</sup>, was designed to study orientation perception at different depths in the processing pipeline. In the article, we demonstrate that the model captures several human characteristics of orientation perception across granularities and styles. Indeed, abstract art, more than other styles, requires spatially extended integration of orientation cues for these to cohere into a reliable orientation estimate. We also tested fragmented stimuli in our experiments with humans, and we found that the detailed operations of the human mechanism are not identical to the modeled counterparts for small fragments, corresponding to superficial layers.

Concerning the core of this manuscript, its main theoretical contribution is the definition of a compositional paradigm (see Chapter.1). We regard this as a foundational step for any serious modeling approach, i.e. the stage at which the modeler attempts to understand something fundamental about the modeled phenomenon. In the machine learning field, the nature of inputs  $\mathbf{x}$  does not really matter as long as the model fulfills its practical requirements. The versatility of deep learning architectures and their associated power support generative latent spaces that can produce art-like artifacts, without explicitly questioning whether those spaces make sense with regard to the wider conceptual framework surrounding the learned data, e.g. the collective history of western painting. It is possible to build a latent space or a representation of any source of graphics, but the possibility of achieving this goal does not in itself guarantee pertinence of the final result, neither with regard to the pictorial material in question, nor with regard to the original approach of the artist to art. In our case, we wanted to answer the fundamental question of whether a space of compositions, in the form of a hyper-compositional object, would be artistically relevant. We have demonstrated that this goal is within reach, provided we accept that artworks may be regarded as the recording matter of compositional regularities, that the specific artistic

---

<sup>1</sup>The VGG is a standard model used in vision and image classification (Simonyan & Zisserman, 2014).

practice under scrutiny is serial and focused on dynamical interaction of graphical elements, and that strokes and compositions are the result of some morphogenetic process, intrinsically defining a continuous space of possibilities. Our approach is not fully consolidated yet, but it already constitutes a strong anchor point for my own artistic work.

A related contribution is to consider compositional practice as the organization of a system (see Subsection.1.2.System complexity and system organization). This idea is appealing in that it provides a clear and rich optimization objective for the arrangement of graphical elements on the plane. Composition becomes a creative *in-between*, where the artist intuitively creates conditional constraints between forms that are neither too weak, nor too strong. Thus, we attribute to our probabilistic approach a deeper implication than determining the *best* or *optimal* art. We also present both compositional space and compositional plane as probabilistic spaces with inherent diversity and richness of alternatives (see Section.1.3). We believe this framework presents an attractive view of machine learning algorithms and their associated optimization objectives. We finally put some effort into the description of the implications associated with high dimensionality in relation to potential fears about *normalizing* art, and corresponding misconceptions about probabilistic maxima.

Even though at present I do not intend for my personal dataset of compositions to become publicly available, I consider this fundamental resource as a contribution. It represents an implicit huge amount of manual work, from its creation to its processing into a dataset. To this end, we have developed custom software to ease operations. Concerning the processing pipeline, we have mostly mobilized existing image processing libraries and reimplemented some algorithms, such as the vectorization routine to cubic Bézier. Nonetheless, this procedure aggregates several small contributions, giving for instance more control over skeleton disentanglement at intersections, and over parameterization and simplification of curved elements (see Section.2.2). Finally, I am particularly proud of the algorithmic block that shuffles graphical elements within a composition down a hierarchical tree, with an option to limit the number of permutations (see Subsection.2.3.Composition permutations).

Compared with previous deep learning models applied to simple line drawings, our work contributes several innovations. The first essential innovation concerns the parameterization of curves. Our approach goes beyond mere encoding as a sequence of line segments and, in doing so, remains closer to motor intentions and artistic gesture of artists. It also offers more flexibility and precision in matching the original curves (see Subsection.2.2.Parametric curves). Secondly, we propose hierarchically nested stroke and composition models (see Section.3.2 and Section.3.3). The primary drive behind this choice of methodology is to capture and capitalize upon the fundamental temporal difference in the nature of these two action sequences: strokes are ordered sequences of Bézier components, while

## Conclusion

compositions can arise from stroke series in no particular order, possibly even incomplete. We specifically attempt to project each family of partially defined compositions onto a unique location in latent space, thus equipping this space with richer encoding power. Finally, we complete our journey by introducing a compositional plane model dedicated to the characterization of conditional constraints associated with graphical elements on the canvas. To create its predictions, this model relies on the two pre-trained nested models: the stroke model and the composition model (see Section.3.4).

In our project, there is no obvious metric to assess model efficiency. We therefore regard our training procedure with associated monitoring metrics as tools for aiding architectural design and hyperparameter selection (see Section.4.1). However, our contribution to the field of neural networks and representation learning is precisely realized through the many training tricks implemented by our procedures, which we have attempted to compile in a comprehensive and cohesive manner with this thesis (see Section.3.5). For instance, we introduced adaptive constraints on  $D_{\text{KL}}(q(\mathbf{z} | \mathbf{x}) || p(\mathbf{z}))$  by adding a nonlinearity per dimension. We also formulated a procedure to optimally re-balance model resources and partially overcome uneven dataset statistics, e.g. sequence lengths. Another significant contribution is represented by the limitation imposed on output variance, and the involvement of a new unit called *BackwardClamp*. Collectively, these technical innovations help to handle small datasets that nonetheless span a (too) diverse native space. They support the construction of a continuous and expressive representation despite operating within the low dimensionality of the chosen format. Our selected dimensionality is indeed smaller than previous works by several folds. This constraint is adopted primarily for easier manipulation and interpretation of model parameters, and with a view towards the perceptual experiments.

Another important contribution relates to the compositional metrics offered by the different models (see Section.4.3). For the composition model, these measurements are elaborated around encoder and decoder distributions. They specify not only position in latent space or on the plane, but also the associated degree of uncertainty, individually for every latent dimension and stroke. In the compositional plane model, we even have access to conditional probabilities on the plane, as a complex field of possible next strokes (location and shapes) given existing graphical elements, and a target position in the compositional space. For the purposes of this thesis, we restrict the presentation to visualization methods and insights from the different metrics. However, we are encouraged in pursuing this line of research further in the future by our first successful attempt at predicting perceptual scales from Fisher information computed on these compositional metrics (see Subsection.5.3.Perceptual scale prediction from Fisher information).

We have subsequently verified some qualitative aspects of the latent space. We were particularly interested in whether the model had captured regularities that align with important structural features of human perceptual space. The main difficulty

arises in connection with the dimensionality of the latent space: although relatively low (16 dimensions), it does not lend itself to feasible exhaustive experimental exploration. We have run several simulations with different methods before reaching this conclusion (see Section.5.1). Having taken stock of this limitation, we searched for alternative approaches that would allow us to tackle the issue at least indirectly. In the machine learning community, it has often been reported (but never adequately quantified) that a qualitative drawback of generative models is the existence of low-density regions within their latent space. These regions are often encountered during interpolation. It is therefore reasonable to posit that homogeneous density in latent space should be essential to the perceptual quality of interpolations. We performed perceptual scaling experiments involving similarity judgments with a triplet variant of the MLDS protocol (see Section.5.3). We found unexpected distortions, possibly reflecting discrepancies between human and model representations. Nevertheless, we were able to predict some non-trivial alterations of the perceptual scale using the auxiliary compositional metrics provided by the model (described above). Despite flaws in its representation, the model is thus able to incorporate information that is useful to predict and correct homogeneity in latent space. In short, our results indicate that the model has captured important compositional regularities that are, at least coarsely, aligned with human perception. However, only further iterations on this framework will provide us with a more definitive answer.

One last, but not least, interesting scientific contribution concerns the MLDS methodology itself. To incorporate circular trajectories through latent space, we have extended MLDS to periodic physical spaces (see Section.5.2). This variant is actually possible because most triplets, as defined in the canonical method, are barely informative and can be discarded. As a result, we can also drastically reduce the number of combinations per condition, directly influencing task duration. However, to understand and prove this insight, it was necessary to conduct more theoretical work that had not been originally planned. We were able to derive tighter bounds between Thurstonian scaling and MLDS, and provide an explanation for the variously reported discrepancies between the two methods. This fascinating work exposed theoretical issues associated with MLDS, in particular concerning the non-normal character of the distance metric between compared pairs.

Concerning the reproduction of generated drawings on paper with a pen-plotter, I developed different techniques to render expressive and dynamic stroking of lines. I adopted a strategy that adapts to the scale of the drawing and to the nature of the tool – pen or brush (see Section.6.2). Finally, I proposed different visual principles for representing the diversity and the continuity of compositional space, and for revealing the dynamics of graphical elements (see Section.6.3).

To summarize the whole project, our approach serves to validate a novel modeling framework for pictorial composition. We first introduce a compositional paradigm that supports a working deep learning model, and attests that composition can be

## Conclusion

modeled as a continuous hyper-dimensional object. We then demonstrate that captured compositional regularities and associated metrics present similarities with human perception. The adopted models and experimental protocols are still in their early stage of development and therefore present some limitations. However, our overall results are encouraging and point to the real possibility that complex perceptual phenomena such as art and composition, which are not easily reducible to elementary components, may be studied with some degree of quantitative scrutiny.

At the beginning of this manuscript, I asked whether my artistic practice and the proposed modeling approach could serve as meaningful points of contact between creative endeavors and the methods of scientific research. I hope that the presented work and associated contributions make a compelling case for how the particular can bring insights to more universal knowledge.

## Future works

We have presented most of our contributions at conferences. However, except for the work on orientation perception of paintings<sup>2</sup>, none of this research has been published yet. After my PhD, I plan to focus on writing up the unpublished material. More specifically, I hope to complete an interdisciplinary paper on the use of MLDS to study latent spaces, and Fisher information to correct for resulting distortions. It could be beneficial to scientific communities interested in generative models, such as those operating in visual perception and machine learning. I also plan to make all associated code publicly accessible, which will require some polishing, refactoring and additional commenting. This could be an opportunity for me to contribute back to the open-source community, which gave me so much during this project.

Beyond publishing this research and making it available to the wider community, we can identify a number of future projects at both artistic and scientific levels, ranging from the development of technical tools to the introduction of fundamental new ideas. However, it is not clear what timescale would be involved with these projects and ideas, whether short or long term. For instance, we may find that it is not a priority to improve some major conceptual innovations, as they are already sufficiently developed to support further research. Below, we review a number of technical details that arose during the project, and which we plan to pursue further in the future.

When matching distributions, e.g.  $q(\mathbf{z} | \mathbf{x})$  to  $p(\mathbf{z})$ , we found that using Maximum-Mean Discrepancy could be a useful alternative to Kullback-Leibler divergence. This metric may improve qualitative issues connected with the non-symmetric nature of  $D_{\text{KL}}$ , which leads to over-generalization of modes within the target

---

<sup>2</sup>Lelièvre and Neri, 2021.

distribution. With relation to the optimization function, we plan to investigate the possibility of matching higher moments for the regularizer  $D_{\text{Cov}}(q(\mathbf{z}), p(\mathbf{z}))$ , enforcing  $q(\mathbf{z})$  to match the prior and present a good separation of  $\mathbf{z}$  dimensions. We believe that an approach of this kind could also address issues associated with the bimodality of distributions along some dimensions of the latent space. Finally, on this topic, during our project we found that prediction of the perceptual scale from Fisher information using our encoder presented some re-mapping issues. A given  $\mathbf{z}$ , deterministically decoded to the corresponding  $\mathbf{x}$ , should be mainly re-encoded around  $\mathbf{z}$ , but this is not always the case, especially when an additional stroke pops in or out during interpolation. The introduction of a feedback loop at training could reduce this class of failures. Another, possibly more effective approach to solve this problem may involve a regularizer directly incorporating human perceptual behavior through Fisher information from the decoder. Instead of repeated sampling of encoded  $\mathbf{z}$  at training, we could consider a local region around those  $\mathbf{z}$  as involving a small interpolation on a hypersphere that must correspond to a linear perceptual scale.

We can also identify several future directions for our research on *human perception*. First, we want to better understand why GLM fits are less accurate than their MLE counterparts, especially concerning the  $\sigma$  value of perceptual noise. This quantity is related to the  $\alpha$  clipping value, but the causal implication is unclear. For MLDS and PMLDS, we pragmatically defined a value for  $\tau$  (related to step size) because this seemed optimal based on pilot simulations, however the validation of this choice with real data is more challenging. Further research will be necessary to turn our heuristics into a proper method. As we said in the manuscript, a good start for future works should involve simpler and well-known stimuli, already validated with traditional methods such as 2AFC for the derivation of psychometric functions. We also hope to improve our method for computing Fisher information from the decoder, especially for series of compositions with different stroke numbers.

With regard to more general ideas for future work, we could explore new ways of exploiting different compositional metrics, or identify viable methods to experimentally measure low frequency distortions in latent space. We could also improve the compositional plane model, particularly concerning its tendency toward stroke repetition.

An obvious direction of future work at an artistic level would involve extending and enriching my personal composition dataset. An effort of this kind may also be beneficial to the associated scientific research. Model development critically depends on the amount of input data and density of intermediary exemplars. Even without improving the balance of general sequence length in the dataset, more inputs should increase the diversity of longer strokes. During my PhD, I found that I had less time to draw, but I never really stopped. Thus, I have already accumulated a good amount of new material. Having said all the above, we may decide that it is not advisable to embark on another iteration of the project, and

## Conclusion

that we should instead focus on exhausting the creative possibility of the current trained model, and fully explore the associated visual principles aiming at rendering dynamics.

The artistic questions that sit at the core of this project also call for longer-term developments. First, we would like to define intermediary levels of graphical elements between strokes and composition. In an effort to identify common groups of strokes (e.g. squares and other simple graphical structures), we could benefit from the statistical constraints and regularities discovered by the current compositional plane model. It would also be exciting to implement support for stroke widths. At the model level, this upgrade would only imply marginal modifications of layer dimensionalities. However, it would be very costly at the level of dataset processing, because current vectorization tools do not take this aspect into account (see Subsection.2.2.Surfaces to lines). Furthermore, vector graphic standards and file structures are not currently designed to incorporate this information. For these reasons, it would be difficult to achieve good storage, visualization, and artistic usage without the introduction of significant technological developments. Finally, in relation to these topics, we are fascinated by the possibility of exploring the diversity of the latent space with real-time interactive applications, and we think it would be worth investing time into possible interfaces.

I would also very much like to pursue a more theoretical aspect of my work: the concept of self-organization applied to composition. In particular, I would like to understand how artists manage to overcome the structural and functional complexities that they encounter during the creation phase. I believe there is a lot to explore and learn from the notion of *emergence*, i.e. the spontaneous appearance of a macrostructure that was not predictable from individual knowledge of its sub-elements. In parallel with this line of inquiry, I would like to investigate traditional Chinese painting. Masters of this art form were able to tightly integrate their art with knowledge about the world, unifying the two using the regularities of a single gesture. These different ideas for future research may be re-articulated using more contemporary concepts.

Along a different direction, I can identify another possible strategy for future work on topics of theoretical importance. I have always regarded work on my personal compositions as a proof of concept, a starting point for future generalizations. Thus, instead of consolidating the current model, I could begin to address applications of the model to different problems, e.g. kanji/hànzì calligraphy. I have already used kanji to experiment with some deep learning architectures, thinking it would be easier and/or more objective to determinate if a generated character was meaningful or not. However, the lack of real continuity in character space presented serious challenges, motivating me to consider material of a different kind. Similar challenges also apply to compositions, e.g. concerning the stroke number per drawing, but they are less problematic. This line of future work is interesting and appealing, as it would certainly benefit the study of other compositions,

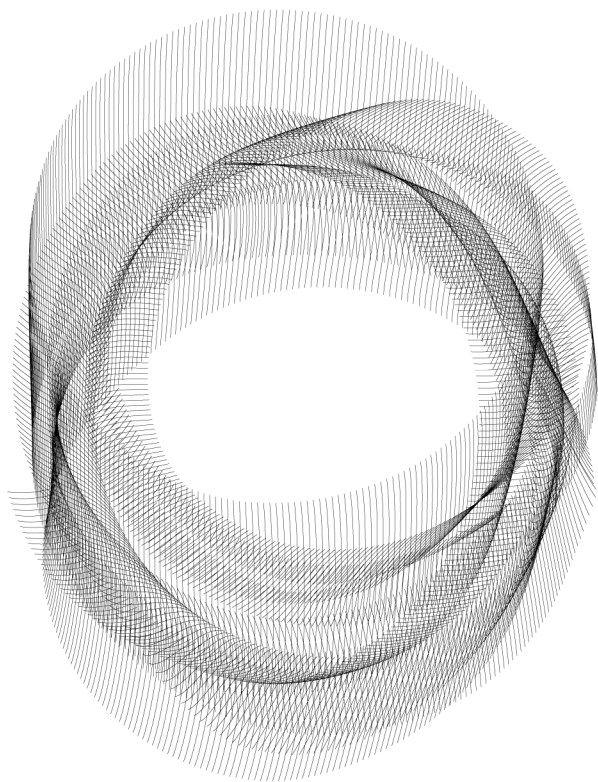
and may serve to demonstrate the general applicability of the whole framework. Furthermore, it could provide me with useful hindsight into some architectural points that may need reconsideration, without coming into too close contact with the manipulated material (I am not a fluent reader of kanji/hànzì yet). Finally, it could offer the chance to start collaborative work, and possibly form the basis of an interesting postdoctoral proposal.

Alongside the above projects, and outside the connection between scientific and artistic activities, I could invest some effort into making the transition from what is currently an introspective practice, to an artistic activity that is identified as such by the community. For instance, I could prepare an exhibition without the constraint of establishing scientifically relevant connections. The change of scale that would go with such an endeavor would probably prompt me to reformulate some questions, and trigger new ones. In addition, I could attempt to sell my artworks, not so much for financial reasons but because this is an important step toward art as it is organized and recognized nowadays. It would be important for me to tackle the task of understanding cultural institutions, and how to use them as stepping stones.

After finishing the writing of this manuscript, which collects and summarizes the work done during my PhD, I have finally come to realize and understand what is coming next. The above paragraphs are characterized by a certain degree of strategic indecision concerning the organization of my future research agenda over the longer term. I believe that this lack of a rigid plan is a source of potential difficulty, but also an opportunity: as a postdoc, I should be able to explore different research directions, as this is an inevitable and necessary path towards becoming a senior scientist. The maturity that would come from this process may simultaneously extend my artistic practice, and allow it to flourish at the same time. A mesmerizing and thrilling horizon of possibilities opens before me.



## Conclusion





# Appendices

## A.1 *A deep learning framework for human perception of composition*

Beyond contributions to the understanding of the orientation judgment of abstract paintings, this initial research, conducted during the first year of my PhD, can be seen as a proof of concept of the method presented in this manuscript, i.e. deep learning modeling combined with psychophysical investigations in humans. It helped me to familiarize with the deep learning library PyTorch, and more importantly with the field of psychophysics. Online experiments also pushed me to implement my own platform for data collection<sup>1</sup>.

Preliminary results have been presented during a talk session at ECVF 2019, the European Conference on Visual Perception<sup>2</sup>. The final paper (Lelièvre & Neri, 2021), published in May 2021, is reproduced as it can be found in *Journal of Vision*<sup>3</sup>. We also include associated supplementary materials.

---

<sup>1</sup>Experiments are still available on my website: <https://plelievre.com/experiments>

<sup>2</sup>Lelièvre and Neri, 2019.

<sup>3</sup>DOI: <https://doi.org/10.1167/jov.21.5.9>

# A deep-learning framework for human perception of abstract art composition

Laboratoire des systèmes perceptifs,  
Département d'études cognitives Science Arts Création  
Recherche (EA 7410), Paris, France  
École normale supérieure, PSL University, CNRS,  
Paris, France



Pierre Lelièvre

Laboratoire des systèmes perceptifs,  
Département d'études cognitives, Paris, France  
École normale supérieure, PSL University, CNRS,  
Paris, France



Peter Neri

**Artistic composition (the structural organization of pictorial elements) is often characterized by some basic rules and heuristics, but art history does not offer quantitative tools for segmenting individual elements, measuring their interactions and related operations. To discover whether a metric description of this kind is even possible, we exploit a deep-learning algorithm that attempts to capture the perceptual mechanism underlying composition in humans. We rely on a robust behavioral marker with known relevance to higher-level vision: orientation judgements, that is, telling whether a painting is hung “right-side up.” Humans can perform this task, even for abstract paintings. To account for this finding, existing models rely on “meaningful” content or specific image statistics, often in accordance with explicit rules from art theory. Our approach does not commit to any such assumptions/schemes, yet it outperforms previous models and for a larger database, encompassing a wide range of painting styles. Moreover, our model correctly reproduces human performance across several measurements from a new web-based experiment designed to test whole paintings, as well as painting fragments matched to the receptive-field size of different depths in the model. By exploiting this approach, we show that our deep learning model captures relevant characteristics of human orientation perception across styles and granularities. Interestingly, the more abstract the painting, the more our model relies on extended spatial integration of cues, a property supported by deeper layers.**

## Introduction

Artistic graphical composition can be roughly defined as the structural organization of pictorial

elements on a canvas. Art history offers some basic rules and heuristics for understanding the qualitative characteristics of this phenomenon; however, it does not codify processes such as segmentation/interaction of pictorial elements to the degree of specification required by quantitative analysis. Modern artists such as Kandinsky or Klee initiated some systematic and almost scientific studies on this topic (Kandinsky, 1989, 1991; Klee, 1961, 1973, 1998), but they struggled with the combinatorial complexity afforded by compositional questions. Despite more recent progress in this area (Arnheim, 2004), composition remains a complex amalgam of different phenomena, highly dependent on context and other aspects that are not easily quantified. Composition also represents a versatile experimental tool for empirical aesthetics (Locher et al., 1999; McManus et al., 1993; Schwabe et al., 2018); however, this approach focuses primarily on aesthetic judgements, rather than the compositional processes associated with those judgements.

Recent advances in machine learning, and particularly deep architectures, have demonstrated the ability of artificial neural networks to extract hidden structure from high-dimensional data and solve complex problems with human-level performance (Dodge & Karam, 2017; Serre, 2019). Our goal is to discover whether deep learning tools can advance our understanding of composition and whether, by relying on those tools, we may define a partial, yet relevant, metric description of this phenomenon that is available for quantitative scrutiny (see Iigaya et al., 2020 for related methodology). To achieve this goal, we rely on a well-defined and robust perceptual judgment of visual orientation that is related to composition: telling whether a painting is hung “right-side up.”

Citation: Lelièvre, P., & Neri, P. (2021). A deep-learning framework for human perception of abstract art composition. *Journal of Vision*, 21(5):9, 1–18, <https://doi.org/10.1167/jov.21.5.9>.

<https://doi.org/10.1167/jov.21.5.9>

Received April 8, 2020; published May 11, 2021

ISSN 1534-7362 Copyright 2021, The Authors

a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License



Under the assumption that the orientation of reference for a painting is that selected by the artist, previous work has demonstrated that humans can perform this task well above chance, even for abstract paintings, and regardless of their level of familiarity with painting material (Lindauer, 1969; Mather, 2012). Therefore, it seems that orientation judgments represent a robust behavioral metric, even for material with no recognizable content. Orientations other than the reference orientation may elicit equally valuable subjective interpretations and/or aesthetic experiences in the viewer; however, existing empirical evidence indicates that part of the orientation judgment is consistent across observers: not necessarily directed *toward* the orientation of reference, but at least directed *away* from some of the alternative options. Furthermore, orientation judgments are of immediate relevance to the study of visual perception, an area where image orientation is often manipulated to selectively target higher level processing (see for example the well-known inversion effect (Neri, 2014; Valentine, 1988) and its numerous applications (Cusack et al., 2015; Gaspar et al., 2008; Kelley et al., 2003; Neri et al., 2006, 2007; Yovel & Kanwisher, 2005)).

The exact mechanisms underlying orientation judgements are not fully understood. Some authors have suggested that the perception of orientation depends more on low-level stimulus properties than higher level object recognition and/or image interpretation (Lindauer, 1987), prompting others to investigate the potential role of relatively simple cues, such as Fourier amplitude spectrum slope (Mather, 2012), or image statistics based on explicit rules gathered from several art theories incorporated into a machine learning algorithm (Liu et al., 2017) (see Elgammal et al., 2018; Rodriguez et al., 2018 for related applications).

In approaching these issues, we do not commit to restrictive assumptions or purpose-built schemes. Our model is structured around a general architecture not originally devised for application to art material. We exploit a large database of paintings to train the model, and in so doing we automatically approximate the perceptual mechanisms underlying composition. Despite not being hand-engineered to tailor our specific problem of interest, the trained model outperforms previous applications and extends to a greater variety of painting styles.

It is generally believed that orientation judgments are supported by global analysis of the scene (Oliva & Torralba, 2006). The role of local cues has been relatively unexplored, and more generally the granularity of this phenomenon is not well-understood (Gong et al., 2018). Within the context of our approach, we can naturally probe the issue of granularity and identify the appropriate scale for understanding pictorial elements. More specifically, by exploiting

the hierarchical architecture of our model, we can explore how information is represented at different depths within the network. We find that the use of small-scale patterns and deeper level features shows qualitative differences between abstract paintings and more realistic pictorial styles.

To validate the applicability of our model to human visual perception, we carried out a web-based experiment with human observers. They were asked to perform the orientation judgment task on whole paintings as well as fragments of different sizes, corresponding with the different extent covered by the receptive field of distinct depth levels in the model. These experiments were designed with the following goals in mind: establish whether human performance on the orientation task can survive a wider range of stimulus manipulations (painting style, abstraction level, fragment size) than previously tested in the literature; and determine whether our model provides a satisfactory account of the human process. We find positive answers to both questions, although we did identify some discrepancies between human and simulated results, which serve as useful starting points for us to elaborate on how the proposed model may be augmented in future work.

## Methods

### Database

Our image database is derived from the WikiArt web encyclopedia (WikiArt). The associated API returns metadata such as artist identification and painting styles of each image. At the time of this experiment (May 2019), the WikiArt database contained 157,291 entries. We excluded non-painting styles (e.g., performing arts) and pictures of painting details, reducing this figure to 141,892 items. To make our results directly comparable with those reported by Mather (2012), we manually added 18 entries and moved all paintings from this paper in the validation set. Because our interest is mainly in how model performance varies with style (e.g. abstract vs. figurative), we ensured that different styles and artists were comparably distributed between the training and validation sets. With a target validation ratio of 0.1, the final split is 126,451/15,459. We grouped entries into the genres and styles detailed in Supplementary Tables S1 and S2. Representative examples from this selection are shown in Figure 1. Chosen classification is largely unambiguous, but there are instances for which the specific choice of genre/style may be disputable from historic and/or artistic perspectives. For instance, abstract style is often associated with modern/contemporary Western movements; from such a viewpoint, our decision to

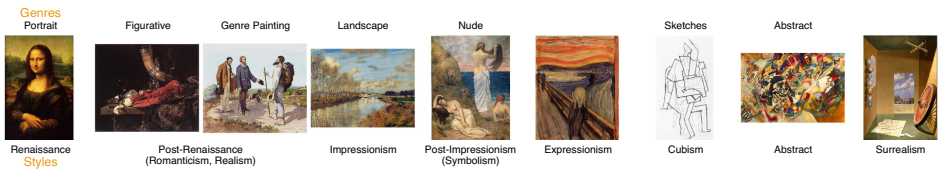


Figure 1. Gallery of genres and styles mentioned throughout the paper. Ordering is chronological. (*Mona Lisa* by *Leonardo da Vinci* (1503-1519), *Still-Life with Drinking-Horn* by *Willem Kalf* (1653), *The Meeting (Bonjour Monsieur Courbet)* by *Gustave Courbet* (1854), *Argenteuil seen from the small arm of the Seine* by *Claude Monet* (1872), *Young Girls on the Edge of the Sea* by *Pierre Puvis de Chavannes* (1879), *The Scream* by *Edvard Munch* (1893), *Seated man with his arms crossed* by *Pablo Picasso* (1915), *Komposition VII* by *Wassily Kandinsky* (1913), *A Naturalist's Study* by *Pierre Roy* (1928)).

include Native art in the abstract category may seem questionable. This decision, however, is motivated by our focus on visual abstraction, rather than abstraction as defined by historical criteria. Furthermore, the questionable instances represent <1% of the total, rendering this issue of little concern. A more probable source of bias is represented by the portrait/landscape aspect ratio. We address this issue in the Supplementary Material, where we demonstrate that this bias is negligible and that the aspect-ratio distribution is well-balanced for abstract paintings, the class we are most interested in.

