



HAL
open science

Improving Image Quality using High Dynamic Range and Aesthetics Assessment

Mathieu Chambe

► **To cite this version:**

Mathieu Chambe. Improving Image Quality using High Dynamic Range and Aesthetics Assessment. Image Processing [eess.IV]. Univ Rennes, 2023. English. ⟨NNT : ⟩. ⟨tel-04187749⟩

HAL Id: tel-04187749

<https://hal.science/tel-04187749v1>

Submitted on 25 Aug 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.



Distributed under a Creative Commons CC BY-ND 4.0 - Attribution - No Derivative Works - International License

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601

*Mathématiques, Télécommunications, Informatique, Signal, Systèmes,
Électronique*

Spécialité : *Informatique*

Par

Mathieu CHAMBE

Improving Image Quality using High Dynamic Range and Aesthetics Assessment

Thèse présentée et soutenue à Rennes, le 16 Juin 2023

Unité de recherche : IRISA UMR 6074

Rapporteurs avant soutenance :

Frédéric DUFAUX Directeur de Recherche, Université Paris-Saclay

Karol MYSZKOWSKI Senior Researcher, MPI Informatik

Composition du Jury :

Présidente : Élixa FROMONT Professeur des Universités, Université de Rennes

Examineurs : Frédéric DUFAUX Directeur de Recherche, Université Paris-Saclay

Daniel MÉNEVEAUX Professeur des Universités, Université de Poitiers

Karol MYSZKOWSKI Senior Researcher, MPI Informatik

Christophe SCHLICK Professeur des Universités, Université de Bordeaux

Dir. de thèse : Zoltan MIKLOS Maître de Conférences HDR, Université de Rennes

Co-dir. de thèse : Rémi COZOT Professeur des Universités, Université Côte d'Opale

Kadi BOUATOUCH Professeur émérite, Université de Rennes

TABLE OF CONTENTS

Résumé en Français	5
1 Introduction	11
2 General notions	15
2.1 Introduction	15
2.2 Human Visual Perception	15
2.2.1 The eye	15
2.2.2 Radiometry and Photometry	17
2.2.3 High Dynamic Range imaging	18
2.2.4 Tone Mapping and Inverse Tone Mapping	19
2.2.5 Metrics	21
2.3 Aesthetics	21
2.3.1 Definition	22
2.3.2 Application in computer science	23
2.4 Conclusion	24
3 Related work	25
3.1 Aesthetics assessment	25
3.1.1 Prediction models	25
3.1.2 Aesthetics assessment datasets	27
3.2 High Dynamic Range Imaging	31
3.2.1 HDR datasets	31
3.2.2 HDR Image Generation	32
3.2.3 Our approach	34
4 Assessing aesthetics using professional photographs	37
4.1 Introduction	37
4.2 Testing models and datasets	39
4.2.1 NIMA model	40

TABLE OF CONTENTS

4.2.2	Ranking Network model	41
4.2.3	Datasets of professional photography	41
4.2.4	Overview of models	43
4.3	Presentation of experiments	45
4.3.1	Hypotheses	45
4.3.2	Fine-tuning of Nasnet-based model	45
4.4	Results	46
4.4.1	Statistical analysis	46
4.4.2	Does fine-tuning Nasnet-based NIMA model improve the overall prediction capabilities?	48
4.4.3	Comparison with weighted CNN	49
4.4.4	Discussion	52
4.5	Conclusion	52
5	HDR-LFNet: Inverse Tone Mapping by Fusion Methods	55
5.1	Introduction	55
5.2	Our fusion network	57
5.2.1	Overview of 3D convolutions	58
5.2.2	Architecture	59
5.2.3	Loss function	63
5.2.4	Postprocessing	64
5.2.5	Training dataset	65
5.2.6	Choice of input iTMO	67
5.3	Results	69
5.3.1	Implementation details	69
5.3.2	State-of-the-art comparison	69
5.3.3	Ablation study	76
5.3.4	Subjective user study	78
5.4	Conclusion	81
6	Conclusion	83
	Bibliography	87