### Model architecture

The task of orienting an image can be thought of as a simple classification problem with four classes, each class corresponding with one possible orientation for the painting. Within this family of machine learning problems, the classification of items from ImageNet (Russakovsky et al., 2015) has led to the development of several deep learning models dedicated to image processing, in particular convolutional neural networks. There is now extensive evidence highlighting similarities between convolutional neural networks and the mammalian visual pathway (Kriegeskorte, 2015; Yamins & DiCarlo, 2016). Among such artificial neural architectures, the most popular are AlexNet (Krizhevsky et al., 2012) and VGG (Simonyan & Zisserman, 2014). Based on its complexity and reported accuracy on ImageNet, we selected VGG-16 (PyTorch implementation; Paszke et al., 2019) as an appropriate starting point for this study.

Figure 2 shows the schematic architecture of our network. All convolutional blocks in gray (1–5) are directly ported from VGG. They consist of multiple convolutional layers with rectified linear units (ReLU) activation functions followed by max-pooling. Our implementation does not use batch-normalization and we removed the original linear layers of the classifier to

be replaced by a custom-designed classifier-5, composed of a convolutional layer (kernel size = 7, stride = 3) and linear layers (sizes = [512, 128, 32]). ReLU activation functions and dropout units are applied to all layers except for the last one, to which we applied a softmax function for classification purposes. The dropout rate is of 0.30, except for units before the last layer with a rate of 0.15.

The main feature of our network is that its linear layers are convolutional with kernel size 1. We adopted this formulation to enable inspection of the spatial distribution associated with classified outputs. The consequence on classifier-5 is null because, at this depth in the network, its output (height = 1, width = 1, classes = 4) is generated by a receptive field covering the entire input image. The implication for the other classifiers (1–4), inserted after each convolutional block corresponding to earlier visual areas, is that they have access to small receptive fields. As a consequence, classifier-1 (earliest level) produces for example a classification output of shape (36, 36, 4), as if the network simultaneously judged the orientation of multiple fragments across the picture. This architecture makes it possible for us to inspect network behavior at different depth and for cues of differing granularity.

### Training procedure

Input images conform to the VGG format with resolution  $224 \times 224$  pixels and color normalization computed from the ImageNet database. In principle, all parts of a painting may be relevant to judging its orientation, making it inappropriate to crop images into a square shape. We therefore scaled images so that their largest dimension was 224, and fill the remaining empty space with the ImageNet mean value (Figure 3a). These manipulations raise two possible concerns. First, downsampling to a lower resolution may leave out useful orientation cues from the original

## A.1 A deep learning framework for human perception of composition

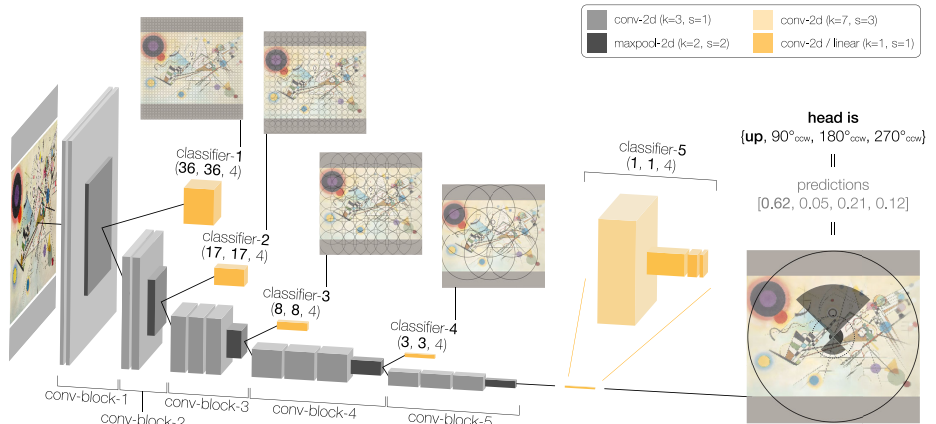


Figure 2. Schematic architecture of the multilevel orientation classification model employed in this study. Each of five convolutional blocks is associated with a classifier (indicated by classifier- $n$  with  $n = 1$  to 5). The output dimensionality of each classifier is indicated by  $(x, x, 4)$ , where  $x$  is the number of samples across each spatial dimension (see density of circle array within insets overlaying local filters onto painting), and 4 is the number of orientation labels  $\{\text{up}, 90^\circ, 180^\circ, 270^\circ\}$ . The four values within [ ] show one example of the categorical distribution generated by the network for *Komposition VIII* by *Wassily Kandinsky* (1923). In the legend,  $k/s$  stand for kernel/stride size.



Figure 3. Effect of median filtering on network attention, visualized through guided error back-propagation. Error map is inverted and thresholded for legibility. Light gray indicates pixels where attention reaches at least 1% of its maximum (moderate attention); dark gray indicates pixels where it exceeds 10% (high attention). (a) shows original images used for training. (b) shows directed attention in the absence of median filtering applied to the borders, (c) in the presence of median filtering. Two examples by *Paul Klee* are shown: *The Place of the Twins* (1929) and *After Annealing* (1940).

image. This is possible; however, general considerations about the nature of the images, combined with cursory inspection of representative examples, indicates that composition is a global property that is retained at the adopted resolution. For example, the images shown in Figure 1 are downsampled using the same algorithm we used for the experiments: these paintings are still highly recognizable and understandable. Furthermore, our study is designed as a comparative behavioral experiment between humans and a deep learning model; we expect that the two systems should be similarly impacted by downsampling. The second

potential concern relates to color normalization of the paintings. If carried out incorrectly, this procedure may disrupt the perceptual analysis of color and partially alter compositional effects. To avoid this undesirable outcome, we compute mean and standard deviation per channel at the dataset level, not at the level of individual images. Therefore, when normalization is carried out using these mean and standard deviation values, relative color differences and local contrast are conserved at the painting level.

We minimized overfitting using simple data-augmentation techniques: first, we applied random



color gamma correction  $output_c = input_c^{\gamma_c}$  where  $\gamma_c = 2^{0.5a+0.25b}$ ,  $a$  a random scalar and  $b$  a random vector sampled from uniform distributions over  $[-1, 1]$ . Second, images are randomly rotated by up to  $5^\circ$  in either direction and randomly shifted along their shorter dimension within a range such that the whole image remains visible. To accelerate training, we relied on the pretrained model provided by PyTorch. Parameters for the convolutional blocks are not fixed, so they are fine tuned for painting material during training. When filter parameters are fixed, performance is substantially reduced (see Supplementary Material). We used cross-entropy loss for optimisation as is customary in classification problems. Optimisation is performed by an Adam algorithm with learning rate  $1e^{-4}$  and a scheduler that decreases this learning rate by a factor of 10 when network accuracy remains stable across two epochs.

## Testing procedure

At the adopted resolution of the input images, some pictures retained spurious cues to their original orientation, such as artist signatures or handwritten titles near the border. We solved this issue as follows. We initially relied on guided back-propagation to visualize regions emphasized by the model during a preliminary training procedure, and found that the network directed attention to artist signatures and other written characters usually within the bottom region of paintings (Figure 3b). These cues can be trivially exploited to determine picture orientation, but are not connected with composition, so our goal was to remove them as effectively as feasible in automated fashion (manual editing was not an option for such a large database). We applied a median filter with a ramp along all borders of each painting (filter of size 5, full on the outer 5% of the image and with a ramp to zero up to the 20% point). Median filtering is preferable to Gaussian filtering because it removes high-frequency noise while retaining sharp edges. This border-based median-filtering procedure is only applied during validation because it is not useful during training: the network is still able to learn residual artefacts associated with signatures. Figure 3c demonstrates that, even though the network has learned to exploit signatures during training, it successfully reallocates its attention to other parts of the painting when median filtering is applied to borders during validation. Results reported in this article (most importantly validation scores) are averaged separately for each painting over four presentations of that painting in every possible orientation. Model performance refers to average top-1 scores. Top-1 accuracy is 1 if the most probable predicted class is the targeted class, 0 otherwise.

## Web-based experiments

We developed a dedicated website for human data collection. Before accessing the experimental platform, participants registered and specified their age as well as their general knowledge of art material. In the first experiment, participants were required to select the original orientation of randomly picked abstract paintings successively presented in blocks of 10. Each painting was presented in isolation and could be oriented interactively by the user; once the participant was satisfied with a particular orientation, this was selected by pressing a button and triggered presentation of the next painting in the sequence. If any element in the painting could serve as obvious hint to the correct orientation, like a word or a signature, people were asked to report it via a dedicated button. After each series, a figurative painting of obvious orientation was inserted into the sequence to check whether participants were meaningfully engaging with the task. To motivate their interest and maintain their focus, participants were provided with feedback at the end of each series detailing performance scores and information about the paintings. In the second experiment, participants saw fragments of both abstract and figurative paintings. The fragments were sized to span the approximate size and location of fragments accessible to the network for each classifier. Under these conditions, the task was perceived as challenging and sometimes puzzling owing to the fragments often being small and blurry; however, it produced interesting results for understanding compositional perception at different granularities. Because we sought to randomly sample paintings from the same style distribution as the model dataset, we excluded categories with a small number of entries to avoid unreliable measurements. More specifically, the abstract category included the following styles (in decreasing order of representation): Abstract Expressionism, Abstract Art, Art Informel, Color Field Painting, Minimalism and Lyrical Abstraction; the figurative category only included Romanticism. We collected an average of 50 trials per participant from 71 participants aged between 15 and 67 and coming from 8 different countries. As an indication that our sample is representative of those commonly used in the literature, our measured average accuracy of 47% (Figure 9b) is highly consistent with values reported by existing studies (Lindauer, 1969; Mather, 2012). We excluded eight participants with scores of less than 0.75 for figurative styles and of less than 0.25 for abstract styles who had typically collected fewer than 10 trials. The inclusion of these participants lowers overall accuracy to 46%, but does not alter the general pattern of the results and their interpretation. We also recorded reaction time, age and general knowledge of art material (as self-reported via questionnaire); these factors are tangential to the present study, so

## A.1 A deep learning framework for human perception of composition

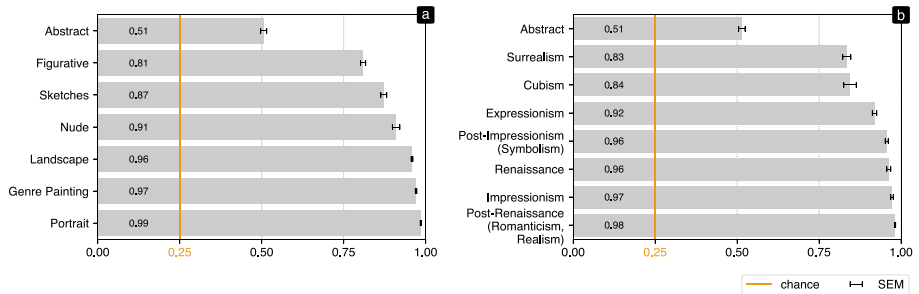


Figure 4. Model performance on whole paintings grouped by genre (a) and style (b).

they are only briefly discussed in Supplementary Material.

### Guided back-propagation

This neural network visualization technique consists of back-propagating the true class/label (binary one-hot distribution) as an error through the network all the way back to the input image. Because the network applies more correction to regions of the input image where information is most useful for achieving categorization of the back-propagated class, those regions map out the equivalent of attentional deployment by the network (Figure 3c, Figure 5b). The *guided* variant of back-propagation was introduced by Springenberg et al. (2015) to improve back-propagation of the gradient through ReLU activation units.

### Cross-entropy

Given target probability distribution  $p$  and estimated probability distribution  $q$ , cross-entropy is defined as  $H(p, q) = H(p) + D_{\text{KL}}(p \parallel q)$  where  $H(p)$  is the entropy of the target distribution (i.e., the average amount of uncertainty/information about  $p$ ) and  $D_{\text{KL}}(p \parallel q)$  is the Kullback–Leibler divergence from  $q$  to  $p$  (a measure of the difference between the two distributions). When target distribution  $p$  is the final classified label (binary one-hot distribution),  $H(p) = 0$  and cross-entropy simplifies to  $D_{\text{KL}}(p \parallel q)$ ; to optimize this function, the model simply pushes the  $q$  estimate to match  $p$  as closely as possible. We also compute cross-entropy for target distributions other than the final one-hot label; more specifically, we compute distributions for fragments at level  $n$  ( $q$  in notation above) and measure their predictive power for target

distributions of closest fragments at level  $n + 1$  ( $p$  in notation above). The goal of this between-level metric is to measure redundancy between distributions at different levels. To produce a more interpretable metric in Figure 8, redundancy is defined as  $\exp[-H(p, q)]$ . The maximum redundancy is 1, corresponding with 0 cross-entropy. A chance level can also be defined as the cross-entropy between equiprobable distributions, simplifying to a redundancy of 0.25 with four classes.

## Results

### Model performance on whole paintings

Model performance on whole paintings of the abstract genre is around 50% (Figure 4a), in excellent agreement with human measurements from existing literature (Lindauer, 1969; Mather, 2012). Performance also progressively improves from abstract to objects, landscapes through to portraits (Figure 4a). Qualitatively speaking, this progression seems to be related to the characteristics of possible orientation cues, such as their diversity and reliability. For example, Portraits (e.g., *Mona Lisa* in Figure 1) contain faces that are almost exclusively in the upright orientation, making for highly stereotyped and reliable cues. Genre Paintings often display people in standing position, during battles, religious ceremonies or everyday life (e.g., *The Meeting (Bonjour Monsieur Courbet)* in Figure 1); cues are still primarily restricted to human characters, but are less stereotyped due to different (potentially conflicting) body poses. Landscapes and Figurative genres display greater diversity of cues, more abundant but certainly less reliable: trees and clouds can be seen via water reflections and objects may not be associated with specific orientations. Along

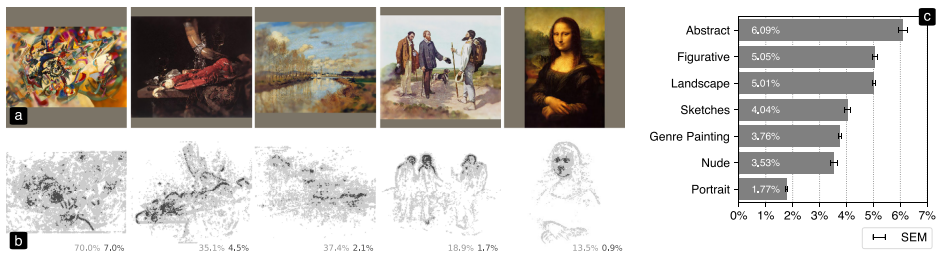


Figure 5. Network attention through guided error back-propagation (see Methods). (a) Five examples of original inputs for validation (*Komposition VII* by Wassily Kandinsky (1913), *Still-Life with Drinking-Horn* by Willem Kalf (1653), *Argenteuil seen from the small arm of the Seine* by Claude Monet (1872), *The Meeting (Bonjour Monsieur Courbet)* by Gustave Courbet (1854), *Mona Lisa* by Leonardo da Vinci (1503-1519)). (b) Error maps with inverted and thresholded intensity. Light gray indicates pixels where attention reaches at least 1% of its maximum (moderate attention); dark gray indicates pixels where it exceeds 10% (high attention). Numeric values report light and dark pixel percentages over the entire painting surface. (c) Average surface ratio of high attention, plotted separately for different genres.

this qualitative scale, Nude is perhaps the only genre that seems to be misplaced (right before Landscape in Figure 4a), because one may expect that it should be similar to Genre Painting. Looking at *Young Girls on the Edge of the Sea* in Figure 1, Nude paintings seem to explore an extended range of body poses, making body orientation a potentially unreliable cue.

To investigate this interpretation more quantitatively, we can visualize the network's error back-propagation, a technique that exposes regions where the network directs its attention during evaluation. The spatial organization of attentional deployment offers useful insight into the diversity of available cues. Consider *Mona Lisa* in Figure 5b: the most active attentional areas, indicated by dark gray pixels, are highly localized and limited to facial details. In comparison, Kandinsky's *Komposition VII* prompts the model to gather information across the entire image. For Monet's landscape and Kalf's still life, the model operates in a manner that appears to sit halfway between those two extremes, in line with the hypothesis described earlier. We attempt to quantify this trend by simply measuring the proportion of image pixels where the back-propagated attentional signal exceeds 10%. When plotted separately for the different genres (Figure 5c), this quantity is well aligned with the genre ordering of Figure 4a. If we adopt pixel area as a proxy for cue numerosity, the network model uses nearly 3.5 times more cues for Abstract paintings than Portraits. In this ranking, Nude is closer to Genre Painting, as expected from our earlier qualitative considerations. Finally, Sketches may be expected to occupy a position closer to the Abstract genre; however, the sparseness of line

content over the flat canvas may explain the lower ratio reported in Figure 5c.

A related concept for ordering model performance on different art material is the reliability/interpretability of available orientation cues, which may reflect the purported importance of “meaningful” content for orientation judgements. From this perspective, painting style (rather than genre) may offer better insight into the role of image content. For example, portraits from Leonardo da Vinci and Picasso (see Figure 1) encompass different degrees of ambiguity. With this notion in mind, Figure 4b demonstrates a lawful relationship between performance and abstraction level (concreteness): from abstract style to Cubism, Symbolism, and post-Renaissance realism. Therefore, taken as a genre or a style, abstraction is in both cases the most difficult material to orientate.

These observations may be summarized by the notion that, although abstract orientation cues are widely distributed across the canvas, they seem to carry limited predictive power. By and large, these visual features are likely employed by artists regardless of their orientation; nonetheless, the associated performance in the orientation judgment task is well above chance. A recent study (Specker et al., 2020) reports that human observers share artistic judgment more effectively in relation to whole abstract artworks as opposed to isolated elements (lines and colors). Therefore, it appears that, in the absence of preferred orientation for individual elements, the only effective source of information must come from the combination of the different cues into specific arrangements that may or may not be represented at the level of the perceptual/neural process. The progressive

## A.1 A deep learning framework for human perception of composition

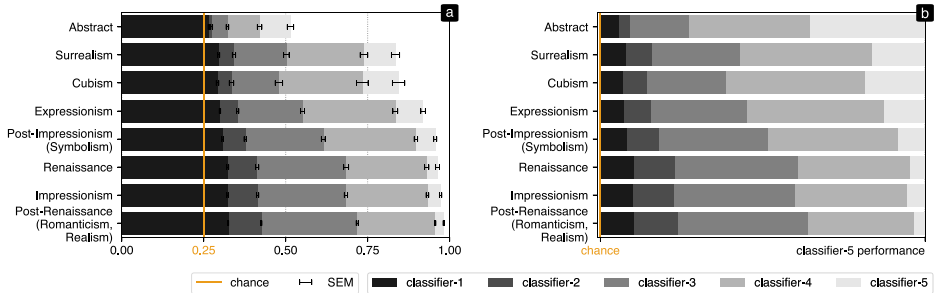


Figure 6. Model performance across classifiers. Values are grouped by style (as in Figure 4b) and displayed separately for the five distinct classifiers. (b) plots values from (a) after rescaling between chance and maximum value for given style (corresponding to performance of classifier-5).

construction/representation of compositional patterns is a phenomenon of central interest to our study, and one which we hope to understand further by examining the role of local image cues/fragments in greater detail (discussed further in the next Section).

### Model performance on fragments

We are in a position to study model behavior for earlier layers via inspection of classifier-1:4 (Figure 6a). Model performance shifts toward chance as its spatial resolution is restricted to smaller receptive fields and could be rephrased as “more is better.” We interpret this trend as reflective of the commonly regarded high-level nature of the orientation judgement (Neri, 2014; Valentine, 1988). Perhaps related to this observation, a recent investigation of human aesthetic judgement viewed from a neural network perspective (Igaya et al., 2020) reports that judgments of “concreteness” become increasingly dominant with neuronal integration. We also find that, when values are normalized by the performance level associated with classifier-5 (Figure 6b), the dependence on deeper layers increases with abstraction level of the painting.

We can gain more insight into the issue of granular representation within the model by plotting predicted orientations from individual receptive field units (Figure 7). The first and most obvious characteristic of these results is that figurative paintings are more spatially redundant than abstract paintings: they offer orientation cues more uniformly spread across the image down to small scales. Further to this observation, although the results for figurative paintings at coarser scales can be roughly predicted from those at finer scales via simple integration of local cues, this rule does not

seem applicable to abstract paintings: a large fragment is not reflected by simple averaging of smaller related fragments.

To quantify redundancy between adjacent classifiers, we measure how well distributions at level  $n$  describe those at level  $n + 1$  using rescaled cross-entropy (see Methods). This quantity is plotted in Figure 8; it ranges between chance (level  $n + 1$  cannot be predicted by level  $n$ ) and ceiling performance (level  $n + 1$  can be fully predicted by level  $n$ ). First, we notice that redundancy increases as we transition from the earlier layers to the later layers, meaning that redundancy increases along the processing pipeline. For example, redundancy between classifiers 1-2 and 2-3 remains near chance across all styles. As we transition to later levels (description of classifier-5 from classifier-4), figurative paintings show a strong correlation between classifiers, while abstract paintings remain close to chance.

A different (but related) way of thinking about Figure 8 is to consider the progressively expanding horizontal bars for figurative paintings as reflecting a gradual emergence of a structured representation that is largely shared across layers. Whatever properties are being represented by the network to support classification, their representation is constructed incrementally along the processing hierarchy and is therefore distributed across layers. In the case of abstract art, representation of relevant properties does not appear to emerge gradually along the pathway. Classifier-5 seems to represent a global property of abstract art that is not transparently available from earlier layers, and which we speculate may be connected with composition. It is true that earlier layers support an appreciable level of task performance (see Figure 6a), but our cross-entropy analysis indicates that

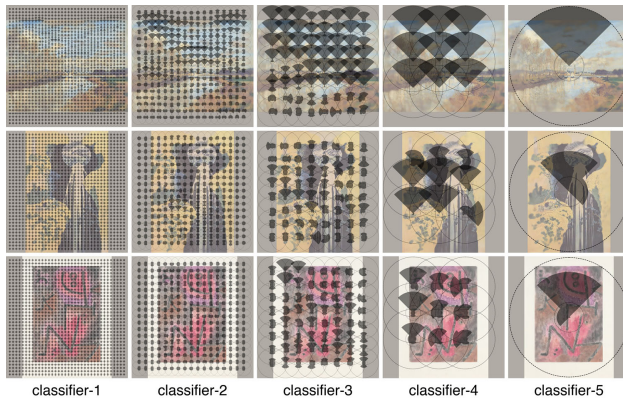


Figure 7. Predicted orientations from individual receptive field units within each classifier. Different classifiers (1–5) are plotted from left to right. Relative size of the four wedges within each circle reflects prediction strength across the four different orientations. Examples are shown for three paintings (dates given when known): *Argenteuil seen from the small arm of the Seine* by Claude Monet (1872), *The Waterfall of Amida behind the Kiso Road* by Katsushika Hokusai, *After Annealing* by Paul Klee (1940).

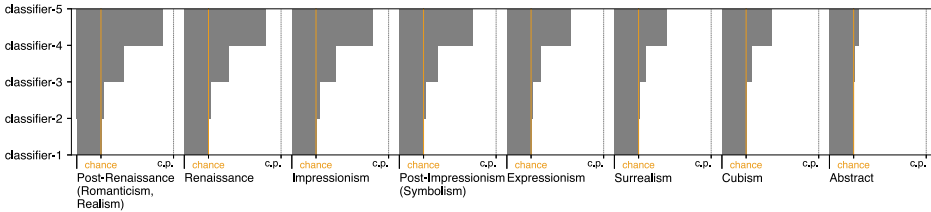


Figure 8. Redundancy between adjacent classifiers, grouped by style. This metric corresponds to rescaled cross-entropy between classifier distributions at level  $n$  and those at level  $n + 1$  (see Methods). Values are averaged across fragments. Along x axis, c.p. stands for ceiling performance.

this is achieved via representation of other task-relevant properties that do not share characteristics with those represented by classifier-5.

To summarize these results, it appears that abstract art suffers from higher local variability of compositional effects, requiring spatially extended integration of orientation cues for them to cohere into a reliable orientation estimate. Deeper layers must represent emergent global properties that are not necessarily available to previous layers; these properties may be connected with Gestalt principles associated with abstract material, for which the whole is more than the sum of its parts. It is true that we measured performance levels that are relatively low (albeit well

above chance), and that this observation alone prompts caution in potentially overstating the universality of this phenomenon; nonetheless, it also implies the existence of a mechanism that is clearly structured to a measurable extent (i.e. stands above chance). Partial, but systematic, neural integration of image features has also been described for other aesthetic judgments (Iigaya et al., 2020). To determine whether these findings are idiosyncratic to our model or, as we hope, they reflect real compositional mechanisms of more general relevance to cognition, we report on human behavioral experiments designed to retain the closest possible connection with the above characterization of the network model.

## A.1 A deep learning framework for human perception of composition

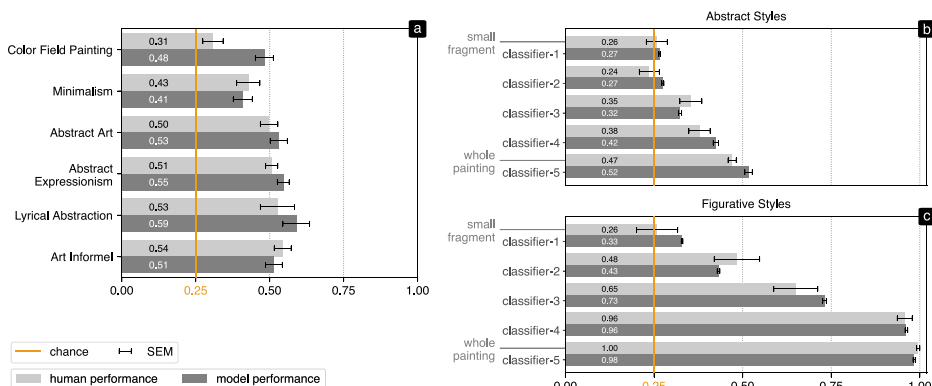


Figure 9. Human versus model performance for whole paintings and fragments. In (a), model performance from classifier-5 is plotted alongside human performance on whole paintings (dark versus light bars, respectively), grouped by style. In (b–c), model performance from different classifiers (1–5) is plotted alongside human performance on image fragments, separately for abstract (b) and figurative styles (c).

### Human experiments and comparison with network model

Among abstract styles, human observers have the most difficulty determining the orientation of images from Color Field Painting (e.g., Mark Rothko) and Minimalism (e.g., Francois Morellet), as these styles provide less pictorial content and more perfect symmetries. Figure 9a demonstrates that results from the neural network are well-aligned with the corresponding human results (except for Color Field Painting).

The model-human correspondence also exists on a per-classifier basis (Figures 9b, c). For this analysis, we compare model performance from different layers with human performance for different fragment sizes. We emphasize that values for the model are not obtained by presenting the model with fragments (as for example in Rodriguez et al. (2018)): here the model is always presented with full-size images. Different values refer to different classifiers at different depths. We can establish a one-to-one pairing between network layer and fragment size because, when selecting fragment size in the human experiments, the different sizes were tailored to the receptive-field size of different layers within the model. Other than that, there is no obvious connection between model and human results, meaning that it is not trivially expected that values obtained from different network depths should mirror those obtained from human measurements at different fragment sizes.

We find good correspondence between the two sets of results: abstract and figurative styles show the same progression of performance across different fragment/receptive-field sizes ( $r^2 = 0.976$  with  $p < 0.001$ ). One implication of this result is that, if we assume that the network model represents an acceptable approximation to the human visual pathway (Kriegeskorte, 2015; Yamins & DiCarlo, 2016), we should be able to probe activity at different levels within the pathway by simply restricting fragment size in a behavioral experiment. Although this result may seem trivial on the surface, it is not to be taken for granted when the output metric is a relatively complex perceptual judgment (see Discussion for more in-depth consideration of these issues). Further experiments using different behavioral tasks would be necessary to confirm/disprove the generality of this result.

Our proposed model is not only able to replicate the extent to which humans produce correct responses, but also specific patterns according to which humans produce incorrect responses. Figure 10 plots normalized frequency of incorrect predictions (three orientations other than the upright orientation of reference) across classifiers (for model in a and b) and fragment size (for humans in c). It is evident that, when incorrect responses are produced, there is a tendency on the part of both model and humans to select the orientation  $180^\circ$  away from the orientation of reference (painting in upside-down configuration) more often than those orthogonal to it. This anisotropic effect applies to all styles for the model (Figure 10a) and is particularly

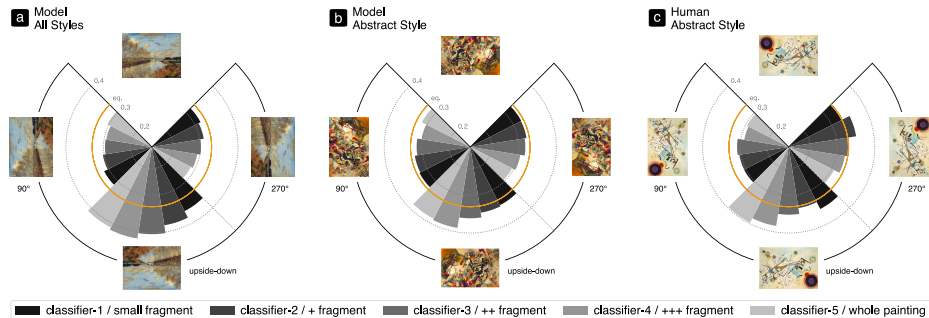


Figure 10. Normalized frequency of incorrectly predicted orientations across classifiers for model, all styles (a) and abstract style (b); across fragment size for humans, abstract styles (c). *eq.* stands for equi-frequency. Examples are shown for three paintings: *Argenteuil seen from the small arm of the Seine* by Claude Monet (1872), *Komposition VII* by Wassily Kandinsky (1913), *Komposition VIII* by Wassily Kandinsky (1923).

pronounced when analysis is restricted to abstract paintings (Figures 10b, c); for this type of art material, model and human behavior are well-correlated across classifiers and fragment sizes ( $r^2 = 0.745$  with  $p < 0.001$ ).

The correspondence between human and model behavior for incorrect responses indicates that, in both cases, some image features present horizontal/vertical compositional cues that support alignment of the image along either horizontal or vertical axes, without providing useful information for determining how the image should be mirror-flipped around the chosen axis. Consider, for example, an image containing a mountain reflected against a lake in front of the mountain; clearly, a human observer is able to orient this image so that one mountain is above, and the other one is below. However, if the observer were asked to determine which mountain should be on top and which below, he or she may be unable to make such a determination (in the assumption that the lake produces a nearly perfect reflection of the mountain above it). Similarly, if the observer were asked to determine whether the image should be flipped left-right or not, he or she may be unable to produce an informed answer. Our results indicate that cues of this kind are available from the image database we constructed, and that both model and human are able to exploit them in similar fashion. In Figure 10, the upside-down confusion also seems to be more pronounced for later/larger layers/fragments, suggesting that the horizontal/vertical position emerges as a consequence of spatially broad cue integration. On abstract material, across classifiers and fragment sizes, a Cuzick’s test (Cuzick, 1985) confirms this trend with  $p = 0.012$ .

### Human/model comparison on a per-painting basis

So far, we have considered the behavior of humans and model without referring to individual paintings. For example, when we say that model performance matches human performance for orienting abstract art, we mean that out of 100 abstract paintings, the model responds on average as correctly as the human observers. This finding does not mean that model and human responses match at the level of individual paintings: the model may be correct for 50 out of 100 paintings, and so may be the human observer, but the 50 paintings for which the model is correct may be those 50 for which the human is incorrect. To address this possibility, below we consider model versus human responses on a per-painting basis.

Figure 11 plots the density distribution of joint orientation choices generated by model and humans for individual abstract paintings. If model and humans were to agree on the orientation of every painting, modulations would only be present within the diagonal bins; all other values should be zero. Because all values must sum to 1 in each plot, we can take the sum of the diagonal values as an indication of model–human agreement (the sum is 1 when model and humans fully agree, 0 when they consistently disagree). The diagonal sum is significantly different from the null prediction only for whole paintings (Figure 11e); when data are plotted for humans orienting smaller fragments and model responses from more superficial layers (Figures 11a–d), agreement decreases to around chance. But how do we assess significance in relation to the statements above?



## A.1 A deep learning framework for human perception of composition

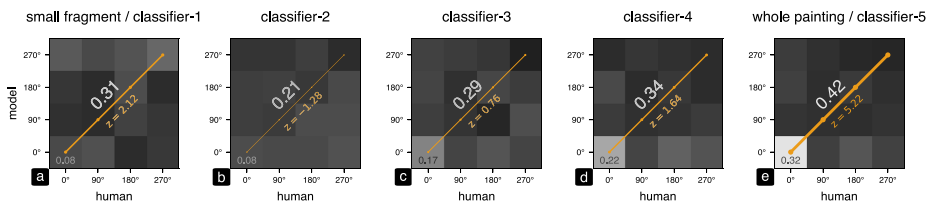


Figure 11. Density distribution of joint orientation choices generated by model and humans for individual abstract paintings, computed separately for different fragment-size/classifier from small/early (a) to large/late (e). Diagonal values correspond with matching responses (humans and model generate the same response); the diagonal sum (indicated by large white digits) is therefore termed “mutual agreement.” Its value is z-scored against the null hypothesis of human/model independence of choices (see main text for clarification). Intensity of white digits and thickness of diagonal orange line scale with corresponding z score. Bottom-left value reports agreement on target orientation.

There are at least two different ways of defining a null hypothesis against which to test significance of the agreement value. The simplest approach is to define the null hypothesis as one where both humans and network respond randomly; in this case, the expected value for each pixel in the  $4 \times 4$  surface plots of Figure 11 is simply  $1/16$ , and the expected sum across the diagonal is  $1/4$ . Although this approach may be appropriate for evaluating whether humans/models perform above chance, we find that it is inadequate for the purpose of addressing the specific issue we formulated at the beginning of this section. Consider, for example, a scenario in which humans and the model are always correct, regardless of the specific painting that is presented to them; clearly, they are also always in agreement with each other, merely as a consequence of being correct: the diagonal sum would be 1 and, when tested against the null hypothesis as outlined, it would be highly significant. We would then incorrectly conclude that humans and model behave similarly on a per-painting basis. A similar issue arises if, for example, humans and model are always incorrect by consistently reporting the upside-down orientation: again, they will be 100% in agreement, but this outcome does not carry any specificity for distinct paintings. More generally, this problem applies to any non-random pattern of responses on the part of humans/model, including less extreme versions of the scenarios outlined above; that is, ones where a given response is not certain but has an associated probability different than chance. Our goal is to define the null hypothesis in relation to this class of scenarios.

To establish a baseline level of agreement, we calculate expected agreement under the hypothesis that humans and model act independently with relation to specific paintings: on any given trial, we assume that humans produce the four possible responses with

probabilities  $\{p_{\uparrow}, p_{\rightarrow}, p_{\downarrow}, p_{\leftarrow}\}$  regardless of the specific painting that is presented, and the model produces those responses with probabilities  $\{q_{\uparrow}, q_{\rightarrow}, q_{\downarrow}, q_{\leftarrow}\}$ ; using the empirical estimates for these quantities, we calculate the expected value for their agreement  $a_0$  and its standard deviation  $\sigma_0$  on a per-painting basis. We then assess the experimentally measured agreement value  $\hat{a}$  in relation to this baseline via  $(\hat{a} - a_0)/\sigma_0$  (z-score); that is, we determine how far the observed agreement values score over and above their expected level under the hypothesis that humans and model present no per-painting association. When we apply this calculation, we find that the agreement value associated with the whole-painting/classifier-5 dataset (Figure 11e) returns a large z-score ( $>5$ ), whereas the z-scores associated with the other four datasets (Figure 11a-d) barely reject the null hypothesis of independence. We therefore conclude that, although humans and model perform similarly on average across the entire database for all fragment-size-versus-classifier comparisons (see Figure 9), their strategies may differ on a per-painting basis. More specifically, when humans have access to fragmentary information about a specific painting, and the network is restricted to early classifiers, humans adopt a decision strategy that bears little resemblance to the strategy adopted by the network. In contrast, when the whole painting is available to human observers and the network has access to classifier-5 information, their strategies present similarities that are specific to the given painting and extend to both correct and incorrect classifications.

We propose the following explanation for these results. Earlier classifiers (corresponding to smaller receptive fields) only have access to fragment-like information during training; this constraint may steer the classifiers towards discovering local statistical regularities for the purpose of identifying the overall



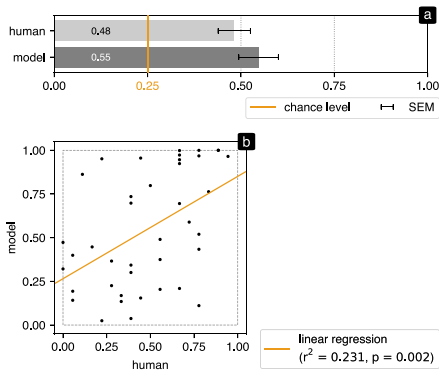


Figure 12. Comparison between our model and the results reported by Mather (2012). (a) The average human and model performance. The original article reports human mean performance per painting. This quantity is not directly comparable to top-1 accuracy of the model, because the latter does not reflect the level of uncertainty for each painting. We have therefore chosen to plot the raw prediction value for the correct orientation as the model metric to plot against human performance (b).

orientation of the painting. The resulting strategy may differ from the way in which humans approach fragments of Abstract art: the human tendency is to consider sub-parts of an abstract painting as a new complete painting, rather than as a fragment. An additional factor that may be relevant in this context is the well-documented inconsistency of aesthetic judgments across observers, especially for abstract material (Leder et al., 2016; Schepman et al., 2015; Specker et al., 2020; Vessel, 2010; Vessel et al., 2018). Although the network model does not suffer from subjective variability in the human sense, it is affected by the stochastic nature of the training protocol. Therefore, it is possible to quantify and compare internal noise between model and humans (Neri, 2010), an endeavor which we hope to pursue in future research.

We find similar results with human data collected by others. Our model is better than human observers for the selection of paintings adopted in Mather (2012) (Figure 12a), similar to the small difference we observe for our own data (Figure 9b). When we plot model-versus-human responses to individual paintings from this prior study (Figure 12b), we find a measurable trend ( $p = 0.002$ ), but the magnitude of the correlation is relatively small ( $r^2 = 0.231$ ) (see Dodge & Karam, 2017 for related results). Clearly, the detailed behavior

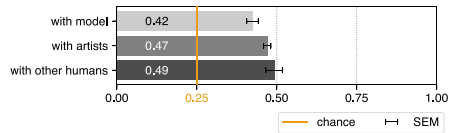


Figure 13. Painting-by-painting human agreement with network model (top), the artists who painted the images used in our study (middle), and other humans from our sample of participants (bottom). This analysis was restricted to abstract material.

of our model on a per-painting basis presents some limitations that will require further investigation.

### Human agreement with model/artist/other humans for abstract paintings

Figure 13 reports human overall agreement with the model (classifier-5), the artist (whose choice is used as correct reference above) and other humans, for judgments made on whole paintings (data from fragments is excluded from this analysis). The model–human agreement is 42%. Agreement with artists is the same as human performance (already reported in Figure 9b) and it is slightly higher at 47%. Because model performance is close to human performance, this difference in agreement is due to the discrepancy already highlighted in Figure 11. Finally, for paintings that have been evaluated multiple times by humans, we can compute mutual agreement via average agreement of all possible pairs of judgements per painting. Defined this way, inter-human agreement reaches 49%, amusingly suggesting that artists themselves may not be the most reliable reference on this task or, more likely, that some artists deliberately choose non-optimal orientations (insofar as optimality is defined with reference to the orientation considered most appropriate by an average human observer).

## Discussion

### Relations to art composition

Despite its rich history, the study of pictorial composition has been hampered by the inherent combinatorial complexity of how graphical elements interact on canvas. Our goal was to determine whether modern computational tools, in particular deep learning, may help to tackle this difficult problem and advance our understanding of abstract art composition. Engaging with a research programme of this kind

brings up an immediate problem: how do we go about quantifying the perception of art composition in humans? It goes without saying that the cognitive phenomena underlying composition, both those exploited by the creating artist and those engaged by the observing spectators, reach far beyond the remit of one scientific study. Moreover, we would like to draw a distinction between “objective” description and “metric” description: a metric description does not necessarily imply an objective description. Our goal is not to objectify the meaning or the interpretation of the composition, neither defining rightness of compositions, but to build a metric of composition based on data, that is, existing paintings from a specified limited spectrum. Objectiveness of the metric is then transparently constrained by the range of the dataset itself, rather than its fundamental correctness. We want to organize composition around measurable dimensions that are relevant to human perception, so that perceptual processes may 1) serve as a guide in the identification of important dimensions for candidate metric(s) and 2) enable quantitative measurements of how pertinent those metrics are to composition. As a consequence, we do not have a definitive answer concerning whether and how the chosen metric is connected with the notion of “objective” description. Possibly, we will never have such an answer because the very concept of objective description may not exist. This consideration has forced us to focus on a relatively simple, yet critical metric of perceptual judgment relating to art material: determining the overall orientation of the picture (Lindauer, 1969, 1987; Liu et al., 2017; Mather, 2012). We view this as a first humble step in the direction of answering the question laid out in this article, and therefore recognize that a satisfactory account of art composition will require further research. Notwithstanding the simplicity of this behavioral metric, we discuss below its merits and its connection with existing literature in vision science.

Anecdotal evidence from the art world provides some relevant points of contact with the judgment task used in this study. Upon returning home, still lost in his thoughts, Kandinsky once noted: “I suddenly saw a painting of indescribable beauty, impregnated with great inner ardor. I was at first dumbfounded, then I quickly reached this mysterious painting on which I only saw shapes and colors and whose subject was incomprehensible.” (Kandinsky, 2014). As a matter of fact, he was looking at one of his own paintings, but set out in unfamiliar orientation. A mere change of orientation in the picture was sufficient to spark a perceptual reaction that would conjure up a novel composition, serving as a cursory indication that image orientation and art composition are somehow connected, albeit in ways that we (or even the artist) may not fully understand. If we accept that this connection may be present, we must then ask whether

orientation judgments of art material are supported by perceptual mechanisms that overlap with those studied by visual psychophysics; in other words, is vision science an appropriate tool for understanding this problem at any meaningful level (Mamassian, 2008)? There is evidence to support this additional connection: portrait artists, for example, are more efficient at certain visual discrimination tasks than non-artists; however, they are equally subject to the well-known face inversion effect (Devue & Barsics, 2016), a phenomenon intimately linked with the perception of overall image orientation. This brings us to the last connecting element between art composition and vision science: if we accept that global orientation judgments are relevant to art composition, and if we accept that judgments of this kind may engage similar mechanisms to those operating in other visual skills, we then ask whether this task is also important for understanding vision in general. Existing literature provides clear answers to this question.

## Relations to existing literature in vision science

Prior studies offer numerous demonstrations of perceptual inversion effects in relation to meaningful visual material, such as faces (Valentine, 1988) or moving bodies (Chang & Troje, 2009; Neri et al., 2006, 2007). In these demonstrations, flipping the stimulus upside-down generally disrupts perceptual analysis by biological observers (human as well as non-human Vallortigara et al., 2005; Vallortigara & Regolin, 2006), even though it is not expected that this manipulation should impact an artificial system for which up and down do not necessarily carry any meaning (unless the system has learnt about gravity). The impact of stimulus inversion is characterized by a distinct developmental trajectory (Zhao et al., 2014) and has been associated with specific regions of visual cortex (Grossman & Blake, 2002). In short, at least within the context of contemporary thinking about higher-level vision, there is no doubt that stimulus orientation represents a valid topic of enquiry for understanding visual perception. More specifically, inversion effects are intimately associated with the notion of holistic processing, often summarized as “the whole is more than the sum of its parts,” a concept that has played a significant role in the study of higher-level vision (Ullman, 1996). Inversion effects have been exploited to selectively probe holistic processes in a number of applications, ranging from natural scene perception (Neri, 2014) to action processing (Taubert et al., 2011; Cusack et al., 2015).

Furthermore, and in direct connection with the present study, previous authors have argued that deep neural networks should prove useful for the study of perceptual inversion effects (VanRullen, 2017). In our

study, perhaps the most pertinent demonstration of the profitability afforded by this computational tool is the stratification of relevant effects across layers (Figure 6a); indeed, it is difficult to imagine how this type of analysis would have been possible using more conventional modeling tools. Collectively, our results indicate that abstract art, more than other styles, relies on global compositional principles that emerge deeper into the network (Figure 6b), and that may bear on the concept of holistic processing outlined above. The term “global” may not encompass overly complex cognitive phenomena, and may to some extent overlap with the notion of ‘spatially extended’ as deeper layers possess larger receptive fields. Nevertheless, we have also shown that there is no simple/naive integration of orientation cues that would explain the observed patterns in our data (Figure 8). The issue of granularity remains largely unanswered at this stage, although we do make some progress in this respect.

### Granularity and receptive field structure in human versus network architectures

By breaking paintings into fragments, our goal was to venture beyond prior studies and begin to consider composition as dynamic interaction of image subelements. As outlined, we find that local features of abstract art are integrated into a global representation that remains hidden from transparent explanation. This may, or may not, conform to artistic intuition. On the one hand, Abstract art explores pictorial composition on a level that is not bound by conventional relationships of experiential space, so it may be expected that the underlying structure should not be available at the level of simple spatial integration. On the other hand, it is often the case that Abstract art seems to be redundant across space (e.g., some applications of action painting), so that it would seem that little should be gained from incorporating more spatially extended information. Furthermore, Figurative art often presents complex spatial relationships on a large scale; indeed, natural scene perception is by no means a phenomenon that can be easily reduced to naive spatial integration of local cues (DiCarlo et al., 2012). We conclude that our demonstration of emergent global encoding at deeper layers for abstract art is not trivially expected based on either conventional ideas about art material, nor on mainstream considerations about receptive-field structure in hierarchical models. We discuss the latter issue further below.

The notion that visual cortex is organized along a hierarchical pathway of visual areas with progressively increasing receptive-field size is established (Dumoulin & Wandell, 2008; Yamins & DiCarlo, 2016); however, it

is not at all understood how information is combined from one area to the next. At this stage, we are perhaps nearing adequate characterization and computational understanding of the transition from V1 to V2 (Freeman et al., 2013), but subsequent transformations remain poorly understood. This picture is further complicated by the known presence of feedback processes (Lamme et al., 1998), which are not implemented in any form within our model. With this in mind, it is somewhat surprising that our model is able to capture some properties of human orientation judgments for isolated fragments by simply restricting its access to more superficial layers. On the face of it, this result indicates that, by designing experiments with tailored fragmented stimuli, we may be in a position to probe human perceptual mechanisms corresponding to different layers in the model and possibly different visual areas along the processing hierarchy. We contend that this result is not trivial, both in consideration of the unresolved issues associated with inter-aerial transformations outlined above, and also in light of the fact that the connection between the notion of receptive/perceptive field on the one hand, and final behavioral response on the other hand, is far from being as straightforward as is often tacitly assumed (Neri & Levi, 2006; Spillmann, 1971). In humans, we cannot simply read out of earlier visual areas using experimental tools; what we can perhaps do is force observers to rely on signals from those earlier areas for the production of behavior (which we can measure). That we may achieve this by tailoring fragment size is not trivially expected, particularly in relation to a behavioral judgment that is not explicitly connected with global integration and that involves higher-level cognitive processes. We do not know whether the same result would be obtained for other perceptual judgments, an issue we hope to address in future research.

Notwithstanding the correspondence between human observers and model responses as discussed, we do find conspicuous differences between the behavior exhibited by the network and that measured from humans. Interestingly, those differences become particularly evident when we consider fragments, less so with whole paintings (Figure 11). We propose that this result should be interpreted in light of the considerations discussed above. As we have already noted, our model is purely feed-forward, that is, its architecture fails to incorporate important recurrent computations that are known to operate in cortex. It is conceivable that related perceptual processes are engaged by humans in our task, possibly contributing to the discrepancy we observe with respect to the model (see also Doerig et al., 2020 for related considerations). Furthermore, the nature of the discrepancy may be specific to the task/protocol we selected for this study and/or to the resolution of our measurements. We do

not have definite answers to these and other related questions, some of which we have highlighted in this Discussion. At this stage, we view our contribution as a starting point for more in-depth studies of art composition adopting a similar framework, namely the integrated application of deep learning models, data-driven extraction of regularities and psychophysical validation in human observers. Our results demonstrate that this approach is feasible and capable of generating non-trivial insights and predictions into the mechanisms underlying art composition in humans.

*Keywords:* machine learning, psychophysics, receptive field, pictorial composition, inversion effect

### Acknowledgments

Supported by grants ANR-16-CE28-0016, ANR-17-EURE-0017, ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL from Agence Nationale de la Recherche.

Commercial relationships: none.

Corresponding author: Pierre Lelièvre.

Email: contact@plelievre.com.

Address: Laboratoire des systèmes perceptifs, Département d'études cognitives Science Arts Création Recherche (EA 7410), Paris, France.



### References



- Arnheim, R. (2004). *Art and Visual Perception – A Psychology of the Creative Eye (2nd edition, 50th Anniversary)*. Berkeley: University of California Press. (Original work published 1954).
- Chang, D. H., & Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision*, 9(1), 1–17.
- Cusack, J. P., Williams, J. H., & Neri, P. (2015). Action perception is intact in autism spectrum disorder. *Journal of Neuroscience*, 35(5), 1849–1857.
- Cuzick, J. (1985). A wilcoxon-type test for trend. *Statistics in Medicine*, 4(1), 87–90, <https://doi.org/10.1002/sim.4780040112>.
- Devue, C., & Barsics, C. (2016). Outlining face processing skills of portrait artists: Perceptual experience with faces predicts performance. *Vision Research*, 127, 92–103.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434.
- Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*, 1–7, <https://doi.org/10.1109/ICCCN.2017.8038465>.
- Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research*, 167, 39–45.
- Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage*, 39, 647–660.
- Elgammal, A. M., Mazzone, M., Liu, B., Kim, D., & Elhoseiny, M. (2018). *The shape of art history in the eyes of the machine*. ArXiv:1801.07729 [Cs, AI], <http://arxiv.org/abs/1801.07729>.
- Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7), 974–981.
- Gaspar, C. M., Bennett, P. J., & Sekuler, A. B. (2008). The effects of face inversion and contrast-reversal on efficiency and internal noise. *Vision Research*, 48, 1084–1095.
- Gong, M., Xuan, Y., Smart, L. J., & Olzak, L. A. (2018). The extraction of natural scene gist in visual crowding. *Scientific Report*, 8(1), 14073.
- Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, 35(6), 1167–1175.
- Igaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2020). Aesthetic preference for art emerges from a weighted integration over hierarchically structured visual features in the brain. *BioRxiv* 2020.02.09.940353, <https://doi.org/10.1101/2020.02.09.940353>.
- Kandinsky, W. (1989). *Du spirituel dans l'art, et dans la peinture en particulier* (P. Sers, N. Debrand, & B. Du Crest, Trans.). Denoël: Gallimard. (Original work published 1912).
- Kandinsky, W. (2014). *Regards sur le passé: Et autres textes, 1912-1922* (J.-P. Bouillon, Ed.). Hermann. (Original work published 1974).
- Kandinsky, W. (1991). *Point et ligne sur plan: Contribution à l'analyse des éléments picturaux* (P. Sers, Ed.; S. Leppien & J. Leppien, Trans.). Gallimard. (Original work published 1926).
- Kelley, T. A., Chun, M. M., & Chua, K. P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision*, 3(1), 1–5.
- Klee, J. (1961). *Notebooks, Volume 1: The thinking eye* (J. Spiller, Ed.; R. Manheim, C. Weidler, & J. Wittenborn, Trans.). Lund Humphries.

- Klee, P. (1973). *Notebooks, Volume 2: The nature of nature* (J. Spiller, Ed.; H. Norden & J. Wittenborn, Trans.). Lund Humphries.
- Klee, P. (1998). *Théorie de l'art moderne* (P.-H. Gonthier, Ed. & Trans.). Denoël: Gallimard. (Original work published 1924)
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.
- Krizhevsky, A., Sutskever, I., Hinton, E., & G. (2012). ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems, 25*, <https://doi.org/10.1145/3065386>.
- Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology, 8*(4), 529–535.
- Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and shared taste in art and face appreciation. *Frontiers in Human Neuroscience, 10*, 155, <https://doi.org/10.3389/fnhum.2016.00155>.
- Lindauer, M. S. (1969). The orientation of form in abstract art. *Proceedings of the Annual Convention of the American Psychological Association, 4*(1), 475–476.
- Lindauer, M. S. (1987). Perceived and preferred orientations of abstract art. *Empirical Studies of the Arts, 5*(1), 47–58, <https://doi.org/10.2190/K1X2-X4VJ-6YN9-BKD8>.
- Liu, J., Dong, W., Zhang, X., & Jiang, Z. (2017). Orientation judgment for abstract paintings. *Multimedia tools and applications, 76*(1), 1017–1036, <https://doi.org/10.1007/s11042-015-3104-5>.
- Locher, P. J., Stappers, P. J., & Overbeeke, K. (1999). An empirical evaluation of the visual rightness theory of pictorial composition. *Acta Psychologica, 103*(3), 261–280, [https://doi.org/10.1016/S0001-6918\(99\)00044-X](https://doi.org/10.1016/S0001-6918(99)00044-X).
- Mamassian, P. (2008). Ambiguities and conventions in the perception of visual art. *Vision Research, 48*(20), 2143–2153.
- Mather, G. (2012). Aesthetic judgement of orientation in modern art. *i-Perception, 3*(1), 18–24, <https://doi.org/10.1068/i0447aap>.
- McManus, I. C., Cheema, B., & Stoker, J. (1993). The aesthetics of composition: A study of Mondrian. *Empirical Studies of the Arts, 11*(2), 83–94, <https://doi.org/10.2190/HXR4-VU9A-P5D9-BPQQ>.
- Neri, P. (2010). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review, 17*, 802–808.
- Neri, P. (2014). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience, 34*, 2374–2388.
- Neri, P., & Levi, D. M. (2006). Receptive versus perceptible fields from the reverse-correlation viewpoint. *Vision Research, 46*, 2465–2474.
- Neri, P., Luu, J. Y., & Levi, D. M. (2006). Meaningful interactions can enhance visual discrimination of human agents. *Nature Neuroscience, 9*, 1186–1192.
- Neri, P., Luu, J. Y., & Levi, D. M. (2007). Sensitivity to biological motion drops by approximately 1/2 log-unit with inversion, and is unaffected by amblyopia. *Vision Research, 47*, 1209–1214.
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc, <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- Rodriguez, C. S., Lech, M., & Pirogova, E. (2018). Classification of style in fine-art paintings using transfer learning and weighted image patches. *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS), 1–7*, <https://doi.org/10.1109/ICSPCS.2018.8631731>.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., . . . Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. arXiv: 1409.0575 [cs]. <http://arxiv.org/abs/1409.0575>.
- Schepman, A., Rodway, P., Pullen, S. J., & Kirkham, J. (2015). Shared liking and association valence for representational art but not abstract art. *Journal of Vision, 15*(5), 11, <https://doi.org/10.1167/15.5.11>.
- Schwabe, K., Menzel, C., Mullin, C., Wagemans, J., & Redies, C. (2018). Gist perception of image composition in abstract artworks. *i-Perception, 9*, 204166951878079, <https://doi.org/10.1177/2041669518780797>.
- Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

- Simonyan, K., & Zisserman, A. (2014, September 4). Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556 [cs], <http://arxiv.org/abs/1409.1556>.
- Specker, E., Forster, M., Brinkmann, H., Boddy, J., Immelmann, B., Goller, J., ... Leder, H. (2020). Warm, lively, rough? Assessing agreement on aesthetic effects of artworks (R. T. H. Ho, Ed.). *PLoS One*, *15*(5), e0232083, <https://doi.org/10.1371/journal.pone.0232083>.
- Spillmann, L. (1971). Foveal perceptive fields in the human visual system measured with simultaneous contrast in grids and bars. *Pflugers Archiv (European Journal of Physiology)*, *326*, 281–299.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015, April 13). Striving for simplicity: The all convolutional net. arXiv: 1412.6806 [cs], <http://arxiv.org/abs/1412.6806>.
- Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research*, *51*(11), 1273–1278.
- Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT Press.
- Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*(Pt 4), 471–491.
- Vallortigara, G., & Regolin, L. (2006). Gravity bias in the interpretation of biological motion by inexperienced chicks. *Current Biology*, *16*, R279–280.
- Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns. *PLoS Biology*, *3*(7), e208.
- VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142.
- Vessel, E. A. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision*, *10*(2), 1–14, <https://doi.org/10.1167/10.2.18>.
- Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, *179*, 121–131, <https://doi.org/10.1016/j.cognition.2018.06.009>.
- WikiArt. (n.d.). WikiArt.org - Visual Art Encyclopedia, <https://www.wikiart.org/>.
- Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.
- Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*, 2256–2262.
- Zhao, J., Wang, L., Wang, Y., Weng, X., Li, S., & Jiang, Y. (2014). Developmental tuning of reflexive attentional effect to biological motion cues. *Scientific Report*, *4*, 5558.