RÉSUMÉ EN FRANÇAIS

La quantité de contenu visuel, que ce soient des images ou des vidéos, est en constante augmentation ces dernières années. En effet, nous estimons qu'environ 300 heures de vidéo sont mises en ligne chaque minute sur le site internet YouTube dans le monde entier. En prenant en compte les images ainsi que les contenus hors ligne, la quantité réelle de contenu visuel est donc bien supérieure à ces 300 heures estimées. Ce grand volume conduit à poser des questions et des problématiques dans différents domaines : comment stocker une telle quantité de données ? Est-ce qu'il est possible de créer des requêtes efficaces pour interroger des bases de données visuelles de grandes tailles ? Et la question qui nous intéresse dans cette thèse : *Y a-t-il des méthodes pour améliorer automatiquement la qualité de contenus visuels ?* En effet, il est important non seulement de créer du contenu, mais surtout de créer du contenu de bonne qualité. Cela soulève une autre question : qu'est-ce qu'une image ou une vidéo de «bonne qualité» ? Selon la réponse que l'on donne à cette question, il existe plusieurs outils qui permettent d'évaluer la qualité (des métriques objectives, basées sur des calculs numériques ; ou bien des études utilisateurs, où les avis d'observateurs extérieurs sont agrégés pour calculer la qualité d'une image) ou bien d'améliorer la qualité (grâce à des méthodes basées matériel ou logiciel).

Dans cette thèse, nous proposons deux définitions différentes de la «qualité d'image». La première définition consiste à considérer les images comme des signaux en deux dimensions. Il y a plusieurs façons de définir ce qu'est un bon signal. De fait, il existe plusieurs façons d'améliorer la qualité d'un signal. Par exemple, utiliser un algorithme de débruitage permet d'avoir un signal plus propre. Dans notre situation, nous considérerons que les images de bonne qualité sont les images avec une grande précision. Il existe deux formes de précision pour les images : la précision spatiale (et dans ce cas, une image avec une grande précision spatiale correspond à une image de haute résolution) et la précision spectrale, à laquelle nous nous intéressons ici. Pour les images numériques, la précision spectrale correspond à la précision des valeurs des pixels. Les images avec une grande précision spectrale ont un meilleur contraste, et un ensemble plus grand de valeurs potentielles de luminosité. De telles images sont appelées images HDR (grande gamme dynamique, *High Dynamic Range* en anglais), en contraste avec les images communément

répandues et qualifiées de SDR (pour *Standard Dynamic Range*, gamme dynamique standard). La problématique principale est alors de générer de telles images HDR. Plusieurs méthodes ont été développées pour générer de nouvelles images HDR. La méthode que nous proposons ici est appelée opérateur de correspondance inverses de couleurs (*Inverse Tone Mapping Operator*, iTMO en anglais). Le but des méthodes dites iTMO est de transformer une image SDR en image HDR. Pour cela, nous entraînons un modèle statistique pour associer efficacement une image HDR à une image SDR donnée. Les principaux modèles statistiques utilisés dans la littérature sont les réseaux de neurones. Les réseaux de neurones contiennent un certain nombre de paramètres qui sont modifiés pendant l'entraînement. Ces paramètres sont ensuite utilisés pour calculer plusieurs représentations de l'image d'entrée, et finalement, ces représentations sont agrégées en une unique image de sortie. Le nombre de paramètres du réseau influe directement sur ses performances : pour améliorer les performances du réseau, il est possible d'augmenter le nombre de paramètres dont il a besoin. Les iTMOs récents sont basés sur des réseaux de neurones qui contiennent aux alentours d'un million de paramètres, mais certains réseaux dépassent les dix millions. Augmenter le nombre de paramètres affine les représentations intermédiaires, et donc améliore la qualité de l'image finale, mais augmente le nombre d'images exemples nécessaires pour avoir une performance correcte, ainsi que la consommation de ressources de manière générale (temps d'entraînement, espace mémoire, temps de traitement).