# A deep-learning framework for human perception of abstract art composition

**Pierre Lelièvre** Laboratoire des systèmes perceptifs, Département d'études cognitives  
 Science Arts Création Recherche (EA 7410)    
 École normale supérieure, PSL University, CNRS, 75005 Paris, France

**Peter Neri** Laboratoire des systèmes perceptifs, Département d'études cognitives  
 École normale supérieure, PSL University, CNRS, 75005 Paris, France  

## Supplementary Material

### Genre and style categories used for model performance

For the purpose of legibility and ease of exposition, we grouped WikiArt genres and styles into new categories that retained consistency with the original classification as closely as possible. Categories are described below in Table.S1 and Table.S2

Genre Painting (0.24)	Genre Painting (0.58), Religious Painting (0.28), History Painting (0.04), Interior (0.03), Allegorical Painting (0.03), Literary Painting (0.03), Battle Painting (0.01)
Landscape (0.23)	Landscape (0.69), Cityscape (0.24), Marina (0.07), Cloudscape (0.01), Panorama (~0)
Portrait (0.19)	Portrait (0.92), Self-portrait (0.08)
Abstract (0.13)	Abstract
Figurative (0.12)	Figurative (0.40), Still Life (0.31), Animal Painting (0.14), Flower Painting (0.14)
Sketches (0.07)	Sketch and Study (0.63), Design (0.37)
Nude (0.03)	Nude painting (Nu)

Table S1: Genre groups used for model performance and corresponding WikiArt genres. Group frequency within the database and genre frequency within the assigned group are indicated in brackets.

### Potential role of portrait/landscape aspect-ratio bias

Portrait/landscape aspect-ratio distribution may represent a source of bias in the database, for example by artificially increasing the chance level away from 0.25. As a way of illustration, consider the extreme scenario in which all paintings in our database come in portrait format. Under this scenario, observers may ignore the content of the painting, and simply re-orient the frame to be in portrait configuration on every trial. This strategy would correspond to a chance level of 0.5. To examine the applicability of this scenario (or related ones) within the context of our database, Fig.S1 plots aspect-ratio distribution across the entire database (a) and separately for different genres (b-e), together with the corresponding model performance (orange bars). Aspect ratio is defined as  $\log_2\left(\frac{width}{height}\right)$ : positive values correspond to landscape configuration, negative values to portrait configuration.

As expected, the Portrait genre contains more vertical paintings (Fig.S1b) and the Landscape genre contains more horizontal ones (Fig.S1c). Across all genres (Fig.S1a), we find a slight preponderance of portrait paintings. Because model performance is globally higher for the Portrait genre (Fig.4a), it is legitimate to ask whether this small bias in aspect-ratio may account for the difference in performance. Under this hypothesis, we expect performance for Figurative and Abstract genres (which are well balanced, see Fig.S1d-e) to be higher than performance for the Landscape genre, but this is clearly not the case for both positive and negative aspect log-ratios (orange bars in Fig.S1d-e are shorter than orange bars in Fig.S1c on either side of each plot).

## A.1 A deep learning framework for human perception of composition

Lelièvre & Neri

2

Post-Renaissance (Romanticism, Realism) (0.36)	Realism (0.35), Romanticism (0.28), Baroque (0.15), Neoclassicism (0.08), Rococo (0.07), Academicism (0.03), Orientalism (0.01), Luminism (0.01), Tenebrism (0.01), Classicism (0.01), Biedermeier (~0), Neo-Rococo (~0), Costumbrismo (~0)
Abstract (0.15)	Abstract Expressionism (0.25), Abstract Art (0.12), Art Informel (0.11), Color Field Painting (0.09), Minimalism (0.08), Lyrical Abstraction (0.06), Op Art (0.05), Concretism (0.04), Hard Edge Painting (0.04), Tachisme (0.03), Symbiotic Art (0.02), Post-Painterly Abstraction (0.02), Nouveau Réalisme (0.01), Spatialism (0.01), Neoplasticism (0.01), Suprematism (0.01), Action painting (0.01), P&D (Pattern and Decoration) (0.01), Post-Minimalism (0.01), Neo-Geo (0.01), Neo-Minimalism (0.01), Maximalism (~0), New Casualism (~0), Light and Space (~0), Native Art (~0), Neo-Concretism (~0), Automatic Painting (~0), Indian Space painting (~0), Neo-Orthodoxism (~0), Perceptism (~0), Synchromism (~0), Excessivism (~0)
Impressionism (0.14)	Impressionism
Post-Impressionism (Symbolism) (0.12)	Post-Impressionism (0.58), Symbolism (0.36), Pointillism (0.04), Cloisonnism (0.01), Synthetism (~0)
Expressionism (0.09)	Expressionism
Renaissance (0.07)	Northern Renaissance (0.38), Mannerism (Late Renaissance) (0.20), High Renaissance (0.19), Early Renaissance (0.19), Proto Renaissance (0.04), Renaissance (~0)
Surrealism (0.06)	Surrealism
Cubism (0.02)	Cubism

Table S2: Style groups used for model performance and corresponding WikiArt styles. Group frequency within the database and style frequency within the assigned group are indicated in brackets.

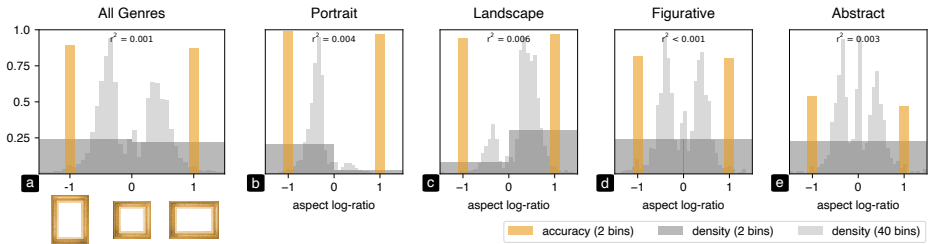


Figure S1: Aspect-ratio imbalance does not explain model performance. Left/right orange bars show model performance on whole paintings averaged across negative/positive (portrait/landscape) aspect log-ratio  $\log_2(\frac{\text{width}}{\text{height}})$ , computed across all genres (a) and separately for 4 genres of specific interest (b-e). Gray histograms plot aspect log-ratio density across paintings for two sampling resolutions: 40 bins (light gray) and 2 bins (dark gray).  $r^2$  correlation values are computed between model performance and corresponding aspect log-ratio on a painting-by-painting basis.

In addition to the above observation, we can exclude a role for aspect-ratio bias more quantitatively by computing the correlation value ( $r^2$ ) between aspect log-ratio and model performance on a painting-by-painting basis; this analysis returns tiny values (0.001 across all genres and no larger than 0.01 for each genre). More specifically, the  $r^2$  value of 0.003 is insignificant for our specific class of interest (abstract paintings), rendering the issue of tangential relevance to the main conclusions of this study.

To further exclude a role for aspect ratio, we can compare performance on whole paintings (Fig.4) with performance on the central fragment accessible from classifier-4. This square fragment is essentially a cropped version of the painting with no aspect-ratio bias. As expected, overall performance is lower than corresponding performance on whole paintings; consider for example performance for the abstract genre at 0.43 in Fig.S2a (central fragments) versus 0.51 in Fig.4a (whole paintings). Notwithstanding this expected drop in model performance, Fig.S2 and Fig.4 demonstrate exactly the same ranking of genres/styles, leading us to conclude that aspect-ratio bias in our database plays no role in relation to our main results.



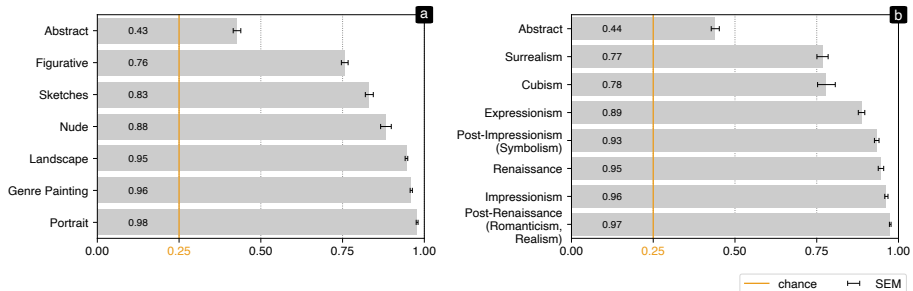


Figure S2: Model performance of classifier-4 on its central fragment for paintings grouped by genre (a) and style (b). Because the central fragment is square in shape, these results are free of any potential aspect-ratio bias.

### Model performance with fixed parameters for convolutional blocks

We attempted transfer learning from a pre-trained model (PyTorch implementation) by freezing the convolutional blocks. Performance for this protocol is substantially reduced (Fig.S3) compared to the one reported in Fig.4b. Notwithstanding this overall drop in classification performance, the overall trend across styles is largely preserved. The only measurable exception applies to Cubism and Surrealism, for which performance was similar under the transfer-learned model.

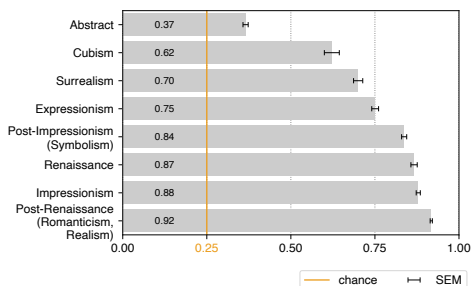


Figure S3: Model performance on whole paintings with frozen convolutional blocks, grouped by style.

### Human performance with respect to art knowledge, age and reaction time

Previous studies (Lindauer, 1969) have demonstrated that humans can perform the orientation task regardless of their degree of familiarity with painting material. During the registration process of our web-based experiment, participants were asked to rate their general knowledge of art material as none, little, medium or significant. We do not report any substantial difference in performance across these four categories (Fig.S4a).

Fig.S4b plots performance for different age groups. Within the 15-67 y.o. range (adults and young adults), we do not observe measurable differences. Prior work (Arnheim, 1954/2004) has reported some differences between adults and children, but our dataset does not support adequate assessment of the age range that is relevant to those observations.

## A.1 A deep learning framework for human perception of composition

Lelièvre & Neri

4

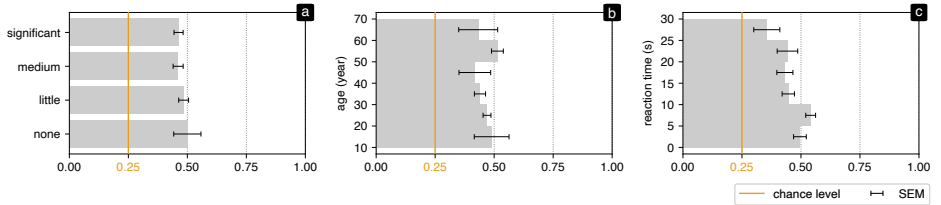


Figure S4: Human performance with respect to art knowledge (a), age (b) and reaction time (c).

We report an inverse relationship between performance and reaction time (Meadmore et al., 2014) (Fig.S4c). It appears that orientation judgement is optimal when produced within 5 seconds of studying the image; decisions associated with longer inspection times approach chance level. To understand which factors may be implicated in determining this relationship between performance and time-to-decision, we plot reaction time (RT) distributions separately for abstract and figurative style (Fig.S5a). The mode of the distribution is clearly shorter for figurative paintings (by  $>3$  seconds), suggesting that lower performance at longer RT's simply reflects overwhelming representation of ambiguous image material (primarily abstract) within this performance/RT range.

Although the above interpretation appears reasonable, we report one piece of evidence that is not trivially compatible with it: RT scales with fragment size, being shorter for smaller fragments and longer for larger fragments (Fig.S5b), a puzzling result given that small fragments are typically ambiguous in their content. One possible interpretation for this result is that reaction time is dependent on the number of available orientation cues needed to produce an overall orientation judgement, rather than ambiguity of content in a semantic sense as suggested by the figurative-vs-abstract analysis from the previous paragraph. Clearly, the latter factor does play a role when orienting figurative material with highly recognizable content, but it is not the sole factor at play. This factor set aside, the results in Fig.S5b suggest that abstract art requires more complex and spatially extended integration of orientation cues to produce a good judgement of overall orientation, in line with our findings (see Results).

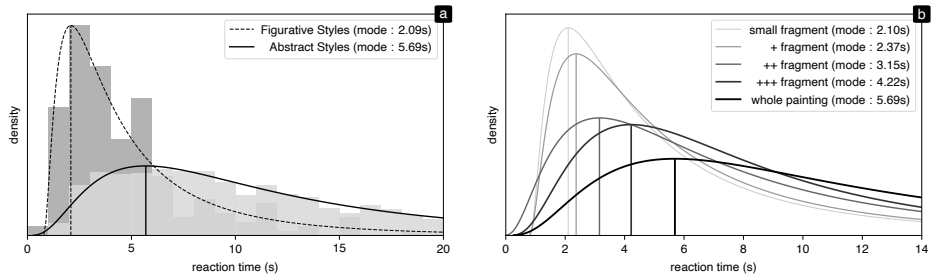


Figure S5: Reaction time density grouped by style (figurative versus abstract in a) and fragment size of abstract paintings (b). Smooth curves show log-normal best-fits, with vertical line indicating mode of distribution. Only fits are shown in (b) for legibility.

### References

- Arnheim, R. (2004). *Art and Visual Perception – A Psychology of the Creative Eye* (2nd edition, 50th Anniversary). University of California Press. (Original work published 1954)
- Lindauer, M. S. (1969). The orientation of form in abstract art. *Proceedings of the Annual Convention of the American Psychological Association*, 4(1), 475–476.
- Meadmore, K. L., Liversedge, S. P., Wenger, M. J., & Donnelly, N. (2014). Exploring the relationship between response time, sensitivity and bias in categorical and coordinate visuospatial processes: Evidence for hemispheric specialisation. *Journal of Cognitive Psychology*, 26(4), 423–432. <https://doi.org/10.1080/20445911.2014.903255>

## A.2 ART@VSAC 2019

ART@VSAC was a curated exhibition highlighting established and emerging artists. It took place in 2019 at BAC Atelier in Leuven, Belgium, and my artworks have been awarded the audience prize.

This exhibition was part of VSAC, the Visual Science of Art Conference, which initially spawned as a satellite event from ECVF. Beside the art exhibition, this conference was also accepting scientific contributions in the form of talks, posters, and joint poster/artwork discussions. However, my goal was to experience the setup of a real exhibition, and to meet the public. It was also a chance to convince myself of being an artist. Therefore, I applied for the main exhibition, without any direct presentation of my scientific work.

For this occasion, I prepared five A3 artworks. They are reproduced in Fig.A.3, Fig.A.4, Fig.A.5, Fig.A.6 and Fig.A.7. Achieved in one month, this work was challenging as I had to process the most recent third of my personnel dataset of compositions, and to finish many algorithmic developments. The extremely positive feedbacks from visitors made me more confident about my PhD project.

Nonetheless, even if the public was somehow a *selected* audience, familiar with digital approaches to art, as soon as I spoke about AI in the presence of my drawing machine (see photos in Fig.A.1), there was a certain confusion and invisibilization of my *true* creative process. In people's view, computer art is necessarily generative, and the moving machine embodies its creative hand and mind. I ultimately believe that technological aspects generally prevent people from looking at graphical proposals themselves. Autonomous machines seem to grab the core attention, and it is difficult to desacralize this fascination. This observation really made me wonder about the necessity to show the pen plotter besides artworks. Even if producing postcards in live, that spectators can keep, is an efficient manner to question the material value of the presented work, there is way to improve the communication about the conceptual work.

To get more insights about presented artworks, please have a look at the exhibition catalogue reproduced in Fig.A.2.

Appendices



Figure A.1: Photographs of the exhibition venue.

**PIERRE LELIÈVRE IS A FRENCH ARTIST AND RESEARCHER, BASED IN PARIS. HIS WORK FOCUSES ON PICTORIAL COMPOSITION AND IS DRIVEN BY AN ITERATIVE PROCESS, MIXING TRADITIONAL DRAWING AND ADVANCED ALGORITHMS.**

**AFTER STUDYING CINEMATOGRAPHY AT THE ENS LOUIS LUMIÈRE, HE BEGAN HIS CAREER AS A R&D ENGINEER FOR CINEMA AND VIDEO GAMES, PUSHING BOUNDARIES OF DIGITAL HUMAN PHOTOREALISM. MEANWHILE, HE STARTED TO DRAW AND COLLECT SMALL COMPOSITIONAL STRUCTURES THAT FLOAT WITHOUT FRAMES. THIS PERSONAL DATABASE PROGRESSIVELY REACHED 5000 ELEMENTS AND BECAME THE STARTING POINT, AS WELL AS THE CORE MATERIAL OF HIS REFLECTION AND ARTISTIC EXPLORATIONS.**

**AS THESE PICTORIAL ELEMENTS INCREASED IN COMPLEXITY, THE INTERNAL HARMONY PRINCIPLES, INTUITIVELY SUPPORTED BY A RECURRENCE OF GESTURES AND SPECIFIC COMBINATIONS, HAVE REMAINED VEILED, LOST BETWEEN ARBITRARY RULES AND TOTAL RANDOMNESS. THE LACK OF DESCRIPTORS AND MEASURING TOOLS TO UNDERSTAND THE BASIC MECHANISMS OF THE COMPOSITION PERCEPTION, FINALLY CRYSTALLIZED IN HIS PHD PROJECT.**

**SINCE LATE 2018, HIS RESEARCH HAS BEEN FUNDED BY SACRE (SCIENCES, ARTS, CRÉATION, RECHERCHE) AND HOSTED BY THE ENS (ÉCOLE NORMALE SUPÉRIEURE) IN PARIS. THE KEY ASPECT OF THIS PROGRAM IS TO PROVIDE AN INTERDISCIPLINARY SPACE WHICH ENCOURAGES RESEARCH BY CREATION. IN THIS ECOSYSTEM, THE ARTISTIC APPROACH OF PIERRE LELIÈVRE IS EVOLVING AS A SHORTCUT BETWEEN THEORY AND PRACTICE, FROM MODELING TO VALIDATION, ENABLING QUICKER ITERATIONS AND DEEPER REFLEXIVE ANALYSIS**

**VSAC CONSTITUTES A PREMIERE FOR HIS RESEARCH::CREATION ARTWORKS IN AN EXHIBITION CONTEXT.**

Pierre Lelièvre's mediums encompass ink on paper, algorithms and machine learning without clear frontiers. His artistic process starts from instinctive drawings, made among notes in an erratic manner. Scanned, cleaned and vectorized, they become data, feeding statistical analysis and innovative computations. In order to send back the meaning of these numerical values to the paper, with the same ink as the original drawings, a mechanical pen plotter completes the creative cycle. The presented artworks are the results of this approach.

Accumulation expresses the puzzling and fascinating contradictory feelings produced by unordered data. Drawings become patterns, forgetting they could have their own Individuality.

Morphology addresses the issue of drawings' digital representation and human perception. From a finished artwork, strokes definition and order are lost, merged into one new entity. Do algorithms and human eyes share strategies to disentangle contours and structural skeletons?

Combinations exposes a hierarchical accumulation of different versions derived from the same initial drawing. The result, produced by multiple cycles through the artist's creative loop with the pen plotter, emphasizes composition complexification and combinatory effects of graphical elements.

Cartography is an experimental map of similarities produced by an artificial neural network. The selected architecture can automatically discover and compress important features in data. Every drawing can therefore be encoded with only few dimensions. This representation is the most efficient one learned by the algorithm, but do we agree? Could imagining the machine's perception improve the understanding of our own?

Figure A.2: Extract from the exhibition catalogue.



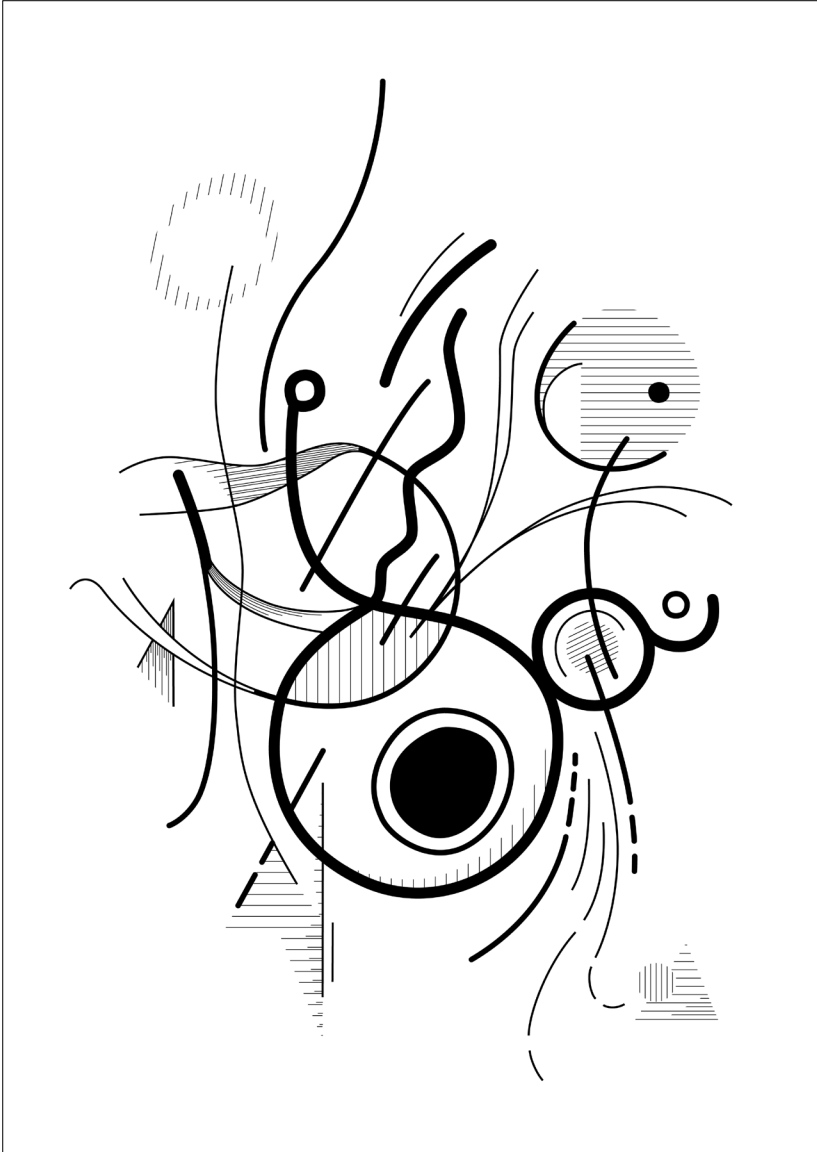


Figure A.4: Individuality



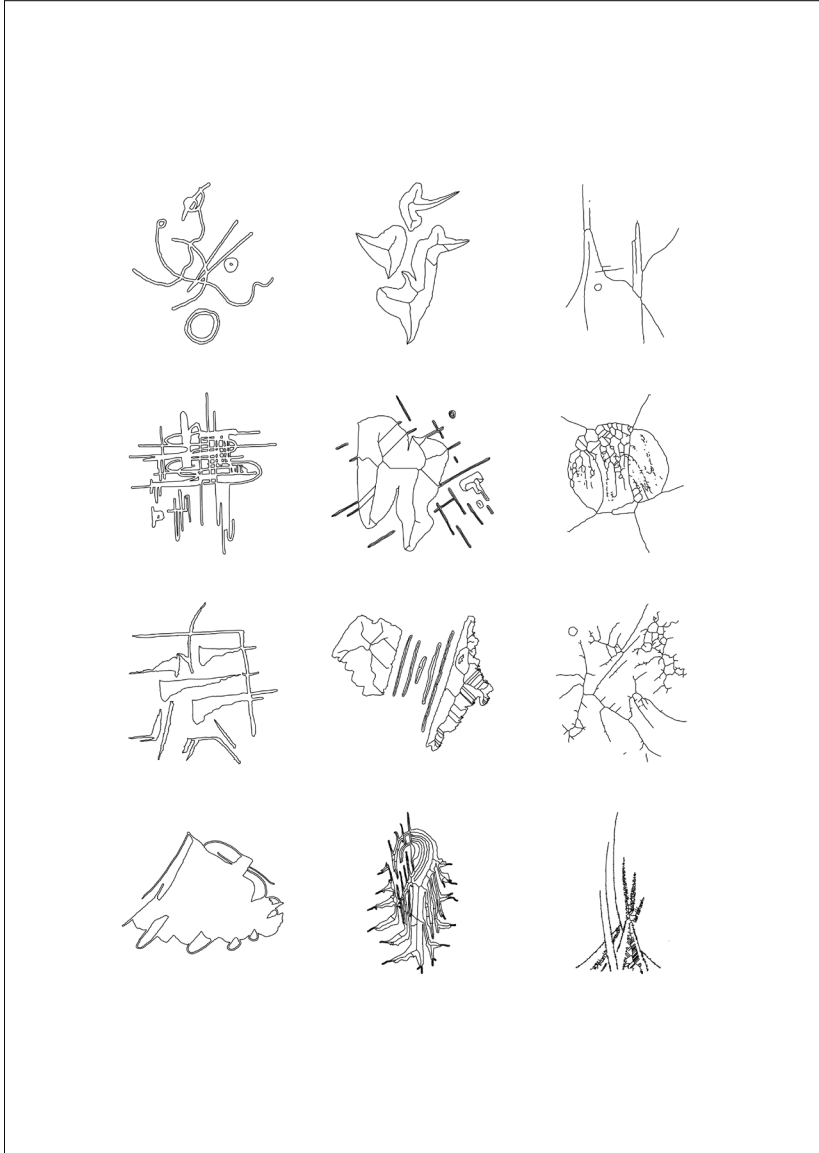


Figure A.5: Morphology

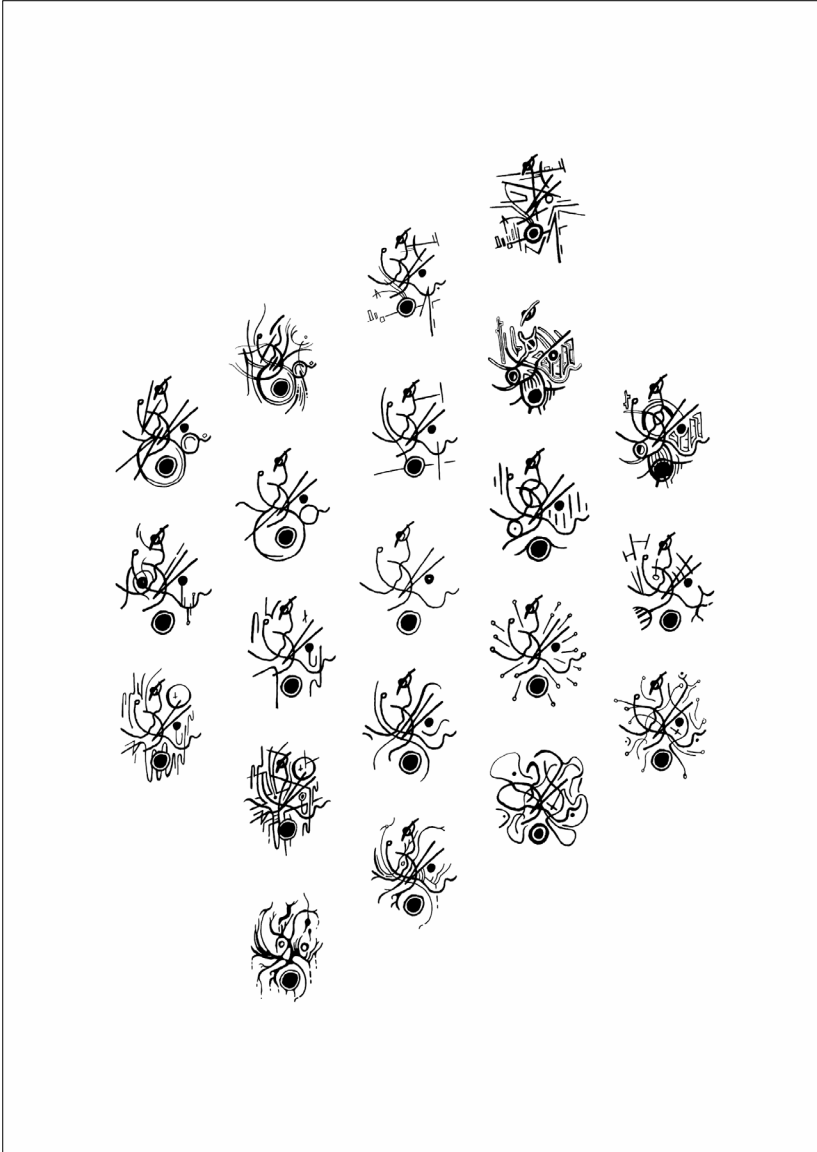


Figure A.6: Combinations

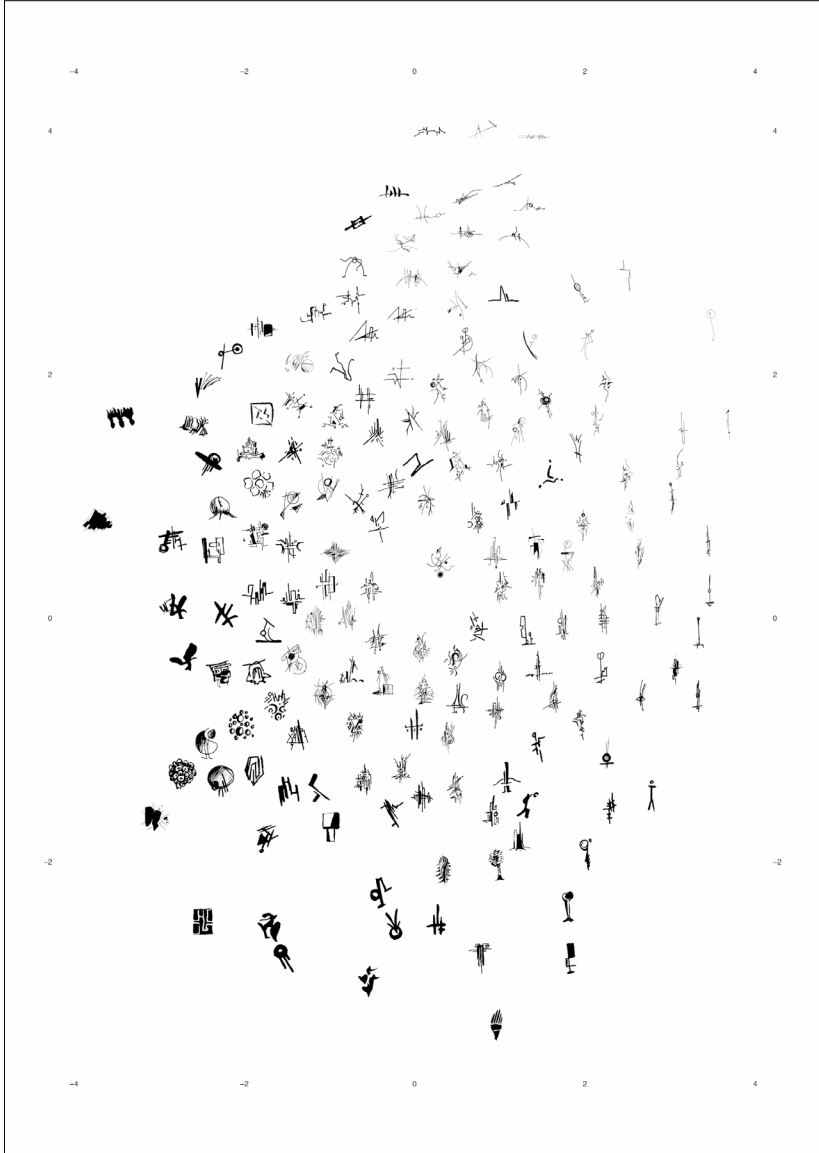


Figure A.7: Cartography

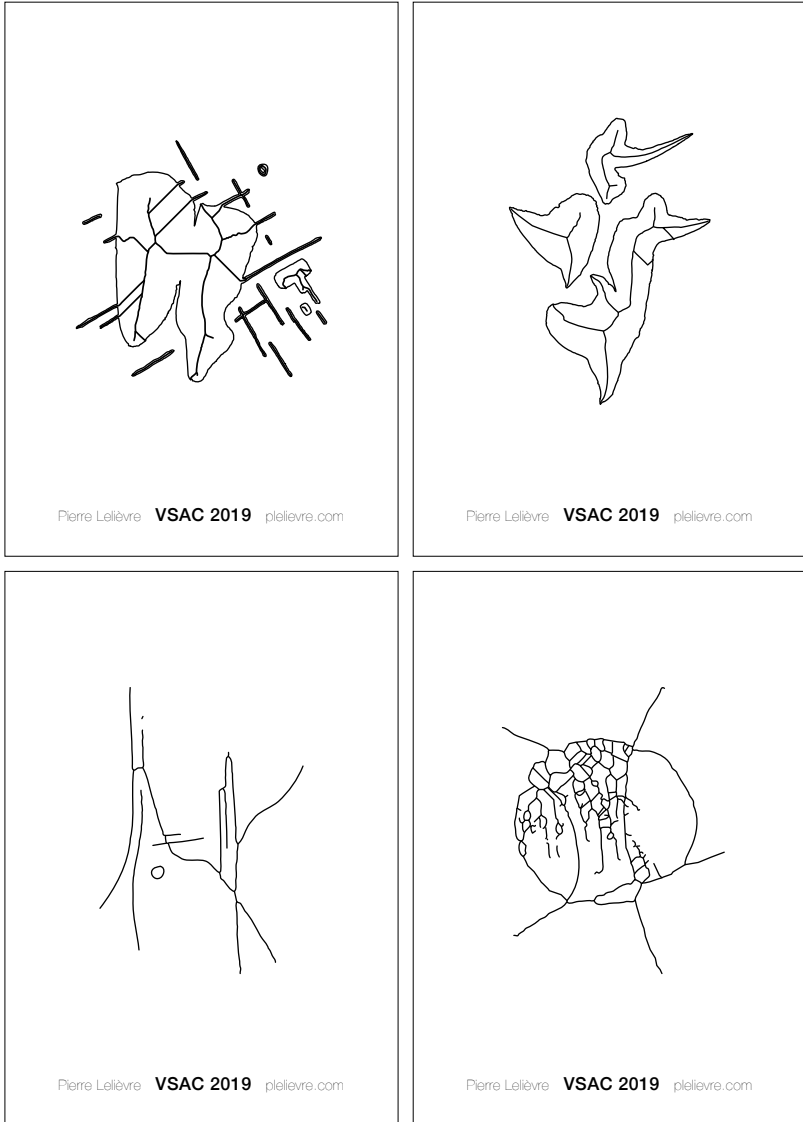


Figure A.8: Postcards that were plotted and offered to the public during the exhibition.



## A.3 Defense exhibition

The defense exhibition, called *Hidden Dynamics* took place in the Bibliothèque Math-Info<sup>4</sup> of the École Normale Supérieure in Paris. It was open to the public from November 21st 2022 to December 2nd 2022<sup>5</sup>, however it has been mainly designed to be an integrated part of the PhD defense. The idea was to propose a sensitive approach of my artistic material prior to the scientific summary of my research. Therefore, jury members had the opportunity to visit the exhibition right before my presentation and the principal discussion.

The presented artworks are plotted with roller pens on technical papers of format  $60 \times 80$ cm, and are essentially based on propositions detailed in Section.6.3. They are divided into two related series, showing variations around the same five principles. Fig.A.9 and Fig.A.10 show in order of presentation: *Trace Circle*, *Trace Line*, *Transition*, *Uncertainty*, and *Surrounding*<sup>6</sup>. Globally, the second series intends to express more complicated and intense dynamics than the first set.

In the exhibition venue, the two series were arranged on each outer side of the main stairs of the library lobby. Photographs of Fig.A.11, Fig.A.12, Fig.A.13 and Fig.A.14 illustrate this scenography. The intention was to invite viewers to focus on details and use their visual memory to investigate dynamical variations, instead of relying on a direct comparative viewpoint on alternatives.

Finally, Fig.A.15 reproduces the poster of the exhibition<sup>7</sup>.

---

<sup>4</sup>The library dedicated to mathematics and computer sciences.

<sup>5</sup>Officially until November 25th 2022.

<sup>6</sup>More details on these principles are available on my website: <https://plelievre.com/projects/phd-exhibition/>

<sup>7</sup>Graphic design by Yunya Hung (<https://yunyahung.com>).

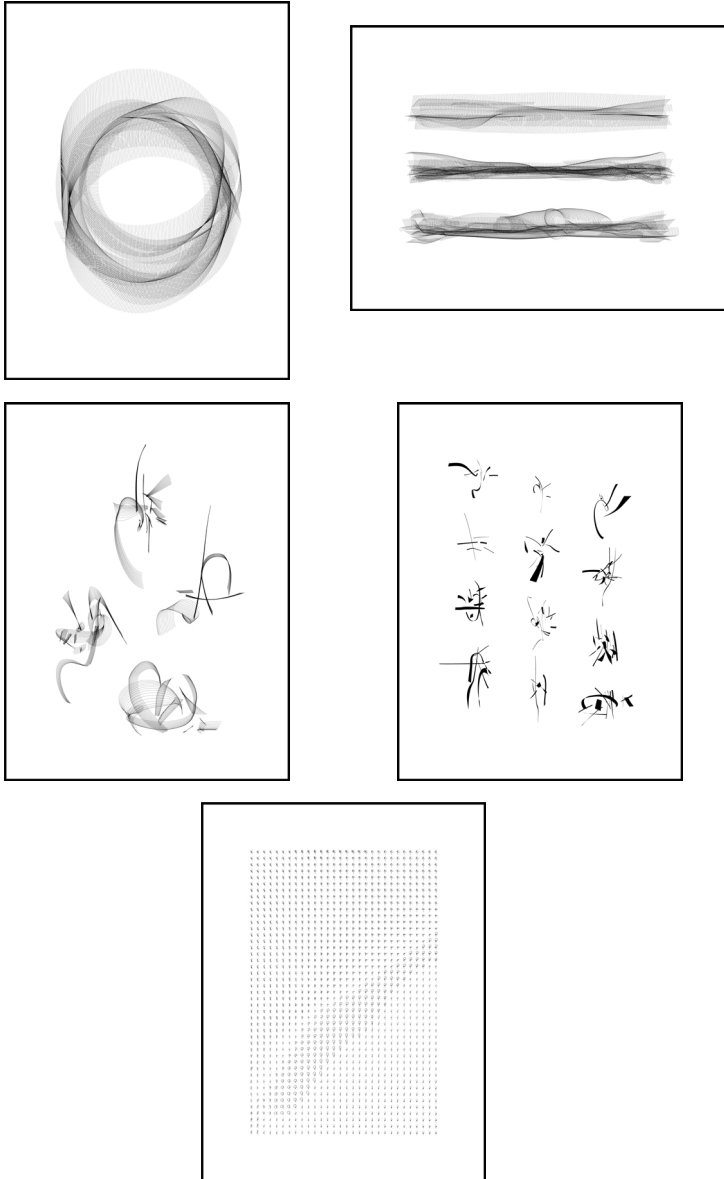


Figure A.9: Digital renderings of series I. In order of presentation: *Trace Circle I*, *Trace Line I*, *Transition I*, *Uncertainty I* and *Surrounding I*.

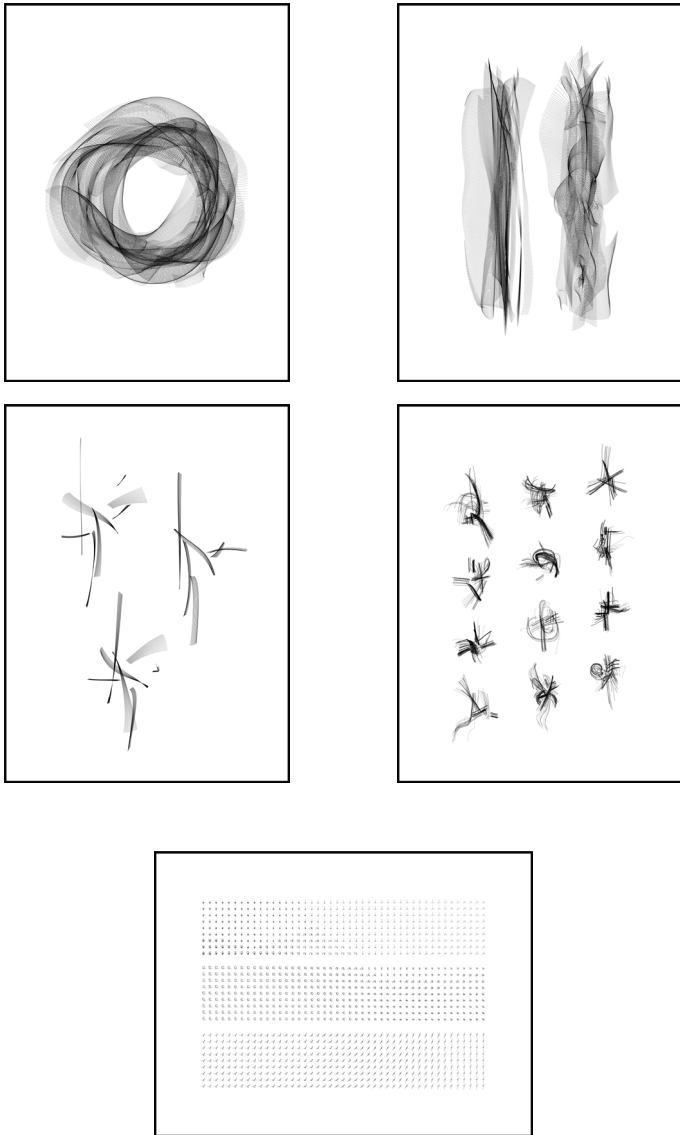


Figure A.10: Digital renderings of series II. In order of presentation: *Trace Circle II*, *Trace Line II*, *Transition II*, *Uncertainty II* and *Surrounding II*.





Figure A.11: Photographs of the exhibition.



Figure A.12: Photographs of the exhibition.



Figure A.13: Photographs of the exhibition.



Figure A.14: Photograph of the exhibition.



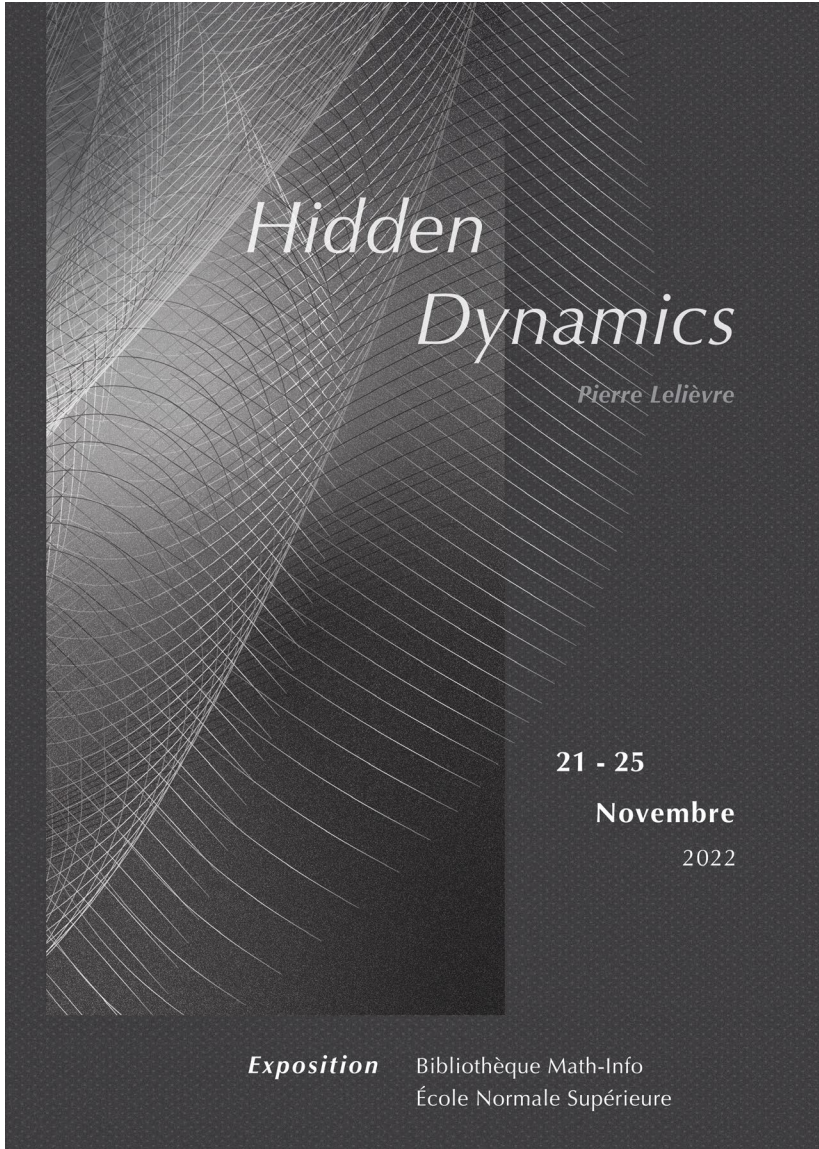


Figure A.15: Exhibition poster (graphic design by Yunya Hung).

## A.4 Long résumé en français

La pratique du dessin est originellement exploratoire. Un enfant dessine sans but, pour le pur plaisir de remplir la feuille de papier. Sa volonté de représenter des personnes, des animaux ou des objets familiers n'émerge que plus tard. Pour ma part, il y avait initialement quelque chose de ce même plaisir graphique primordial, issu d'un assemblage intuitif et non-conscient de lignes simples. Il convient également de noter que mon activité de dessin a bien souvent accompagné des moments de contrainte physique. Je ne dirais pas *d'ennui*, mais bien d'un certain besoin d'être concentré, sans pouvoir me déplacer au-delà des quelques centimètres carrés d'une feuille de note, lors d'un cours, d'une réunion, d'une conférence ou d'un appel téléphonique. J'imagine que l'attention portée sur l'écoute a permis dans les premiers temps à la boucle visuomotrice de s'affranchir d'un objectif figuratif. J'étais alors probablement plus libre d'explorer différentes mécaniques de construction abstraite (Fig.0.2). Par la suite, la logique géométrique a progressivement laissé place à une sensibilité plus fine pour les lignes et à des arrangements plus délicats (Fig.0.3).

Ce n'est peut-être qu'au cours des cinq dernières années que ces structures sont véritablement devenues compositionnelles (Fig.0.4). Je me sens maintenant capable d'articuler des éléments graphiques d'une plus grande diversité et de juger des combinaisons d'une plus haute complexité. La pratique a également progressivement glissé vers des moments de création dédiés. Ce qui n'a pas changé en revanche, c'est la taille physique de ces propositions. Je continue à dessiner de petites figures, approximativement contenues dans un cercle de 4cm de diamètre. Je crois que ce qui me fascine à cette échelle – en m'approchant si près de la feuille de papier – c'est de mieux discerner le contact de l'encre et du papier. Autour de moi, le monde se fait plus lointain et j'ai l'impression d'assister à la vie d'un microcosme – l'œil un peu démiurge. C'est aussi une méthode féconde pour se concentrer sur des phénomènes structurels élémentaires. La taille du stylo ou du pinceau, dans un si petit espace, devient un facteur limitant *ad hoc* du niveau de détail pouvant être exploré. En effet, je considère ma collection de compositions comme des graines à peine germées d'hypothétiques œuvres plus complètes. C'est d'ailleurs l'une des raisons pour lesquelles j'ai peut-être commencé à collecter tous ces fragments, dont le compte dépasse maintenant les 5000 unités.

Ma pratique artistique est donc initialement plus introspective que véritablement orientée vers les autres. Si la pratique de la composition est un acte de connaissance en soi, alors témoigner de l'évolution de ce processus, de ses questions associées et de ses découvertes est au moins aussi important que la présentation du corpus lui-même. Parce que les interrogations posées par les dessins finaux ne sont qu'implicites, je pense que les visiteurs sont plus susceptibles de saisir mon point de vue en lisant mes publications et en réalisant leurs propres compositions inspirées des principes exposés dans ces écrits, plutôt qu'en observant mes dessins. Les

## Appendices

œuvres d'art sont des suggestions. Je suis toujours heureux de montrer mon travail et de le partager avec d'autres, mais je crois que, dans ces circonstances, le plaisir primaire dépasse la théorie. Être exposé m'apparaît comme un objectif secondaire de ma trajectoire artistique et je conçois principalement la pratique artistique comme un moteur de questionnements personnels.

Mon interrogation initiale fut donc ancrée dans une nécessité pratique ; celle d'expliquer le besoin indescriptible de poser telle ligne à tel endroit plutôt qu'un autre ; celle de comprendre ce qui impose tel élément graphique à la place d'autres possibles d'une infinie variété. Comme la résolution des situations semble à la fois *évidente* et sans logique objective, j'ai commencé à ressentir une réelle tension dans ma pratique et une certaine angoisse à l'idée de composer au hasard. Je voulais donc comprendre les lois dirigeant ma propre pratique et m'assurer d'être cohérent avec moi-même.

Le véritable catalyseur d'une reformulation scientifique de ces questions personnelles a été ma rencontre avec des écrits théoriques d'artistes, en particulier *Théorie de l'art moderne*<sup>8</sup> de Paul Klee et *Du spirituel dans l'art*<sup>9</sup> de Wassily Kandinsky. Leurs travaux faisaient écho à des sensations et des intuitions perçues au cours de ma propre pratique de la composition. Ils ont également ouvert mon imaginaire sur une démarche scientifique dans le domaine de l'art, détachée d'une pure vision historique. Leur pensée est abondamment représentée au fil des pages de ce manuscrit. Malgré des recherches élargies, ce corpus original reste pour moi le plus pertinent en ce qui concerne le paradigme compositionnel que j'ai été amené à développer. Idéalement, j'aimerais pouvoir m'inscrire dans cette famille d'artistes dont la recherche a été principalement auto-réflexive, tout en s'efforçant de partager une méthode expérimentale visant une connaissance objective et inclusive de l'art.

Pour ma part, je m'intéresse désormais plus précisément à recouvrer et comprendre la dimension dynamique des formes picturales. Cette dimension se loge de manière implicite dans les contraintes entre les différents éléments graphiques du plan, ainsi que dans les régularités qui unissent l'ensemble des compositions d'un même artiste. J'ai l'intuition d'une continuité picturale qui fait de chaque dessin un artefact approximatif d'un tout plus complet et cohérent. Les structures compositionnelles appartiennent à un seul et même système, un objet hyper-compositionnel. Ce point de vue est motivé par le désir fondamental de voir les formes picturales comme des formes vivantes, comme des occurrences d'un même organisme complexe. Cet objet hautement dimensionnel est bien sûr difficile à visualiser dans notre esprit et à représenter explicitement comme un objet explorable. Au moyen de simulations informatiques, je cherche donc à soulever, même légèrement, le voile sur ces dimensions qui nous sont encore masquées. Indirectement, cela implique de pouvoir mesurer et quantifier les régularités compositionnelles, puis possiblement

---

<sup>8</sup>Klee, 1924/1998.

<sup>9</sup>Kandinsky, 1912/1989.

de les mettre à disposition des artistes pour qu'ils puissent faire des choix plus conscients (s'ils le souhaitent évidemment). Concrètement, je rêve de lignes en mouvement, animées par leurs forces internes, et d'outils pour voyager à travers l'espace continu des compositions.

### Méthode

La méthodologie analytique traditionnelle est essentiellement réductionniste. L'idée est de décomposer un problème en autant de petits morceaux que nécessaire. Preuve après preuve, le scientifique remonte alors la chaîne causale et en synthétise les découvertes. Par nature, la composition est un objet avec une délimitation floue, le nexus de considérations perceptives et esthétiques. Il s'agit également d'un concept couvrant à la fois une pratique artistique et un état d'organisation spatiale macroscopique de sous-éléments, dont chacun semble essentiel *a priori*. Dès lors, comment segmenter la composition sans la dénaturer ? Découper un problème en mauvais sous-éléments peut rapidement en augmenter la difficulté<sup>10</sup>. Nous utilisons donc une approche scientifique plus globale et projective : la modélisation. Cette méthode consiste principalement en la construction d'un outil de simulation d'un phénomène complexe, sans nécessairement le diviser en sous-éléments.

La composition est un aspect essentiel de l'art pictural. Cependant, chaque époque et courant artistique a su élaborer un ensemble de méthodes et de critères d'appréciation différents pour en résoudre les problèmes sous-jacents. Cela rend intrinsèquement difficile la réalisation d'une modélisation de la composition suffisamment universelle pour en couvrir toutes les manifestations. Est-ce nécessaire par ailleurs ? L'art, comme la modélisation, sont des pratiques qui nécessitent de prendre position sur le réel et le matériau étudié. À l'exigence d'universalité, se substitue alors un devoir de clarté sur la démarche adoptée. Cette exigence est concrétisée par la définition d'un paradigme compositionnel. La modélisation s'articule ainsi parfaitement avec la recherche-crédation, pour qui la transparence de la méthode est le seul moyen de légitimer le général logé dans le particulier d'une pratique individuelle.

Dans le contexte de la composition, la difficulté pour l'artiste n'est pas tant de se libérer de la nature subjective de sa propre vision, mais plutôt d'avoir le recul nécessaire pour en extraire les régularités. Ce problème complexe se pose d'ailleurs pour le spectateur avec la même difficulté, et ce alors qu'il possède un regard tout à fait extérieur. Ce phénomène s'estompe peut-être seulement lorsque l'artiste et le visiteur ont vécu à des époques différentes. Nous faisons donc le choix de requérir l'aide d'un tiers possédant une capacité d'analyse automatique, i.e. l'apprentissage automatique et en particulier l'apprentissage profond qui s'appuie sur des réseaux de neurones artificiels. D'une part, l'apprentissage profond peut être considéré

---

<sup>10</sup>Le Moigne, 1977/2006, p. 34.



comme un outil statistique froidement objectif, car il est capable d'extraire des régularités dans des données de manière complètement non supervisée. D'autre part, cette approche s'inspire souvent de l'architecture du cerveau biologique, ce qui la rend particulièrement pertinente pour la composition picturale.

Cependant, *objectif* ne doit pas s'entendre ici comme *universel*. Le champ d'application qui en résulte est initialement et essentiellement celui relatif à l'ensemble de données étudié. Par ailleurs, les régularités observées n'ont peut-être aucun rapport avec la réalité perceptive humaine. À cette fin, les sciences cognitives deviennent essentielles et particulièrement la psychophysique. Cette approche consiste précisément à mener une analyse quantitative entre un stimulus physique, réel ou simulé, et sa perception. Pour déterminer la plage de validité d'un modèle donné et son potentiel de généralisation, ces facteurs doivent être évalués au moyen d'une procédure expérimentale conduite avec des participants humains. Alors que certains chercheurs tentent de modéliser explicitement les mécanismes cérébraux, notre projet s'est développé sous des contraintes plus souples. Notre modèle est principalement conçu comme un outil de simulation génératif, servant de base à des expériences sur la perception, et un instrument de mesure statistique de la composition.

Notre programme de recherche est structuré autour des éléments suivants : création active, paradigme compositionnel, implémentation computationnelle et vérification expérimentale. L'horizon d'application de ce travail se situe à l'intersection entre l'héritage d'artistes qui ont fait de la pratique de la composition le sujet central de leur œuvre, et de nécessités issues d'une pratique personnelle. Ce projet se retrouve alors à la croisée de nombreux domaines : recherche-crédation, traitement d'images, apprentissage automatique et psychophysique. La nature interdisciplinaire de ce programme implique nécessairement des difficultés de rédaction lorsque chaque spécialité use d'un vocabulaire et d'outils qui lui sont propres. Cela demande un équilibre difficile entre explication et détails. J'espère que le compromis adopté dans ce manuscrit est adapté à l'intérêt de chaque public.

Malgré une certaine proximité de contenu, ce projet ne s'inscrit pas dans la théorie esthétique ou l'histoire de l'art. Même lorsque qu'il est fait référence à des mouvements artistiques marquants ou que je cite des essais d'artistes, mon objectif n'est pas de retracer l'histoire de la composition. Je ne possède ni les outils nécessaires, ni les connaissances requises pour une telle entreprise. Par ailleurs, je voudrais éloigner la méthodologie proposée du domaine de l'esthétique empirique, et plus précisément des approches quantitatives utilisant des jugements esthétiques directs, tels que ceux portant sur la *beauté*. Ces approches conduisent à la découverte de préférences extrêmement générales (pour la symétrie plutôt que l'asymétrie ou pour les courbes plutôt que pour les angles). Ces tendances *moyennes* masquent parfois de grandes variabilités interpersonnelles. Je préfère donc aborder l'esthétique depuis les artefacts artistiques eux-mêmes. Les jugements humains peuvent porter sur un matériau artistique, mais nous devons nous limiter

à des jugements ayant un objectif clair, tel que la *similarité*. Toute *préférence* doit être évaluée avec un but précis. Par exemple, dans notre étude sur l'orientation des peintures abstraites, les participants étaient invités à déterminer l'orientation optimale de compositions abstraites. Ainsi, il ne s'agissait pas d'une préférence esthétique en général, mais d'une préférence restreinte à une situation non ambiguë (voir Annexe.A.1).

La structure générale de ce manuscrit est articulée autour de deux parties principales. Nous nous concentrons premièrement sur le développement d'un modèle de la composition (Partie.I). Le Chapitre.1 détaille le paradigme compositionnel adopté, dans le but de clarifier notre position sur notre objet d'étude et d'en exposer le champ d'application. Le Chapitre.2 décrit les étapes de traitement nécessaires à la numérisation de mon dataset personnel de compositions, ainsi que la représentation choisie – les spécifications structurelles – adaptée à un usage algorithmique. Avant d'obtenir un outil fonctionnel, la dernière étape nécessaire est une implémentation efficace avec des réseaux de neurones artificiels. Le Chapitre.3 détaille ainsi les différents choix architecturaux intervenus dans le développement du réseau et les nombreuses stratégies d'optimisation qui ont été nécessaires pour rendre l'entraînement du modèle réalisable.

La deuxième partie de cette thèse est consacrée à l'exploration (Partie.II). Le Chapitre.4 établit un inventaire des résultats bruts et des caractéristiques offertes par le modèle, en particulier celles disponibles pour des mesures quantitatives. Le Chapitre.5 se concentre sur la vérification de la continuité de l'objet hypercompositionnel instancié par le modèle principal. Dans ce chapitre, nous présentons les méthodes et les résultats des expériences de psychophysique déployées en ligne pour quantifier l'échelle perceptive des similitudes locales le long d'interpolations dans ce nouvel objet. Le dernier chapitre (6) est consacré aux recherches d'ordre artistique avec notre modèle de la composition. J'y explore les façons de révéler les dimensions compositionnelles cachées et la manière de rendre sensible leur caractère fondamentalement dynamique. J'y fais également le récit d'un retour à l'espace matériel, lieu de rencontre unique de l'encre et du papier.

Malgré tout, ce programme de recherche ambitieux n'en est encore qu'à ses débuts. Au fil des chapitres, nous n'avons eu l'occasion d'introduire qu'une première itération d'un processus complexe – le façonnage de petites briques nécessaires à la construction d'un résultat préliminaire. Il y a eu tellement d'étapes inévitables sur le chemin de la sécurisation d'un outil fonctionnel, que celles-ci ont monopolisé tous mes efforts. On aurait pu s'attendre à des conclusions plus pratiques sur la perception de la composition, ou à des discussions plus avancées sur des questions artistiques, mais je suis déjà fier du travail accompli. L'ensemble de l'approche a été validée et son utilité a été démontrée, au point de mener de véritables expériences avec des participants humains et la réalisation d'œuvres pour une exposition. En définitive, la réalisation de ces objectifs, ainsi que la large palette des contributions – idées théoriques, découvertes scientifiques, outils

quantitatifs, logiciels et propositions artistiques – ont de loin dépassé mes attentes initiales.

### Contributions

Chronologiquement, notre première contribution se situe dans le domaine perceptif avec un article scientifique sur l'orientation des peintures abstraites<sup>11</sup> (voir Annexe.A.1). Il s'agit d'un projet de moins grande envergure, conçu comme une preuve de concept pour la méthode principale introduite dans cette thèse, i.e. une modélisation par apprentissage profond, combinée à des vérifications perceptives chez l'humain. Notre modèle d'apprentissage profond, caractérisé par l'ajout de classificateurs insérés après chaque bloc convolutif d'un VGG pré-entraîné<sup>12</sup>, a été conçu pour étudier la perception de l'orientation à différentes profondeurs du réseau de neurones. Dans l'article, nous démontrons que le modèle capture plusieurs caractéristiques humaines de la perception de l'orientation pour différentes granularités et pour de multiples styles artistiques. Il apparaît également que l'art abstrait, plus que les autres mouvements, nécessite une intégration spatialement étendue des indices d'orientation pour que ceux-ci soient intégrés et fournissent une estimation fiable de l'orientation. Nous avons aussi testé des stimuli fragmentés dans nos expériences chez l'humain, et nous avons constaté que les opérations détaillées des mécanismes perceptifs ne sont pas complètement identiques à leurs homologues artificiels pour les petits fragments, correspondant aux couches superficielles.

En ce qui concerne le cœur de ce manuscrit, la principale contribution théorique est la définition d'un paradigme compositionnel (voir Chapitre.1). Nous considérons cette étape comme essentielle pour toute approche de modélisation sérieuse, i.e. quand le modélisateur tente de comprendre quelque chose de fondamental sur le phénomène modélisé. Dans le domaine de l'apprentissage automatique, la nature des entrées  $x$  n'a pas vraiment d'importance tant que le modèle répond à ses exigences pratiques. La polyvalence des architectures d'apprentissage profond et leur puissance associée, fournissent des espaces génératifs produisant des artefacts semblables à de l'art, sans se demander explicitement si ces capacités ont un sens au regard du cadre conceptuel plus large entourant les données utilisées (par exemple vis-à-vis de l'histoire collective de la peinture occidentale). Il est en fait possible de construire un espace latent ou une représentation de n'importe quelle source graphique, mais la possibilité d'atteindre cet objectif ne garantit pas en soi la pertinence des résultats finaux en ce qui concerne le matériau pictural en question, ni l'approche initiale de l'artiste. Dans notre cas, nous voulions répondre à la question fondamentale de savoir si un espace de compositions, sous la forme d'un objet hyper-compositionnel, était artistiquement pertinent. Nous

---

<sup>11</sup>Lelièvre and Neri, 2021.

<sup>12</sup>Le VGG est un modèle visuel standard, utilisé dans la classification des images en fonction de leur contenu.

avons démontré que cet objectif est possible, à la condition d'accepter que chaque œuvre d'art soit un moyen d'enregistrement de régularités compositionnelles, que la pratique artistique en question soit sérielle et axée sur l'interaction dynamique des éléments graphiques, et que traits et compositions soient le résultat d'un processus morphogénétique, définissant intrinsèquement un espace continu de possibilités. Notre approche n'est pas encore entièrement consolidée, mais elle constitue déjà un point d'ancrage fort pour mon travail artistique.

Une contribution apparentée consiste à considérer la pratique compositionnelle comme l'organisation d'un système (voir Sous-section.1.2. *System complexity and system organization*). Cette idée est intéressante parce qu'elle fournit un objectif d'optimisation riche et clair à la disposition des éléments graphiques sur le plan. La composition devient alors un *entre-deux* créatif, où l'artiste crée intuitivement des contraintes conditionnelles entre des formes qui ne sont ni trop faibles, ni trop fortes. Ainsi, nous attribuons à notre approche probabiliste une réalité plus profonde que la détermination d'un art *optimal* ou de la *plus belle* œuvre. Nous présentons à la fois l'espace des compositions et celui du plan compositionnel comme des espaces probabilistes intégrant une infinie diversité et une richesse d'alternatives (voir Section.1.3). Nous pensons que ce cadre théorique propose une vision attractive des algorithmes d'apprentissage automatique et de leurs objectifs d'optimisation associés. Nous nous sommes d'ailleurs efforcé de décrire les implications de la dimensionnalité élevée des espaces créés, que ce soit vis-à-vis de la crainte potentielle d'une *normalisation* de l'art, ou des méprises liées aux maxima probabilistes.

Même si, à l'heure actuelle, je n'ai pas l'intention de mettre mon dataset personnel de compositions à la disposition du public, je considère cette ressource fondamentale comme une contribution à part entière. Elle représente une énorme quantité de travail manuel, de sa création à son traitement dans un dataset. Nous avons d'ailleurs développé nos propres logiciels pour en faciliter les opérations. En ce qui concerne le pipeline de traitement, nous avons principalement mobilisé des bibliothèques de traitement d'images existantes et réimplémenté certains algorithmes, tels que la routine de vectorisation par courbes de Bézier cubiques. Cette dernière intègre d'ailleurs plusieurs petites contributions, donnant par exemple plus de contrôle sur le désenchevêtrement des lignes aux intersections, et la simplification des éléments courbes (voir Section.2.2). Enfin, je suis particulièrement fier du bloc algorithmique qui mélange aléatoirement les éléments graphiques au sein d'une composition. Celui-ci tient compte d'un arbre hiérarchique des éléments et présente la possibilité de limiter le nombre de permutations (voir Sous-section.2.3. *Composition permutations*).

Par rapport aux modèles d'apprentissage profond précédents, appliqués à de petits dessins figuratifs, notre travail apporte plusieurs innovations. La première est fondamentale et concerne le paramétrage des traits. Notre approche va au-delà d'un simple encodage par séquences de segments de lignes et reste ainsi au plus

proche des intentions motrices de l'artiste. Ce détail offre également plus de flexibilité et de précision dans l'approximation des courbes d'origine (voir Section.2.2.Parametric curves). Deuxièmement, nous proposons des modèles de trait et de composition hiérarchiquement imbriqués (voir Section.3.2 et Section.3.3). La principale motivation derrière ce choix méthodologique est de capturer et de capitaliser sur la différence de nature temporelle entre ces deux types de séquences. Les traits sont des séquences ordonnées de composantes de courbes de Bézier, tandis que les compositions peuvent être le résultat de séries de traits sans ordre particulier, voire être des séries incomplètes. Nous essayons en fait de projeter chaque famille de compositions partiellement définies vers un emplacement unique dans l'espace latent. Cela fournit ainsi une puissance d'encodage plus riche. Enfin, nous introduisons un modèle du plan compositionnel dédié à la caractérisation des contraintes conditionnelles associées à chaque élément graphique. Pour émettre ses prédictions, ce modèle s'appuie sur les deux modèles précédemment indiqués, imbriqués et pré-entraînés (voir Section.3.4).

En ce qui concerne la composition, il n'existe pas de mesure évidente de l'efficacité des modèles. Nous considérons donc notre proposition de procédure d'entraînement et les métriques de surveillance associées comme des outils facilitant grandement la conception architecturale des modèles et la sélection des hyperparamètres (voir Section.4.1). Cependant, nos contributions dans le domaine des réseaux de neurones et de l'apprentissage de représentation se situe dans les nombreuses astuces d'entraînement mises en œuvre, et que nous avons tenté de compiler de manière cohérente et complète (voir Section.3.5). Par exemple, nous avons introduit des contraintes adaptatives sur  $D_{KL}(q(\mathbf{z} | \mathbf{x}) \parallel p(\mathbf{z}))$  en ajoutant une non-linéarité par dimension. Nous avons également implémenté une procédure rééquilibrant les ressources du modèle de manière optimale afin de surmonter partiellement les statistiques inégales du dataset (e.g. la longueur des séquences d'entrée). Une autre contribution significative est caractérisée par la limitation imposée à la variance de sortie du modèle, par l'intermédiaire d'une nouvelle unité appelée *BackwardClamp*. Collectivement, toutes ces innovations techniques permettent d'adresser les petits datasets témoignant d'une (trop) grande diversité. Elles soutiennent la construction d'une représentation continue et expressive tout en fonctionnant en dimensionnalité réduite. La dimensionnalité choisie est en effet plus petite de plusieurs dizaines de fois que les travaux précédents. Cette contrainte a été adoptée principalement pour faciliter la manipulation et l'interprétation des paramètres du modèle, et en vue de conduire des expériences sur la perception.

Une autre contribution importante concerne les mesures compositionnelles offertes par les différents modèles (voir Section.4.3). Pour le modèle de la composition, ces mesures sont élaborées autour des distributions de l'encodeur et du décodeur. Elles spécifient non seulement des positions dans l'espace latent ou sur le plan, mais aussi le degré d'incertitude associé, individuellement pour chaque dimension latente et chaque trait. Grâce au modèle dédié au plan compositionnel, nous avons même accès aux probabilités conditionnelles des éléments graphiques. Cela prend

la forme d'un champ de probabilité complexe (position et forme) des prochains traits, sachant les éléments graphiques existants et une position cible dans l'espace des compositions. Dans le temps contraint de cette thèse, nous avons dû limiter notre exploration de ses mesures compositionnelles à la présentation de méthodes de visualisation et de quelques pistes quant à leur usage. Cependant, nous sommes encouragés à poursuivre cette ligne de recherche par la réussite de notre première tentative, i.e. la prédiction des échelles perceptives à partir de l'information de Fisher calculées sur ces mesures compositionnelles (voir Sous-section.5.3.Perceptual scale prediction from Fisher information).

Nous avons par la suite vérifié certains aspects qualitatifs de l'espace latent compositionnel. Nous souhaitions particulièrement savoir si le modèle avait capturé des régularités en alignement avec des caractéristiques structurelles perceptives humaines. La principale difficulté de cette ambition provient de la dimensionnalité de l'espace latent. Bien que relativement faible, 16 dimensions, celle-ci ne se prête pas à une exploration expérimentale exhaustive (voir Section.5.1). Après avoir fait le point sur cette limitation, nous avons cherché des approches alternatives nous permettant de résoudre la question, au moins indirectement. Dans la communauté de l'apprentissage automatique, il a souvent été rapporté (mais jamais quantifié de manière adéquate) qu'un inconvénient qualitatif des modèles génératifs est l'existence de régions de faible densité dans leur espace latent. Ces régions se rencontrent particulièrement lors d'interpolations. Il est donc raisonnable de supposer qu'une densité homogène dans l'espace latent est essentielle à la qualité perceptive des interpolations. Nous avons effectué des expériences de caractérisation de l'échelle perceptive s'appuyant sur des jugements de similarité avec une variante du protocole MLDS par triplet (voir Section.5.3). Nous avons alors constaté des distorsions inattendues, indiquant potentiellement des divergences entre la représentation du modèle et la représentation humaine. Néanmoins, nous avons pu prédire certaines altérations non triviales de l'échelle perceptive avec l'aide des mesures compositionnelles auxiliaires fournies par le modèle (décrites ci-dessus). Ainsi, malgré des défauts dans sa représentation, le modèle est en mesure d'incorporer des informations utiles pour prédire et corriger l'homogénéité de son espace latent. En bref, nos résultats indiquent que le modèle a capturé d'importantes régularités compositionnelles qui sont, au moins sommairement, alignées sur la perception humaine. Cependant, seules de futures itérations avec notre approche fourniront une réponse plus définitive.

Une dernière contribution scientifique, et non la moindre, concerne la méthodologie MLDS elle-même. Pour investiguer des trajectoires circulaires à travers l'espace latent, nous avons étendu la MLDS aux espaces physiques périodiques (voir Section.5.2). Cette variante est en fait possible parce que la plupart des triplets expérimentaux, tels que définis dans la méthode canonique, ne sont en fait que très peu informatifs et qu'ils peuvent ainsi être omis. En conséquence, nous pouvons également réduire considérablement le nombre de combinaisons évaluées par condition expérimentale et contrôler directement la durée de la tâche. Cependant,

## Appendices

pour comprendre et prouver cette hypothèse, il a été nécessaire de mener des travaux théoriques qui n'avaient pas été prévus à l'origine. Nous avons alors redéfini les contours de la méthode de Thurstone et de la MLDS, et pu fournir une explication quant aux divergences signalées, mais souvent exagérées, entre les deux méthodes. Ce travail fascinant a enfin exposé des problèmes théoriques associés à la MLDS, en particulier concernant le caractère non-normal de la distribution de la mesure perceptive de distance entre des paires de stimuli.

Dans le but de reproduire des dessins générés par le modèle sur du papier avec un traceur numérique, j'ai développé différentes techniques créant une épaisseur expressive et dynamique des lignes. J'ai adopté une stratégie s'adaptant à l'échelle du dessin et à la nature de l'outil – stylo ou pinceau (voir Section.6.2). Enfin, j'ai proposé différents principes visuels pour représenter la diversité et la continuité de l'espace compositionnel, ainsi que pour révéler la dynamique des éléments graphiques (voir Section.6.3).

Pour résumer l'ensemble du projet, notre approche sert à valider une nouvelle approche de la modélisation de la composition picturale. Nous introduisons pour commencer un paradigme compositionnel qui soutient un modèle d'apprentissage profond fonctionnel, et qui atteste que la composition peut être modélisée comme un objet continu et hyper-dimensionnel. Nous démontrons ensuite que les régularités compositionnelles capturées et les mesures associées présentent des similitudes avec la perception humaine. Les modèles et protocoles expérimentaux adoptés en sont encore à un stade précoce de développement et présentent donc certaines limites. Cependant, nos résultats sont encourageants et indiquent la possibilité réelle que des phénomènes perceptifs complexes tels que l'art et la composition, qui ne sont pas facilement réductibles à des composantes élémentaires, peuvent dans une certaine mesure être étudiés de manière quantitative.

Au début de ce manuscrit, je m'interroge sur le fait que ma pratique artistique et l'approche de modélisation proposée puissent servir de point de contact significatif entre des efforts créatifs de nature artistique et des méthodes de la recherche scientifique. J'espère que le travail présenté et les contributions associées plaident de manière convaincante sur la façon dont le particulier peut apporter des avancées à des connaissances plus universelles.

# Bibliography

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. <https://www.tensorflow.org/>
- Abbott, E. A. (1884). *Flatland: A Romance of Many Dimensions*. Seeley. <https://github.com/lvesvdf/flatland>
- Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018, February 13). *Fixing a Broken ELBO*. <http://arxiv.org/abs/1711.00464>
- Apel, U. (2009). *KanjiVG*. <http://kanjivg.tagaini.net>.
- Arnheim, R. (2004). *Art and Visual Perception – A Psychology of the Creative Eye* (2nd edition, 50th Anniversary). Berkeley : University of California Press. (Original work published 1954)
- Atlan, H. (2006). *L'organisation biologique et la théorie de l'information*. Éditions du Seuil. <http://banq.prenumerique.ca/accueil/isbn/9782021331349> (Original work published 1972)  
OCLC: 991493880
- Audry, S. (2021). *Art in the Age of Machine Learning*. The MIT Press.
- Audry, S., & Ippolito, J. (2019). Can Artificial Intelligence Make Art without Artists? Ask the Viewer. *Arts*, 8(1), 35. <https://doi.org/10.3390/arts8010035>
- Bayer, J., & Osendorfer, C. (2015, March 5). *Learning Stochastic Recurrent Networks*. <http://arxiv.org/abs/1411.7610>
- Behrman, B. W., & Brown, D. R. (1968). Multidimensional Scaling of Form: A Psychophysical Analysis. *Perception & Psychophysics*, 4(1), 19–25. <https://doi.org/10.3758/BF03210441>
- Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015, September 23). *Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks*. <http://arxiv.org/abs/1506.03099>
- Bengio, Y., Boulanger-Lewandowski, N., & Pascanu, R. (2012, December 13). *Advances in Optimizing Recurrent Networks*. <http://arxiv.org/abs/1212.0901>
- Berio, D., Akten, M., Leymarie, F. F., Grierson, M., & Plamondon, R. (2017). *Calligraphic Stylisation Learning with a Physiologically Plausible Model of Movement and Recurrent Neural Networks*. <https://doi.org/10.1145/3077981.3078049>
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag. <http://www.springer.com/us/book/9780387310732>
- Boschman, M. C. (2001). DifScal: A tool for analyzing difference ratings on an ordinal category scale. *Behavior Research Methods, Instruments, & Computers*, 33(1), 10–20. <https://doi.org/10.3758/BF03195343>
- Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2016, May 12). *Generating Sentences from a Continuous Space*. <http://arxiv.org/abs/1511.06349>
- Brillouin, L. (1988). *La science et la théorie de l'information*. Éditions Jacques Gabay. <http://www.sudoc.fr/001255002> (Original work published 1959)



## Bibliography

- Burda, Y., Grosse, R., & Salakhutdinov, R. (2016, November 7). *Importance Weighted Autoencoders*. <http://arxiv.org/abs/1509.00519>
- Chatonsky, G., Joyeux-Prunel, B., & Cadain, A. (2017, September 25). *Qu'est-ce que l'imagination (artificielle)? – Postdigital*. <http://postdigital.ens.fr/archives/portfolio/imagination>
- Chen, T. Q., Li, X., Grosse, R., & Duvenaud, D. (2018, February 13). *Isolating Sources of Disentanglement in Variational Autoencoders*. <http://arxiv.org/abs/1802.04942>
- Chen, X., & McMains, S. (2005). Polygon Offsetting by Computing Winding Numbers. *Volume 2: 31st Design Automation Conference, Parts A and B*, 565–575. <https://doi.org/10.1115/DETC2005-85513>
- Cheng, F. (1989). *Souffle-Esprit: textes théoriques chinois sur l'art pictural*. Seuil.
- Cheng, F. (2006). *Vide et plein: le langage pictural chinois*. Seuil  
OCLC: 255037435.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, September 2). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. <http://arxiv.org/abs/1406.1078>
- Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018, December 3). *Deep Learning for Classical Japanese Literature*. <https://doi.org/10.20676/00000341>
- Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015, November 23). *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. <http://arxiv.org/abs/1511.07289>
- Da Vinci, L. (1955). *The notebooks of Leonardo da Vinci* (E. McCurdy, Ed.). G. Braziller. <https://archive.org/details/noteboo00leon>
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>
- Doersch, C. (2016, June 19). *Tutorial on Variational Autoencoders*. <http://arxiv.org/abs/1606.05908>
- Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2), 112–122. <https://doi.org/10.3138/FM57-6770-U75U-7727>
- Eitz, M., Hays, J., & Alexa, M. (2012). How Do Humans Sketch Objects? *ACM Trans. Graph.*, 31(4), 44:1–44:10. <https://doi.org/10.1145/2185520.2185540>
- Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017, June 21). *CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms*. <http://arxiv.org/abs/1706.07068>
- Favreau, J.-D., Lafarge, F., & Bousseau, A. (2016). Fidelity vs. Simplicity: A Global Approach to Line Drawing Vectorization. *ACM Trans. Graph.*, 35(4), 120:1–120:10. <https://doi.org/10.1145/2897824.2925946>
- Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig : Breitkopf und Härtel. <http://archive.org/details/elementederpsych001fech>
- Feldman, J., & Singh, M. (2006). Bayesian Estimation of the Shape Skeleton. *Proceedings of the National Academy of Sciences*, 103(47), 18014–18019. <https://doi.org/10.1073/pnas.0608811103>
- Fischer, A., & Plamondon, R. (2015). A Dissimilarity Measure for On-Line Signature Verification Based on the Sigma-Lognormal Model.
- Flash, T., & Hogan, N. (1985). The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 5(7), 1688–1703.

- Floreano, D., & Mattiussi, C. (2008, August 22). *Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies*. The MIT Press.
- Focillon, H. (1934). *Vie des formes*. P.U.F.
- Fournier, A., & Barsky, B. A. (1985a). Geometric Continuity with Interpolating Bézier Curves. In N. Magnenat-Thalmann & D. Thalmann (Eds.), *Computer-Generated Images* (pp. 153–158). Springer Japan. [https://doi.org/10.1007/978-4-431-68033-8\\_14](https://doi.org/10.1007/978-4-431-68033-8_14)
- Fournier, A., & Barsky, B. A. (1985b). Geometric Continuity with Interpolating Bézier Curves. *Graphics Interface '85, Montréal*, 5 pages. <https://doi.org/10.20380/GI1985.48>
- Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10), 2451–2471. <https://doi.org/10.1162/089976600300015015>
- Glorot, X., & Bengio, Y. (2010, May 13–15). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterton (Eds.), *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256). PMLR. <https://proceedings.mlr.press/v9/glorot10a.html>
- Goodfellow, I. (2016, December 31). *NIPS 2016 Tutorial: Generative Adversarial Networks*. <http://arxiv.org/abs/1701.00160>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014, June 10). *Generative Adversarial Networks*. <http://arxiv.org/abs/1406.2661>
- Goyal, K., Dyer, C., & Berg-Kirkpatrick, T. (2017, April 23). *Differentiable Scheduled Sampling for Credit Assignment*. <http://arxiv.org/abs/1704.06970>
- Graves, A. (2013, August 4). *Generating Sequences With Recurrent Neural Networks*. <http://arxiv.org/abs/1308.0850>
- Ha, D., & Eck, D. (2017, April 11). *A Neural Representation of Sketch Drawings*. <http://arxiv.org/abs/1704.03477>
- He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). *Deep Residual Learning for Image Recognition*. <http://arxiv.org/abs/1512.03385>
- Hertzmann, A. (2018). Can Computers Create Art? *Arts*, 7(2), 18. <https://doi.org/10.3390/arts7020018>
- Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., & Lerchner, A. (2016, June 17). *Early Visual Concept Learning with Unsupervised Deep Learning*. <http://arxiv.org/abs/1606.05579>
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). B-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 22.
- Hilare, X., & Tombe, K. (2006). Robust and Accurate Vectorization of Line Drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6), 890–904. <https://doi.org/10.1109/TPAMI.2006.127>
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012, July 3). *Improving neural networks by preventing co-adaptation of feature detectors*. <http://arxiv.org/abs/1207.0580>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hoffman, M. D., & Johnson, M. J. (2016). ELBO Surgery: Yet Another Way to Carve up the Variational Evidence Lower Bound. *NIPS 2016*, 4. <http://approximateinference.org/accepted/HoffmanJohnson2016.pdf>
- Hübner, R., & Fillinger, M. G. (2016). Comparison of Objective Measures for Predicting Perceptual Balance and Visual Aesthetic Preference. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00335>

## Bibliography

- Hurtut, T., Gousseau, Y., Cheriet, F., & Schmitt, F. (2011). Artistic Line-drawings Retrieval Based on the Pictorial Content. *J. Comput. Cult. Herit.*, 4(1), 3:1–3:23. <https://doi.org/10.1145/2001416.2001419>
- Huszár, F. (2015, November 16). *How (not) to Train your Generative Model: Scheduled Sampling, Likelihood, Adversary?* <http://arxiv.org/abs/1511.05101>
- Introduction à la calligraphie chinoise*. (1997). Ed. du centenaire  
OCLC: 40805953.
- Jang, E., Gu, S., & Poole, B. (2017, August 5). *Categorical Reparameterization with Gumbel-Softmax*. <http://arxiv.org/abs/1611.01144>
- Janssen, R., & Vossepoel, A. (1997). Adaptive Vectorization of Line Drawing Images. *Computer Vision and Image Understanding*, 65, 38–56. <https://doi.org/10.1006/cviu.1996.0484>
- Jongejan, J., Rowley, H., Kawashima, T., Kim, J., & Fox-Gieg, N. (2016). *The Quick, Draw! - A.I. Experiment*. <https://quickdraw.withgoogle.com/>
- Kandinsky, W. (1889). *Du spirituel dans l'art, et dans la peinture en particulier* (P. Sers, Ed.; N. Debrand & B. Du Crest, Trans.). Denoël : Gallimard. (Original work published 1912)
- Kandinsky, W. (1991). *Point et ligne sur plan: contribution à l'analyse des éléments picturaux* (P. Sers, Ed.; S. Leppien & J. Leppien, Trans.). Gallimard. (Original work published 1926)
- Kandinsky, W. (2014). *Regards sur le passé: et autres textes, 1912-1922* (J.-P. Bouillon, Ed.). Hermann. (Original work published 1974)
- Kholmatov, A., & Yanikoglu, B. (2009). SUSIG: An on-line signature database, associated protocols and benchmark results. *Formal Pattern Analysis & Applications*, 12, 227–236. <https://doi.org/10.1007/s10044-008-0118-x>
- Kim, B., Wang, O., Öztireli, A. C., & Gross, M. (2018). Semantic Segmentation for Line Drawing Vectorization Using Neural Networks. *Computer Graphics Forum (Proc. Eurographics)*, 37(2), 329–338.
- Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics a Practical Introduction*. Elsevier Science & Technology Books  
OCLC: 892784285.
- Kingma, D. P., & Ba, J. (2017, January 29). *Adam: A Method for Stochastic Optimization*. <http://arxiv.org/abs/1412.6980>
- Kingma, D. P., & Welling, M. (2013, December 20). *Auto-Encoding Variational Bayes*. <http://arxiv.org/abs/1312.6114>
- Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4), 307–392. <https://doi.org/10.1561/22000000056>
- Klee, P. (1961). *Notebooks, Volume 1 : The Thinking Eye* (J. Spiller, Ed.; R. Manheim, C. Weidler, & J. Wittenborn, Trans.). Lund Humphries  
OCLC: 502482338.
- Klee, P. (1998). *Théorie de l'art moderne* (P.-H. Gonthier, Ed. & Trans.). Denoël : Gallimard. (Original work published 1924)
- Knoblauch, K., & Maloney, L. T. (2008). MLDS : Maximum Likelihood Difference Scaling in R. *Journal of Statistical Software*, 25(2). <https://doi.org/10.18637/jss.v025.i02>
- Kraft, D. (1988). *A Software Package for Sequential Quadratic Programming* (Tech. Rep. DFVLR-FB 88-28). DFVLR  
Open Library ID: OL18926873M.
- Krizhevsky, A., Sutskever, I., & E. Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25. <https://doi.org/10.1145/3065386>

- Kumar, A., Sattigeri, P., & Balakrishnan, A. (2018, December 27). *Variational Inference of Disentangled Latent Concepts from Unlabeled Observations*. <http://arxiv.org/abs/1711.00848>
- Lamb, A., Goyal, A., Zhang, Y., Zhang, S., Courville, A., & Bengio, Y. (2016, October 27). *Professor Forcing: A New Algorithm for Training Recurrent Networks*. <http://arxiv.org/abs/1610.09038>
- Lao Tseu. (2008). *Tao te king* (S. Mitchell & B. Labayle, Trans.). Synchronique éd. OCLC: 495234553.
- Le Moigne, J.-L. (2006). *La théorie du système général: théorie de la modélisation* (4ème Ed (1994) - Les Classiques du Réseau Intelligence de la Complexité (2006)). P.U.F. <http://www.mcxapc.org/inserts/ouvrages/0609tsgtm.pdf> (Original work published 1977) OCLC: 11665596
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.*, 1(4), 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>
- LeCun, Y., Cortes, C., & Burges, C. J. (1998). *The MNIST Database of Handwritten Digits*. <http://yann.lecun.com/exdb/mnist/>
- Lee, T., Kashyap, R., & Chu, C. (1994). Building Skeleton Models via 3-D Medial Surface Axis Thinning Algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6), 462–478. <https://doi.org/10.1006/cgip.1994.1042>
- Lee, Y. J., Lawrence Zitnick, C., & Cohen, M. (2011). ShadowDraw: Real-Time User Guidance for Freehand Drawing. *ACM Trans. Graph.*, 30, 27. <https://doi.org/10.1145/2010324.1964922>
- Leiva, L. A., Martín-Albo, D., & Plamondon, R. (2015). Gestures À Go Go: Authoring Synthetic Human-Like Stroke Gestures Using the Kinematic Theory of Rapid Movements. *ACM Trans. Intell. Syst. Technol.*, 7(2), 15:1–15:29. <https://doi.org/10.1145/2799648>
- Lelièvre, P., & Neri, P. (2019). Abstract Painting Composition : A Deep Learning Model of the Orientation Perception and Judgement. *Perception*, 48 (2supp), 77–77. <https://doi.org/10.1177/0301006619863862>
- Lelièvre, P., & Neri, P. (2021). A Deep-Learning Framework for Human Perception of Abstract Art Composition. *Journal of Vision*, 21(5), 9. <https://doi.org/10.1167/jov.21.5.9>
- Lelièvre, P., & Neri, P. (2022a). Perceptual Exploration of Latent Space for Pictorial Composition. <https://doi.org/10.1167/jov.22.14.3287>
- Lelièvre, P., & Neri, P. (2022b). Perceptual Exploration of Latent Space for Pictorial Composition.
- LeWitt, S. (1967). Paragraphs on conceptual art. *Artforum*, 5(10), 79–83.
- Liao, C.-W., & Huang, J. S. (1990). Stroke Segmentation by Bernstein-Bezier Curve Fitting. *Pattern Recognition*, 23(5), 475–484. [https://doi.org/10.1016/0031-3203\(90\)90068-V](https://doi.org/10.1016/0031-3203(90)90068-V)
- Lindbloom, B. (2017). *Color Space Transformations*. <http://www.brucelindbloom.com/>
- Lioret, A. (2005). Being Paintings. *ACM SIGGRAPH 2005 Electronic Art and Animation Catalog*, 186–190. <https://doi.org/10.1145/1086057.1086145>
- Lories, D., & Lenain, T. (2002). *Esthétique et philosophie de l'art: repères historiques et thématiques* (L. d'esthétique, Ed.). De Boeck.
- Maddison, C. J., Mnih, A., & Teh, Y. W. (2017, March 5). *The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables*. <http://arxiv.org/abs/1611.00712>
- Malevich, K. S. (2003). *The Non-Objective World: The Manifesto of Suprematism*. Dover Publications. (Original work published 1927)
- Maloney, L. T., & Yang, J. N. (2003). Maximum Likelihood Difference Scaling. *Journal of Vision*, 3(8), 5. <https://doi.org/10.1167/3.8.5>

## Bibliography

- Masters, D., & Luschi, C. (2018, April 20). *Revisiting Small Batch Training for Deep Neural Networks*. <http://arxiv.org/abs/1804.07612>
- Mathieu, E., Rainforth, T., Siddharth, N., & Teh, Y. W. (2019, June 12). *Disentangling Disentanglement in Variational Autoencoders*. <http://arxiv.org/abs/1812.02833>
- Matisse, H. (2014). *Écrits et propos sur l'art* (D. Fourcade, Ed.). Hermann  
OCLC: 894424258.
- McManus, I. C., Stöver, K., & Kim, D. (2011). Arnheim's Gestalt Theory of Visual Balance: Examining the Compositional Structure of Art Photographs and Abstract Images. *i-Perception*, 2(6), 615–647. <https://doi.org/10.1068/i0445aap>
- Mishra, B., Meyer, G., & Sepulchre, R. (2011). Low-Rank Optimization for Distance Matrix Completion. *IEEE Conference on Decision and Control and European Control Conference*, 4455–4460. <https://doi.org/10.1109/CDC.2011.6160810>
- Moreno, A. (1998). Information, Causality and Self-Reference in Natural and Artificial Systems. *AIP Conference Proceedings*, 437(1), 202–206. <https://doi.org/10.1063/1.56301>
- Moreno, A. (2004). Auto-organisation, autonomie et identité. *Revue internationale de philosophie*, 228(2), 135–150. <https://doi.org/10.3917/rip.228.0135>
- Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807–814.
- Neri, P. (2014). Semantic Control of Feature Extraction from Natural Scenes. *Journal of Neuroscience*, 34(6), 2374–2388. <https://doi.org/10.1523/JNEUROSCI.1755-13.2014>
- Noris, G., Hornung, A., Sumner, R. W., Simmons, M., & Gross, M. (2013). Topology-driven Vectorization of Clean Line Drawings. *ACM Trans. Graph.*, 32(1), 4:1–4:11. <https://doi.org/10.1145/2421636.2421640>
- O'Reilly, C., & Plamondon, R. (2008). Automatic Extraction of Sigma-Lognormal Parameters on Signatures. *ICFHR Proceedings*, 7. <http://www.cenparmi.concordia.ca/ICFHR2008/Proceedings/papers/cr1020.pdf>
- Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2014, April 24). *How to Construct Deep Recurrent Neural Networks*. <http://arxiv.org/abs/1312.6026>
- Pascanu, R., Mikolov, T., & Bengio, Y. (2012, November 21). *On the Difficulty of Training Recurrent Neural Networks*. <http://arxiv.org/abs/1211.5063>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Plamondon, R. (1995a). A Kinematic Theory of Rapid Human Movements: Part I. Movement Representation and Generation. *Biological Cybernetics*, 72(4), 295–307. <https://doi.org/10.1007/BF00202785>
- Plamondon, R. (1995b). A Kinematic Theory of Rapid Human Movements: Part II. Movement Time and Control. *Biological Cybernetics*, 72(4), 309–320. <https://doi.org/10.1007/BF00202786>

- Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, *38*(33), 7255–7269. <https://doi.org/10.1523/JNEUROSCI.0388-18.2018>
- Ramer, U. (1972). An Iterative Procedure for the Polygonal Approximation of Plane Curves. *Computer Graphics and Image Processing*, *1*(3), 244–256. [https://doi.org/10.1016/S0146-664X\(72\)80017-0](https://doi.org/10.1016/S0146-664X(72)80017-0)
- Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016, May 6). *Sequence Level Training with Recurrent Neural Networks*. <http://arxiv.org/abs/1511.06732>
- Redies, C., Brachmann, A., & Hayn-Leichsenring, G. U. (2015). Changes of Statistical Properties During the Creation of Graphic Artworks. *Art & Perception*, *3*(1), 93–116. <https://doi.org/10.1163/22134913-00002017>
- Redies, C., Brachmann, A., & Wagemans, J. (2017). High Entropy of Edge Orientations Characterizes Visual Artworks from Diverse Cultural Backgrounds. *Vision Research*, *133*, 130–144. <https://doi.org/10.1016/j.visres.2017.02.004>
- Rezende, D. J., Mohamed, S., & Wierstra, D. (2014, May 30). *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*. <http://arxiv.org/abs/1401.4082>
- Rezende, D. J., & Viola, F. (2018, October 1). *Taming VAEs*. <http://arxiv.org/abs/1810.00597>
- Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Springer-Verlag. <http://www.springer.com/us/book/9783540605058>
- Ross, D. W. (1907). *A Theory of Pure Design; Harmony, Balance, Rhythm*. Boston, Houghton, Mifflin. <http://archive.org/details/theoryofpuredesi00rossuoft>
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015, January 29). *ImageNet Large Scale Visual Recognition Challenge*. <http://arxiv.org/abs/1409.0575>
- Schmidhuber, J. (2009, April 15). *Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes*. <http://arxiv.org/abs/0812.4360>
- Schneider, P. J. (1990). An Algorithm for Automatically Fitting Digitized Curves. *Graphics Gems* (pp. 612–626). Academic Press Professional, Inc.
- Schrödinger, E. (2013). *What is Life? The Physical Aspect of the Living Cell with Mind and Matter & Autobiographical Sketches* (14th printing). Cambridge Univ. Press. (Original work published 1944)  
OCLC: 1012743572
- Schuster, M., & Paliwal, K. (1997). Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.*, *45*(11), 2673–2681. <https://doi.org/10.1109/78.650093>
- Schwabe, K., Menzel, C., Mullin, C., Wagemans, J., & Redies, C. (2018). Gist Perception of Image Composition in Abstract Artworks. *i-Perception*, *9*, 204166951878079. <https://doi.org/10.1177/2041669518780797>
- Selinger, P. (2019). *Potrace*. <http://potrace.sourceforge.net>
- Sérisier, P. (1921). *A B C de la peinture*. La Douce France - Henri Floury. <https://gallica.bnf.fr/ark:/12148/bpt6k1170360n>
- Shoemake, K. (1985). Animating Rotation with Quaternion Curves. *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '85*, 245–254. <https://doi.org/10.1145/325334.325242>
- Simonyan, K., & Zisserman, A. (2014, September 4). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. <http://arxiv.org/abs/1409.1556>

## Bibliography

- Steinbruner, J. D. (2021). *The Cybernetic Theory of Decision: New Dimensions of Political Analysis*. <https://doi.org/10.1515/9781400823796> (Original work published 1974)  
OCLC: 1266229387
- Thurstone, L. L. (1927a). A Law of Comparative Judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Thurstone, L. L. (1927b). A Mental Unit of Measurement. *Psychological Review*, 34(6), 415–423. <https://doi.org/10.1037/h0071456>
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-Image: Image Processing in Python. *PeerJ*, 2, e453. <https://doi.org/10.7717/peerj.453>
- Vatti, B. R. (1992). A generic solution to polygon clipping. *Communications of the ACM*, 35(7), 56–63. <https://doi.org/10.1145/129902.129906>
- Verostko, R. (1990). Epigenetic Painting: Software as Genotype. *Leonardo*, 23(1), 17. <https://doi.org/10.2307/1578459>
- von Foerster, H. (2003). On Self-Organizing Systems and Their Environments. *Understanding Understanding: Essays on Cybernetics and Cognition* (pp. 1–19). Springer. [https://doi.org/10.1007/0-387-21722-3\\_1](https://doi.org/10.1007/0-387-21722-3_1) (Original work published 1960)  
OCLC: 818994191
- Watson, A. B., & Kreslake, L. (2001, June 8). Measurement of visual impairment scales for digital video. In B. E. Rogowitz & T. N. Pappas (Eds.). <https://doi.org/10.1117/12.429526>
- Weber, M. (2016). *Autotrace*. <http://autotrace.sourceforge.net>
- White, T. (2016, December 6). *Sampling Generative Networks*. <http://arxiv.org/abs/1609.04468>
- Wijntjes, M. W. A., Spoiala, C., & Ridder, H. de. (2020). Thurstonian Scaling and the Perception of Painterly Translucency. *Art & Perception*, 8(3-4), 363–386. <https://doi.org/10.1163/22134913-bja10021>
- WikiArt. (n.d.). *WikiArt.org - Visual Art Encyclopedia*. [www.wikiart.org](http://www.wikiart.org). <https://www.wikiart.org/>
- Williams, C. K. (2002). On a Connection between Kernel PCA and Metric Multidimensional Scaling. *Machine Learning*, 46(1/3), 11–19. <https://doi.org/10.1023/A:1012485807823>
- Wilson, A., & Chatterjee, A. (2005). The Assessment of Preference for Balance: Introducing a New Test. *Empirical Studies of the Arts*, 23(2), 165–180. <https://doi.org/10.2190/B1LR-MVF3-F36X-XR64>
- Yamins, D. L., & DiCarlo, J. J. (2016). Using Goal-Driven Deep Learning Models to Understand Sensory Cortex. *Nature Neuroscience*, 19(3), 356–365.
- Yee, C. (1974, January 1). *Chinese Calligraphy: An Introduction to Its Aesthetic and Technique* (3rd Revised & Enlarged edition). Harvard University Press.
- Zhang, T. Y., & Suen, C. Y. (1984). A Fast Parallel Algorithm for Thinning Digital Patterns. *Commun. ACM*, 27(3), 236–239. <https://doi.org/10.1145/357994.358023>
- Zhao, S., Song, J., & Ermon, S. (2017, June 7). *InfoVAE: Information Maximizing Variational Autoencoders*. <http://arxiv.org/abs/1706.02262>

## Supplementary bibliography

- Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired: Science*, 16. <https://www.wired.com/2008/06/pb-theory/>
- Anderson, E. C. (1999, October). Monte Carlo Methods and Importance Sampling. [https://ib.berkeley.edu/labs/slatkin/eriq/classes/guest\\_lect/mc\\_lecture\\_notes.pdf](https://ib.berkeley.edu/labs/slatkin/eriq/classes/guest_lect/mc_lecture_notes.pdf)
- Aubry, M., & Russell, B. (2015, June 3). *Understanding deep features with computer-generated imagery*. <http://arxiv.org/abs/1506.01151>
- Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep Convolutional Networks Do Not Classify Based on Global Object Shape (W. Einhäuser, Ed.). *PLOS Computational Biology*, 14(12), e1006613. <https://doi.org/10.1371/journal.pcbi.1006613>
- Berger, P., & Lioret, A. (2012). *L'art génératif: jouer à Dieu, un droit? un devoir?* L'Harmattan  
OCLC: 816613559.
- Bertsekas, D. P., & Tsitsiklis, J. N. (2002, June 30). *Introduction To Probability*. Athena Scientific.
- Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an Object Through Feature Space. *Nature*, 408(6809), 196–199. <https://doi.org/10.1038/35041567>
- Borillo, M., & Goulette, J.-P. (2002). *Cognition et création: explorations cognitives des processus de conception*. Editions Mardaga.
- Budninskiy, M., Yin, G., Feng, L., Tong, Y., & Desbrun, M. (2018, November 2). *Parallel Transport Unfolding: A Connection-based Manifold Learning Approach*. <http://arxiv.org/abs/1806.09039>
- Buduma, N. (2015, November 30). *Fundamentals of Deep Learning*. O'Reilly.
- Calude, C. S., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, 22(3), 595–612. <https://doi.org/10.1007/s10699-016-9489-4>
- Carbon, C.-C., & Fingerhut, J. (2017). Abstracts from the 5th Visual Science of Art Conference (VSAC). *Art and Perception*, 5(4), 337–426. <https://doi.org/10.1163/22134913-00002099>
- Chang, D. H., & Troje, N. F. (2009). Acceleration Carries the Local Inversion Effect in Biological Motion Perception. *J Vis*, 9(1), 1–17.
- Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The Structures of Letters and Symbols throughout Human History Are Selected to Match Those Found in Objects in Natural Scenes. *The American Naturalist*, 167(5), E117–E139. <https://doi.org/10.1086/502806>
- Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., & Bengio, Y. (2016, April 6). *A Recurrent Latent Variable Model for Sequential Data*. <http://arxiv.org/abs/1506.02216>
- Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4), 305–317. <https://doi.org/10.1016/j.tics.2019.01.009>
- Cools, M., & Peteau, M. (1974). Un programme de stimulation inventive : STIM 5. *Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle*, 8(V3), 5–19. [http://www.numdam.org/item?id=RO\\_1974\\_\\_8\\_3\\_5\\_0](http://www.numdam.org/item?id=RO_1974__8_3_5_0)



## Supplementary bibliography

- Cusack, J. P., Williams, J. H., & Neri, P. (2015). Action Perception Is Intact in Autism Spectrum Disorder. *Journal of Neuroscience*, *35*(5), 1849–1857.
- Cuzick, J. (1985). A wilcoxon-type test for trend. *Statistics in Medicine*, *4*(1), 87–90. <https://doi.org/10.1002/sim.4780040112>
- Das, A., & Geisler, W. S. (2021, July 27). *A method to integrate and classify normal distributions*. <http://arxiv.org/abs/2012.14331>
- Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2018, July 10). *Universal Transformers*. <http://arxiv.org/abs/1807.03819>
- Delaunay, R. (1913). Ueber das Licht (P. Klee, Trans.) [magazine]. *Der Sturm, Wochenschrift für Kultur und die Künste*, (144-145), 255–256.
- Devue, C., & Barsics, C. (2016). Outlining Face Processing Skills of Portrait Artists: Perceptual Experience with Faces Predicts Performance. *Vision Res.*, *127*, 92–103.
- Dodge, S., & Karam, L. (2017). A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*, 1–7. <https://doi.org/10.1109/ICCCN.2017.8038465>
- Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding Reveals Fundamental Differences in Local Vs. Global Processing in Humans and Machines. *Vision Research*, *167*, 39–45.
- Dosovitskiy, A., & Brox, T. (2015, June 8). *Inverting Visual Representations with Convolutional Networks*. <http://arxiv.org/abs/1506.02753>
- Dowling, C. (2014). *Aesthetic Formalism*. Internet Encyclopedia of Philosophy. <https://www.iep.utm.edu/aes-form/>  
The Open University, United Kingdom
- Dubal, S., Lerebours, A.-E., Taffou, M., Pelletier, J., Escande, Y., & Knoblauch, K. (2014). A Psychophysical Exploration of the Perception of Emotion from Abstract Art. *Empirical Studies of the Arts*, *32*(1), 27–41. <https://doi.org/10.2190/EM.32.1.EOV.4>
- Dumoulin, S. O., & Wandell, B. A. (2008). Population Receptive Field Estimates in Human Visual Cortex. *Neuroimage*, *39*, 647–660.
- Dumoulin, V., Shlens, J., & Kudlur, M. (2016, October 24). *A Learned Representation For Artistic Style*. <http://arxiv.org/abs/1610.07629>
- Elgammal, A. M., Mazzone, M., Liu, B., Kim, D., & Elhoseiny, M. (2018). *The Shape of Art History in the Eyes of the Machine*. <http://arxiv.org/abs/1801.07729>
- Fabius, O., & van Amersfoort, J. R. (2015, June 15). *Variational Recurrent Auto-Encoders*. <http://arxiv.org/abs/1412.6581>
- Fiedler, J., & Feierabend, P. (2013, June 9). *Bauhaus*. HF Ullmann Editions.
- Fischer, A., Plamondon, R., O'Reilly, C., & Savaria, Y. (2014). Neuromuscular Representation and Synthetic Generation of Handwritten Whiteboard Notes. *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, 2014*. <https://doi.org/10.1109/ICFHR.2014.45>
- Fourmentraux, J.-P., Schmitt, A., Chatonsky, G., Bianchini, S., & Edric Stanley, D. (2011). Ce que la programmation fait à l'art. *Art++* (p. 14). Éditions HXV.
- Fraccaro, M., Sønderby, S. K., Paquet, U., & Winther, O. (2016, November 13). *Sequential Neural Models with Stochastic Layers*. <http://arxiv.org/abs/1605.07571>
- Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A Functional and Perceptual Signature of the Second Visual Area in Primates. *Nat. Neurosci.*, *16*(7), 974–981.
- Ganin, Y., Kulkarni, T., Babuschkin, I., Eslami, S. M. A., & Vinyals, O. (2018, April 3). *Synthesizing Programs for Images using Reinforced Adversarial Learning*. <http://arxiv.org/abs/1804.01118>

- Gaspar, C. M., Bennett, P. J., & Sekuler, A. B. (2008). The Effects of Face Inversion and Contrast-Reversal on Efficiency and Internal Noise. *Vision Research*, 48, 1084–1095.
- Gatys, L. A., Ecker, A. S., & Bethge, M. (2015a, August 26). *A Neural Algorithm of Artistic Style*. <http://arxiv.org/abs/1508.06576>
- Gatys, L. A., Ecker, A. S., & Bethge, M. (2015b, May 27). *Texture Synthesis Using Convolutional Neural Networks*. <http://arxiv.org/abs/1505.07376>
- Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2414–2423. <https://doi.org/10.1109/CVPR.2016.265>
- Ghent, L. (1961). Form and Its Orientation: A Child's-Eye View. *The American Journal of Psychology*, 74(2), 177–190. <https://doi.org/10.2307/1419403>
- Gombrich, E. H. (1995). *The Story of Art*  
OCLC: 33229586.
- Gong, M., Xuan, Y., Smart, L. J., & Olzak, L. A. (2018). The Extraction of Natural Scene Gist in Visual Crowding. *Sci Rep*, 8(1), 14073.
- Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015, February 16). *DRAW: A Recurrent Neural Network For Image Generation*. <http://arxiv.org/abs/1502.04623>
- Gress, T. (n.d.). *Le retrait de la lumière, libération de l'âme*. <http://www.espritudavant.com/DetailElement.aspx?numStructure=79255&numElement=92211>
- Grossman, E. D., & Blake, R. (2002). Brain Areas Active During Visual Perception of Biological Motion. *Neuron*, 35(6), 1167–1175.
- Ha, D. (2015, November 24). *Mixture Density Networks with TensorFlow*. <http://blog.otoro.net/2015/11/24/mixture-density-networks-with-tensorflow/>
- Hawley-Dolan, A., & Winner, E. (2011). Seeing the Mind Behind the Art: People Can Distinguish Abstract Expressionist Paintings From Highly Similar Paintings by Children, Chimps, Monkeys, and Elephants. *Psychological Science*, 22(4), 435–441. <https://doi.org/10.1177/0956797611400915>
- Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11), 1173–1185. <https://doi.org/10.1038/s41562-020-00951-3>
- Hennig, J., Umakantha, A., & Williamson, R. (2017). Sequence Generation and Classification with VaeS and Rnns, 9.
- Hill, A. (1968). Art and Mathesis: Mondrian's Structures. *Leonardo*, 1(3), 233–242. <https://muse.jhu.edu/article/596571/summary>
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504–507. <https://doi.org/10.1126/science.1127647>
- Hübner, R., & Fillinger, M. G. (2019). Perceptual Balance, Stability, and Aesthetic Appreciation: Their Relations Depend on the Picture Type. *i-Perception*, 10(3), 2041669519856040. <https://doi.org/10.1177/2041669519856040>
- Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2020). Aesthetic Preference for Art Emerges from a Weighted Integration Over Hierarchically Structured Visual Features in the Brain. *bioRxiv 2020.02.09.940353*. <https://doi.org/10.1101/2020.02.09.940353>
- Johnson, J., Alahi, A., & Fei-Fei, L. (2016, March 26). *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*. <http://arxiv.org/abs/1603.08155>
- Kaufmann, A. (1969). L'imagination artificielle (heuristique automatique). *Revue Française d'Informatique et de Recherche Opérationnelle*, 3, 5–24. <https://www.zbmath.org/?q=an:0195.17901>

## Supplementary bibliography

- Kelley, T. A., Chun, M. M., & Chua, K. P. (2003). Effects of Scene Inversion on Change Detection of Targets Matched for Visual Saliency. *Journal of vision*, 3(1), 1–5.
- Kim, S., Pasupathy, R., & Henderson, S. G. (2015). A Guide to Sample Average Approximation. In M. C. Fu (Ed.), *Handbook of Simulation Optimization* (pp. 207–243). Springer New York. [https://doi.org/10.1007/978-1-4939-1384-8\\_8](https://doi.org/10.1007/978-1-4939-1384-8_8)
- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2017, January 30). *Improving Variational Inference with Inverse Autoregressive Flow*. <http://arxiv.org/abs/1606.04934>
- Klee, P. (1973). *Notebooks, Volume 2 : The Nature of Nature* (J. Spiller, Ed.; H. Norden & J. Wittenborn, Trans.). Lund Humphries  
OCLC: 959292154.
- Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1, 417–446.
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-Level Concept Learning Through Probabilistic Program Induction. *Science*, 350(6266), 1332–1338. <https://doi.org/10.1126/science.aab3050>
- Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, Horizontal, and Feedback Processing in the Visual Cortex. *Curr. Opin. Neurobiol.*, 8(4), 529–535.
- Latto, R., Brain, D., & Kelly, B. (2000). An Oblique Effect in Aesthetics: Homage to Mondrian (1872–1944). *Perception*, 29(8), 981–987. <https://doi.org/10.1068/p2352>
- Lawless, H. T. (2013, August 30). *Quantitative Sensory Analysis: Psychophysics, Models and Intelligent Design*. John Wiley & Sons. <https://doi.org/10.1002/9781118684818>
- Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and Shared Taste in Art and Face Appreciation. *Frontiers in Human Neuroscience*, 10, 155. <https://doi.org/10.3389/fnhum.2016.00155>
- Levin, D. N. (2000). A Differential Geometric Description of the Relationships among Perceptions. *Journal of Mathematical Psychology*, 44(2), 241–284. <https://doi.org/10.1006/jmps.1999.1240>
- LeWitt, S. (1969). Sentences on Conceptual Art. *Art-Language*, 1(1).
- Li, C., & Wand, M. (2016, January 18). *Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis*. <http://arxiv.org/abs/1601.04589>
- Lin, J.-y. (2014). *Imperial Taste: The Beauty of Calligraphy*  
OCLC: 900918274.
- Lindauer, M. S. (1969). The Orientation of Form in Abstract Art. *Proceedings of the Annual Convention of the American Psychological Association*, 4(1), 475–476.
- Lindauer, M. S. (1987). Perceived and Preferred Orientations of Abstract Art. *Empirical Studies of the Arts*, 5(1), 47–58. <https://doi.org/10.2190/K1X2-X4VJ-6YN9-BKD8>
- Liu, J., Dong, W., Zhang, X., & Jiang, Z. (2017). Orientation Judgment for Abstract Paintings. *Multimedia Tools and Applications*, 76(1), 1017–1036. <https://doi.org/10.1007/s11042-015-3104-5>
- Liu, X., Wong, T.-T., & Heng, P.-A. (2015). Closure-aware Sketch Simplification. *ACM Trans. Graph.*, 34(6), 168:1–168:10. <https://doi.org/10.1145/2816795.2818067>
- Locher, P., Gray, S., & Nodine, C. (1996). The Structural Framework of Pictorial Balance. *Perception*, 25(12), 1419–1436. <https://doi.org/10.1068/p251419>
- Locher, P., Overbeeke, K., & Stappers, P. J. (2005). Spatial Balance of Color Triads in the Abstract Art of Piet Mondrian. *Perception*, 34(2), 169–189. <https://doi.org/10.1068/p5033>
- Locher, P. J. (2003). An Empirical Investigation of the Visual Rightness Theory of Picture Perception. *Acta Psychologica*, 114(2), 147–164. <https://doi.org/10.1016/j.actpsy.2003.07.001>

- Locher, P. J., Stappers, P. J., & Overbeeke, K. (1999). An empirical evaluation of the visual rightness theory of pictorial composition. *Acta Psychologica*, *103*(3), 261–280. [https://doi.org/10.1016/S0001-6918\(99\)00044-X](https://doi.org/10.1016/S0001-6918(99)00044-X)
- Lopes, R. G., Ha, D., Eck, D., & Shlens, J. (2019, April 4). *A Learned Representation for Scalable Vector Graphics*. <http://arxiv.org/abs/1904.02632>
- Mahendran, A., & Vedaldi, A. (2016). Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images. *International Journal of Computer Vision*, *120*(3), 233–255. <https://doi.org/10.1007/s11263-016-0911-8>
- Mamassian, P. (2008). Ambiguities and Conventions in the Perception of Visual Art. *Vision Res.*, *48*(20), 2143–2153.
- Mather, G. (2012). Aesthetic Judgement of Orientation in Modern Art. *i-Perception*, *3*(1), 18–24. <https://doi.org/10.1068/i0447aap>
- Mazzone, M., & Elgammal, A. (2019). Art, Creativity, and the Potential of Artificial Intelligence. *Arts*, *8*(1), 26. <https://doi.org/10.3390/arts8010026>
- McManus, I. C., Cheema, B., & Stoker, J. (1993). The Aesthetics of Composition: A Study of Mondrian. *Empirical Studies of the Arts*, *11*(2), 83–94. <https://doi.org/10.2190/HXR4-VU9A-P5D9-BPQQ>
- McManus, I. C., Edmondson, D., & Rodger, J. (1985). Balance in Pictures. *British Journal of Psychology*, *76*(3), 311–324. <https://doi.org/10.1111/j.2044-8295.1985.tb01955.x>
- McManus, I. C., & Kitson, C. M. (1995). Compositional Geometry in Pictures. *Empirical Studies of the Arts*, *13*(1), 73–94. <https://doi.org/10.2190/66DJ-GVVJ-A33U-X7AT>
- Meadmore, K. L., Liversedge, S. P., Wenger, M. J., & Donnelly, N. (2014). Exploring the Relationship Between Response Time, Sensitivity and Bias in Categorical and Coordinate Visuospatial Processes: Evidence for Hemispheric Specialisation. *Journal of Cognitive Psychology*, *26*(4), 423–432. <https://doi.org/10.1080/20445911.2014.903255>
- Morin, O. (2018). Spontaneous Emergence of Legibility in Writing Systems: The Case of Orientation Anisotropy. *Cognitive Science*, *42*(2), 664–677. <https://doi.org/10.1111/cogs.12550>
- Neri, P. (2010a). How Inherently Noisy Is Human Sensory Processing? *Psychonomic bulletin & review*, *17*, 802–808.
- Neri, P. (2017). Object Segmentation Controls Image Reconstruction from Natural Scenes. *PLoS biology*, *15*(8), e1002611–e1002611. <https://doi.org/10.1371/journal.pbio.1002611>
- Neri, P., & Levi, D. M. (2006). Receptive Versus Perceptive Fields from the Reverse-Correlation Viewpoint. *Vision Research*, *46*, 2465–2474.
- Neri, P., Luu, J. Y., & Levi, D. M. (2006). Meaningful Interactions Can Enhance Visual Discrimination of Human Agents. *Nature Neuroscience*, *9*, 1186–1192.
- Neri, P., Luu, J. Y., & Levi, D. M. (2007). Sensitivity to Biological Motion Drops by Approximately 1/2 Log-Unit with Inversion, and Is Unaffected by Amblyopia. *Vision Research*, *47*, 1209–1214.
- Neri, P. (2010b). Stochastic Characterization of Small-Scale Algorithms for Human Sensory Processing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *20*(4), 045118. <https://doi.org/10.1063/1.3524305>
- Neri, P. (2011). Coarse to Fine Dynamics of Monocular and Binocular Processing in Human Pattern Vision. *Proceedings of the National Academy of Sciences*, *108*(26), 10726–10731. <https://doi.org/10.1073/pnas.1101246108>
- Neri, P. (2015). The Elementary Operations of Human Vision Are Not Reducible to Template-Matching. *PLoS Computational Biology*, *11*(11), e1004499. <https://doi.org/10.1371/journal.pcbi.1004499>

## Supplementary bibliography

- Oliva, A., & Torralba, A. (2006). Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in Brain Research*, 155, 23–36.
- O'Reilly, C., & Plamondon, R. (2009). Development of a Sigma-Lognormal Representation for On-line Signatures. *Pattern Recognition*, 42(12), 3324–3337. <https://doi.org/10.1016/j.patcog.2008.10.017>
- Plamondon, R., Djioua, M., & O'Reilly, C. (2008). La théorie cinématique des mouvements humains rapides: développements récents. *Traitement du Signal, Numéro Spécial: Le Document Écrit*, 26.
- Quiroga, R. Q., & Pedreira, C. (2011). How Do We See Art: An Eye-Tracker Study. *Frontiers in Human Neuroscience*, 5. <https://doi.org/10.3389/fnhum.2011.00098>
- Rodriguez, C. S., Lech, M., & Pirogova, E. (2018). Classification of Style in Fine-Art Paintings Using Transfer Learning and Weighted Image Patches. *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 1–7. <https://doi.org/10.1109/ICSPCS.2018.8631731>
- Rosen, M., Weibel, P., Fritz, D., Gattin, M., Joanneum), L. ( L., & für Kunst und Medientechnologie Karlsruhe, Z. (Eds.). (2011). *A Little Known Story About a Movement, a Magazine and the Computer's Arrival in Art: New Tendencies and Bit International, 1961-1973*. ZKM/Center for Art and Media ; MIT Press  
OCLC: ocn676725801.
- Sayim, B., & Cavanagh, P. (2011). What Line Drawings Reveal About the Visual Brain. *Frontiers in Human Neuroscience*, 5. <https://doi.org/10.3389/fnhum.2011.00118>
- Schepman, A., Rodway, P., Pullen, S. J., & Kirkham, J. (2015). Shared Liking and Association Valence for Representational Art but Not Abstract Art. *Journal of Vision*, 15(5), 11. <https://doi.org/10.1167/15.5.11>
- Schmidhuber, J. (1997). Low-Complexity Art. *Leonardo*, 30(2), 97. <https://doi.org/10.2307/1576418>
- Seidel, H.-P. (1991). Geometrically Continuous Cubic Bézier Curves. *Graphics Gems II* (pp. 428–434). Elsevier. <https://doi.org/10.1016/B978-0-08-050754-5.50094-3>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016, October 7). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. <http://arxiv.org/abs/1610.02391>
- Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annu Rev Vis Sci*, 5, 399–426.
- Shene, C.-K. (1997). *Introduction to Computing with Geometry*. <http://pages.mtu.edu/~shene/COURSES/cs3621/NOTES/>
- Snapper, L., Oranç, C., Hawley-Dolan, A., Nissel, J., & Winner, E. (2015). Your Kid Could Not Have Done That: Even Untutored Observers Can Discern Intentionality and Structure in Abstract Expressionist Art. *Cognition*, 137, 154–165. <https://doi.org/10.1016/j.cognition.2014.12.009>
- Specker, E., Forster, M., Brinkmann, H., Boddy, J., Immelmann, B., Goller, J., Pelowski, M., Rosenberg, R., & Leder, H. (2020). Warm, Lively, Rough? Assessing Agreement on Aesthetic Effects of Artworks (R. T. H. Ho, Ed.). *PLOS ONE*, 15(5), e0232083. <https://doi.org/10.1371/journal.pone.0232083>
- Spiller, J. (1974). Paul Klee : la pensée créatrice. *Communication & Langages*, 21(1), 18–34. <https://doi.org/10.3406/colan.1974.4071>
- Spillmann, L. (1971). Foveal Perceptive Fields in the Human Visual System Measured with Simultaneous Contrast in Grids and Bars. *Pflugers Archiv (European Journal of Physiology)*, 326, 281–299.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015, April 13). *Striving for Simplicity: The All Convolutional Net*. <http://arxiv.org/abs/1412.6806>
- Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The Role of Holistic Processing in Face Perception: Evidence from the Face Inversion Effect. *Vision Res.*, 51(11), 1273–1278.

- Tayebi Arasteh, S., & Kalisz, A. (2021). Conversion Between Cubic Bezier Curves and Catmull–Rom Splines. *SN Computer Science*, 2(5), 398. <https://doi.org/10.1007/s42979-021-00770-x>
- Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter Perception Emerges from Unsupervised Deep Learning and Recycling of Natural Image Features. *Nature Human Behaviour*, 1(9), 657–664. <https://doi.org/10.1038/s41562-017-0186-2>
- Thurstone, L. L. (1927c). Three Psychophysical Laws. *Psychological Review*, 34(6), 424–432. <https://doi.org/10.1037/h0073028>
- Truong, N., Yuksel, C., & Seiler, L. (2020). Quadratic Approximation of Cubic Curves. *Proc. ACM Comput. Graph. Interact. Tech.*, 3(2). <https://doi.org/10.1145/3406178>
- Ullman, S. (1996). *High-Level Vision*. Cambridge, MA: MIT Press.
- Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017, November 29). *Deep Image Prior*. <http://arxiv.org/abs/1711.10925>
- Vacher, J., Davila, A., Kohn, A., & Coen-Cagli, R. (2020). Texture Interpolation for Probing Visual Perception. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 22146–22157). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2020/file/fba9d88164f3e2d9109ee770223212a0-Paper.pdf>
- Valentine, T. (1988). Upside-down Faces: A Review of the Effect of Inversion Upon Face Recognition. *British Journal of Psychology*, 79 ( Pt 4), 471–491.
- Vallortigara, G., & Regolin, L. (2006). Gravity Bias in the Interpretation of Biological Motion by Inexperienced Chicks. *Curr. Biol.*, 16, R279–280.
- Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually Inexperienced Chicks Exhibit Spontaneous Preference for Biological Motion Patterns. *PLoS Biol.*, 3(7), e208.
- VanRullen, R. (2017). Perception Science in the Age of Deep Neural Networks. *Front Psychol*, 8, 142.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention Is All You Need*. <http://arxiv.org/abs/1706.03762>
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017, October 30). *Graph Attention Networks*. <http://arxiv.org/abs/1710.10903>
- Verostko, R. (2010). Form, Grace and Stark Logic: 30 Years of Algorithmic Drawing. *Leonardo*, 43(3), 230–231. <https://doi.org/10.1162/leon.2010.43.3.230>
- Vessel, E. A. (2010). Beauty and the Beholder: Highly Individual Taste for Abstract, but Not Real-World Images. *Journal of Vision*, 10(2), 1–14. <https://doi.org/10.1167/10.2.18>
- Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger Shared Taste for Natural Aesthetic Domains Than for Artifacts of Human Culture. *Cognition*, 179, 121–131. <https://doi.org/10.1016/j.cognition.2018.06.009>
- von Petzinger, G. (2005). *Making the Abstract Concrete: The Place of Geometric Signs in French Upper Paleolithic Parietal Art*. University of Victoria.
- Williams, A., Srnicek, N., & Citton, Y. (2014). Manifeste accélérationniste. *Multitudes*, (56), 23–35. <https://doi.org/10.3917/mult.056.0023>
- Xie, N., Hachiya, H., & Sugiyama, M. (2013). Artist Agent: A Reinforcement Learning Approach to Automatic Stroke Generation in Oriental Ink Painting. *IEICE Transactions on Information and Systems*, E96.D(5), 1134–1144. <https://doi.org/10.1587/transinf.E96.D.1134>
- Xie, N., Zhao, T., Tian, F., Zhang, X., & Sugiyama, M. (2015). Stroke-based Stylization Learning and Rendering with Inverse Reinforcement Learning. *Proceedings of the 24th International Conference on Artificial Intelligence*, 2531–2537. <http://dl.acm.org/citation.cfm?id=2832581.2832603>
- Yovel, G., & Kanwisher, N. (2005). The Neural Basis of the Behavioral Face-Inversion Effect. *Current Biology*, 15, 2256–2262.

## Supplementary bibliography

- Zhang, X.-Y., Yin, F., Zhang, Y.-M., Liu, C.-L., & Bengio, Y. (2016, June 21). *Drawing and Recognizing Chinese Characters with Recurrent Neural Network*. <http://arxiv.org/abs/1606.06539>
- Zhao, J., Wang, L., Wang, Y., Weng, X., Li, S., & Jiang, Y. (2014). Developmental Tuning of Reflexive Attentional Effect to Biological Motion Cues. *Sci Rep*, *4*, 5558.
- Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017, March 30). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. <http://arxiv.org/abs/1703.10593>

## List of figures

0.1	Photograms from a short animation of a painting . . . . .	2
0.2	Drawings of low complexity . . . . .	4
0.3	Drawings of medium complexity . . . . .	4
0.4	Drawings of high complexity . . . . .	5
1.1	Standard ratios and construction lines . . . . .	17
1.2	Terrestrial and celestial compositions . . . . .	20
1.3	Weight and balance of graphical elements . . . . .	21
1.4	Morphogenesis . . . . .	28
1.5	Composition alterations . . . . .	32
1.6	Linear and final compositions by Wassily Kandinsky . . . . .	33
1.7	Continuous space . . . . .	41
1.8	Hyper-dimensional reconstruction . . . . .	43
1.9	Vectorial color spaces . . . . .	45
1.10	Probabilistic space . . . . .	48
2.1	Raw drawings among notes . . . . .	56
2.2	Two samples of excluded drawings . . . . .	58
2.3	Figurative drawings . . . . .	59
2.4	Interaction of a line with the B.P. . . . .	59
2.5	Central composition . . . . .	61
2.6	Diversity of pen, paper, and support in the dataset . . . . .	63
2.7	Colored drawings to grayscale . . . . .	64
2.8	Close-up of scanned drawings . . . . .	66
2.9	Binary map processing . . . . .	67
2.10	Cleaning of binary maps . . . . .	68
2.11	Surface-to-line considerations . . . . .	69
2.12	Skeletonization . . . . .	71
2.13	Skeleton disentanglement . . . . .	73
2.14	Polylines from RDP algorithm . . . . .	75
2.15	<i>Sigma Lognormal</i> parameterization . . . . .	76
2.16	Tensions of the line and the curve . . . . .	77
2.17	Imaginary trajectories associated with arrows . . . . .	78
2.18	Bézier Curves and Bernstein polynomials . . . . .	79
2.19	Geometric continuity . . . . .	82
2.20	Fitting difficulties associated with quadratic Bézier curves . . . . .	83



## List of figures

2.21	Cubic Bézier curves . . . . .	84
2.22	Dataset application . . . . .	87
2.23	Spatial standardization . . . . .	89
2.24	Spatial standardization, examples . . . . .	89
2.25	Stroke simplification . . . . .	91
2.26	Splitting of long strokes . . . . .	93
2.27	Stroke encoding details . . . . .	93
2.28	Composition length limitation . . . . .	93
2.29	Composition permutations . . . . .	95
2.30	Composition permutations, gallery . . . . .	96
3.1	Simple generative RNN architecture . . . . .	111
3.2	Simple VAE architecture . . . . .	113
3.3	Stroke model architecture . . . . .	117
3.4	Conditioning length ratio, exponential decay . . . . .	121
3.5	Composition model architecture . . . . .	123
3.6	Compositional plane model architecture . . . . .	126
3.7	Leveraging $\mathcal{L}_{D_{\text{KL}}}$ . . . . .	132
3.8	Balancing model resources, length weighting matrices . . . . .	137
3.9	Balancing model resources, smoothing function . . . . .	137
3.10	Probabilistic vs. deterministic optimization scenarios . . . . .	140
3.11	Output variance limiting function . . . . .	141
3.12	Table of layer sizes . . . . .	145
4.1	Training logs, stroke model . . . . .	150
4.2	Training logs, composition model . . . . .	151
4.3	Training logs and score distributions, compositional plane model . . . . .	152
4.4	Score distributions, stroke model . . . . .	154
4.5	Score distributions, composition model . . . . .	154
4.6	Scores w.r.t the dimensionality $K$ of $\mathbf{z}$ , stroke model . . . . .	155
4.7	Scores w.r.t the dimensionality $K$ of $\mathbf{z}$ , composition model . . . . .	155
4.8	Reconstruction of strokes . . . . .	157
4.9	Reconstruction of strokes in a compositional context . . . . .	157
4.10	Reconstruction of compositions . . . . .	158
4.11	Compositional plane model predictions . . . . .	159
4.12	Latent space distribution, stroke model . . . . .	159
4.13	Latent space distribution, composition model . . . . .	160
4.14	Generation of strokes . . . . .	161
4.15	Generation of strokes, local explorations . . . . .	161
4.16	Generation of compositions . . . . .	162
4.17	Generation of compositions, local explorations . . . . .	163
4.18	Generation of compositions, compositional plane model . . . . .	164
4.19	Generation of compositions, comp. plane model, family explorations . . . . .	165
4.20	Hyperspace density . . . . .	167

4.21	Counter-intuitive hyperspace characteristics . . . . .	168
4.22	Stroke and composition norm distributions . . . . .	169
4.23	Linear interpolation . . . . .	170
4.24	Linear interpolations per dimension of the stroke model . . . . .	170
4.25	Spherical linear interpolation . . . . .	172
4.26	<i>lerp</i> vs <i>slerp</i> stroke interpolations . . . . .	172
4.27	<i>lerp</i> vs <i>slerp</i> composition interpolations . . . . .	173
4.28	<i>slerp</i> quad-interpolation of strokes . . . . .	174
4.29	<i>slerp</i> quad-interpolation of compositions . . . . .	175
4.30	Measurements, stroke model . . . . .	177
4.31	Measurements, composition model . . . . .	178
4.32	Measurements, compositional plane model (1) . . . . .	183
4.33	Measurements, compositional plane model (2) . . . . .	184
5.1	Rotations in the latent space . . . . .	188
5.2	Difference threshold and psychometric function . . . . .	190
5.3	Similarity judgment threshold estimation . . . . .	192
5.4	Unsigned angle distribution between random samples . . . . .	193
5.5	Multidimensional scaling . . . . .	194
5.6	Isomap . . . . .	195
5.7	Perceptual Scaling . . . . .	197
5.8	Tuning sensitivity . . . . .	201
5.9	MLDS task setups . . . . .	201
5.10	MLDS pair measurement normality verification . . . . .	203
5.11	Fitting methods comparison . . . . .	206
5.12	Theoretical local discriminative probabilities . . . . .	209
5.13	Optimal $\tau$ , number of $\sigma$ per step size . . . . .	210
5.14	TS and MLDS combination constraints . . . . .	211
5.15	PMLDS combination constraints . . . . .	212
5.16	Theoretical perceptual scale . . . . .	215
5.17	Stimuli generation, <i>circle</i> condition . . . . .	216
5.18	Experimental stimuli . . . . .	218
5.19	Online experimental platform . . . . .	219
5.20	Perceptual scaling results, <i>norm</i> condition . . . . .	222
5.21	Perceptual scaling results, <i>circle</i> condition . . . . .	223
5.22	Perceptual distortions and resamplings, <i>norm</i> condition . . . . .	224
5.23	Perceptual distortions and resamplings, <i>circle</i> condition . . . . .	225
5.24	Fitted perceptual noise standard deviation . . . . .	227
5.25	Distribution of response probabilities . . . . .	228
5.26	Perceptual scale prediction from Fisher information . . . . .	232
6.1	Pen-plotter . . . . .	240
6.2	Reproducing a drawing at different scales . . . . .	243
6.3	Offsetting lines and filling surfaces . . . . .	243

## List of figures

6.4	Pen thickness correction and different filling intervals . . . . .	244
6.5	Calligraphic strokes . . . . .	245
6.6	Stroke profiles . . . . .	247
6.7	Stroke profile I on generated compositions . . . . .	248
6.8	Stroke profile II on generated compositions . . . . .	249
6.9	Brush plots with the stroke profile II . . . . .	250
6.10	Constant depth brush plot . . . . .	251
6.11	Trial and error with brush plots . . . . .	251
6.12	Grids . . . . .	255
6.13	Generated compositions over a homogeneous circular grid . . . . .	256
6.14	Dynamics – Uncertainty – 1 . . . . .	259
6.15	Dynamics – Uncertainty – 2 . . . . .	260
6.16	Dynamics – Thickness – 1 . . . . .	261
6.17	Dynamics – Thickness – 2 . . . . .	262
6.18	Dynamics – Transition – 1 . . . . .	263
6.19	Dynamics – Transition – 2 . . . . .	264
6.20	Dynamics – Transition – 3 . . . . .	265
6.21	Dynamics – Trace – 1 . . . . .	266
6.22	Dynamics – Trace – 2 . . . . .	267
A.1	Photographs of the exhibition venue . . . . .	306
A.2	Exhibition catalogue . . . . .	307
A.3	Accumulation . . . . .	308
A.4	Individuality . . . . .	309
A.5	Morphology . . . . .	310
A.6	Combinations . . . . .	311
A.7	Cartography . . . . .	312
A.8	Exhibition postcards . . . . .	313
A.9	Digital renderings of series I . . . . .	316
A.10	Digital renderings of series II . . . . .	317
A.11	Photographs of the exhibition . . . . .	318
A.12	Photographs of the exhibition . . . . .	319
A.13	Photographs of the exhibition . . . . .	320
A.14	Photograph of the exhibition . . . . .	321
A.15	Exhibition poster . . . . .	322

## List of algorithms

2.1	Drawing to binary map . . . . .	97
2.2	Binary map to individual lines . . . . .	98
2.3	Composite cubic Bézier curves fitting . . . . .	99
2.4	Composite cubic Bézier curves simplification . . . . .	101
2.5	Splitting of long strokes . . . . .	102
2.6	Composition permutations . . . . .	102
3.1	Backward Clamp . . . . .	141
5.1	PMLDS3 combinations and periodicity correction vector . . . . .	213
6.1	Stroke profile I . . . . .	247
6.2	Stroke profile II . . . . .	247
6.3	Line stroking . . . . .	252





## RÉSUMÉ

---

La composition picturale, entendue comme la disposition des éléments graphiques sur le plan, est généralement associée à des règles qualitatives et des heuristiques. Bien qu'instructives pour les artistes et leur pratique, ces normes n'agissent que comme des contraintes externes sur le plan. Nous pensons que les œuvres d'art sont capables de fixer des caractéristiques de composition plus fondamentales dans leur matière picturale même. Nous développons donc un paradigme supposant toutes les œuvres d'un-e artiste comme les vues partielles d'une représentation en plus grandes dimensions, agrégeant des régularités compositionnelles intrinsèques. Nous choisissons de matérialiser cet objet hyper-compositionnel théorique par un espace continu, vectoriel et probabiliste. Notre objectif est de rendre ces régularités explicites pour un usage artistique et d'établir des mesures quantitatives pour des études scientifiques. Notre recherche s'inscrit donc pleinement dans un programme réflexif de recherche-crédation; fondé à la fois sur un matériau artistique personnel, riche d'une pratique de plus de 10 ans de la composition abstraite; et sur une approche interdisciplinaire projective, combinant une modélisation itérative par apprentissage automatique et des vérifications perceptives avec de la psychophysique. La nature séquentielle et non stationnaire du processus de composition, ainsi que la définition complexe et évolutive de ses unités fonctionnelles sous-jacentes, se combinent en un phénomène perceptif qui ne se modélise pas facilement par les modèles d'apprentissage profond basés sur des pixels, e.g. CNNs. Nous adoptons une stratégie différente, construite autour d'une définition paramétrique de l'exécution des traits, et de RNN-VAEs (Recurrent Variational Auto-Encoders) imbriqués hiérarchiquement, permettant à notre modèle d'aborder la matière picturale en alignant son comportement sur le geste artistique. Plus précisément, cette architecture extrait les régularités compositionnelles en compressant les dessins en un nombre réduit de dimensions indépendantes, alignées dans l'idéal sur la représentation intérieure construite par les artistes et les observateurs. Ces réseaux neuronaux artificiels sont entraînés sur plus de 5000 compositions abstraites personnelles et vectorisées par des courbes de Bézier. Bien que cet ensemble de données soit important pour un seul artiste, son échelle reste relativement réduite pour l'entraînement de réseaux profonds. Nous abordons cette problématique en introduisant de nouvelles contraintes qui encouragent un espace latent à la fois compact, cohésif et expressif. Nous étudions ensuite l'espace compositionnel résultant à travers des jugements perceptifs de trajectoires interpolées entre des points précis de cet espace. Nous vérifions particulièrement l'homogénéité de la densité latente en mesurant l'échelle perceptive produite par des participants humains jugeant la similarité entre des compositions. Nous limitons notre exploration à des coupes circulaires d'hypersphères, dont la densité latente est relativement stable, et des progressions linéaires orthogonales le long de la norme, provoquant des distorsions perceptives plus importantes. Nous utilisons une variante de la méthode MLDS, que nous avons restreinte à des triplets locaux et étendue aux espaces physiques périodiques. Les échelles perceptives mesurées empiriquement présentent des régularités qui sont capturées de manière satisfaisante par la notion d'information de Fisher calculée à partir des métriques fournies par le modèle. Les algorithmes qui en résultent permettent aux artistes d'explorer l'interaction dynamique des éléments graphiques en fonction non seulement de leurs propres régularités de composition, mais aussi des régularités perceptives intrinsèques de ceux qui voient leur art. Nous terminons enfin ce cycle en révélant les dimensions compositionnelles cachées avec de l'encre et du papier, via des créations par traceur numérique.

## MOTS-CLÉS

---

Art Pictural, Composition, Abstraction, Recherche-Création, Modélisation, Apprentissage Machine, Réseaux de Neurones, Apprentissage Profond, RNN, VAE, Perception, Psychophysique, MLDS

## ABSTRACT

---

Pictorial composition, understood as the arrangement of graphical elements on the plane, is typically associated with qualitative rules and heuristics. Although informative for artists and their practice, these norms and guidelines only act as external constraints on the canvas. We believe that artworks are able to fix more fundamental compositional features in their pictorial matter. We therefore develop a paradigm in which every artwork of an artist represents a partial view of a higher-dimensional representation, aggregating intrinsic compositional regularities. We choose to materialize this theoretical hyper-compositional object as a continuous, vectorial and probabilistic space. Our objective is to make regularities explicit for artistic purposes, and to build quantitative metrics for scientific scrutiny. Our research is therefore articulated around a reflexive research-creation agenda: it is grounded in personal artistic material drawing from more than 10 years of practice in abstract composition, it expands along a projective interdisciplinary framework that combines iterative modeling with machine learning, and it engages in perceptual validation using psychophysical techniques. The sequential non-stationary nature of the compositional process, together with the complex and evolving definitions of its underlying functional units, coalesce into a perceptual phenomenon that cannot be readily modeled through pixel-based deep learning models, such as CNNs. We adopt a different strategy, constructed around a parametric definition of stroke execution and hierarchically nested RNN-VAEs (Recurrent Variational Auto-Encoders), enabling our network to tackle pictorial material by aligning its behavior with the artistic gesture. More specifically, this network architecture extracts compositional regularities by compressing inputs into a reduced number of independent dimensions, ultimately aligned with the representation entertained by artists and observers. These artificial neural networks are trained on >5k personal abstract compositions vectorized as Bézier curves. Although this dataset is large for a single artist, its scale remains relatively small for training large networks. We address this issue by introducing new constraints that support a compact latent space that is both cohesive and expressive. We then study the resulting compositional space through perceptual judgments of interpolated trajectories spanning targeted locations within this space. In particular, we characterize latent density homogeneity by measuring the perceptual scale adopted by human participants when judging sample similarity. We limit our exploration to circular slices of hyperspheres, along which latent density can be regarded as reasonably stable, and orthogonal linear progressions along the norm, which imply larger perceptual distortions. We employ a variant of the MLDS method, which we have restricted to local triplets and extended to periodic physical spaces. The empirically measured perceptual scales present regularities that are satisfactorily captured by the notion of Fisher information computed on metrics provided by the model. The resulting algorithms enable artists to explore the dynamical interaction of graphical elements in accordance not only with their own compositional regularities, but also with the perceptual regularities intrinsic to those who view their art. We then come full circle by revealing the hidden compositional dimensions with ink and paper through digitally pen-plotted creations.

## KEYWORDS

---

Pictorial Art, Composition, Abstraction, Research-Creation, Modeling, Machine Learning, Neural Networks, Deep Learning, RNN, VAE, Perception, Psychophysics, MLDS