Notre première contribution est une nouvelle méthode iTMO appelée HDR-LFNet. Notre méthode a comme but d'être aussi performante que les méthodes de l'état de l'art, tout en consommant significativement moins de ressources. Pour ce faire, nous avons conçu un réseau de neurones qui fusionne plusieurs images HDR en une image HDR. Cette transformation étant plus simple que la transformation de SDR en HDR, le nombre de paramètres nécessaires pour le réseau de neurones est ainsi réduit. Pour traiter une image SDR donnée, il faut donc en premier lieu l'augmenter de plusieurs manières différentes. Nous utilisons plusieurs iTMOs existants qui ne sont pas basés sur des réseaux de neurones, afin de ne pas augmenter la complexité de notre méthode. La fusion de ces différentes méthodes grâce à un réseau de neurones permet d'obtenir une image HDR de meilleure qualité que chaque image HDR en entrée du réseau. L'image obtenue en sortie du réseau est à nouveau traitée (augmentation du contraste et ajout de couleurs) pour obtenir une image HDR aussi proche de la réalité que possible. En plus de proposer une implémentation de notre méthode, nous mettons à disposition la base de données d'images HDR utilisée pour entraîner notre réseau de neurones. Il s'agit d'une base de données de plus de 490 images

HDR en très haute résolution pouvant être utilisée à toutes fins utiles. Enfin, notre étude comporte une évaluation de notre méthode par rapport aux méthodes existantes dans l'état de l'art. Nous montrons grâce à une batterie de métriques objectives et une étude subjective que notre méthode atteint des performances similaires à l'état de l'art, mais consomme beaucoup moins de ressources.

Au lieu de considérer les images comme de simples signaux à deux dimensions, une autre solution est de les considérer comme une unité visuelle atomique. Nous nous intéressons alors à l'avis d'observateurs humains à propos de cette image. Dans cette situation, il est impossible d'avoir une mesure de qualité purement objective : la qualité va fortement dépendre de l'observateur, de ses goûts, de l'environnement ambiant (luminosité de la pièce, dimensions de l'écran, etc...) et d'un certain nombre d'autres facteurs qui empêchent la modélisation parfaite de la situation. Notre intérêt principal ici est l'interaction qui existe entre l'observateur et l'objet considéré (ici, une image). De cette interaction naît la notion d'esthétique : ainsi, une image de bonne qualité sera une image de bonne qualité esthétique. Il faut noter que la qualité en tant que signal et la qualité esthétique sont deux notions complètement différentes et indépendantes (la Figure 1 illustre cette différence). Cela explique pourquoi une nouvelle étude de la qualité esthétique est importante. Plusieurs algorithmes permettant de mesurer la qualité esthétique d'images ont déjà été proposés. La grande majorité des méthodes existantes sont basées sur des techniques d'apprentissage supervisé. Un des facteurs importants d'un algorithme basé sur l'apprentissage supervisé est la base de données d'entraînement. La base de données la plus utilisée pour l'évaluation d'esthétique est *AVA (Aesthetics Visual Analysis [MMP12])*. Nous souhaitons donc évaluer la pertinence de cette base de données.

Notre seconde contribution consiste à évaluer *AVA*. Nous avons collecté plusieurs photographies professionnelles de différentes sources, afin d'obtenir une diversité de contenus et d'intentions artistiques. Ces six catégories de photographies (Mode ; Automobile ; Guerre ; Nature ; Architecture ; Sport) contiennent aux alentours de 100 photographies (exceptée la catégorie Mode qui en contient plus de 1.000) d'esthétiques différentes. En utilisant ces photographies professionnelles, nous voulons étudier les performances de modèles d'esthétique entraînés sur *AVA*. Les deux modèles que nous considérons ici sont *NIMA [TM18a]* et le *Ranking Network [Kon+16a]*. Notre étude montre ainsi qu'il existe des types d'image qui ne sont pas représentés dans *AVA*. L'impact de ce manque de diversité est surtout visible sur *NIMA* qui est un modèle générique, entraîné en premier lieu sur une autre tâche que l'évaluation d'esthétique. Nous remarquons que les scores



FIGURE 1 – La qualité d’image et la qualité esthétique sont deux notions différentes. (gauche) *La Suerte de Capa*, Ernst Haas (1956) : la faible netteté de l’image est une intention du photographe. Cela réduit la qualité de l’image, mais la photographie en elle-même est considérée de bonne qualité esthétique ; (droite) une image provenant de la base de données AVA [MMP12] avec un score esthétique moyen de 4.95/10. Même si l’image n’a pas d’erreur flagrante, le score esthétique est moyen.

de NIMA ne sont pas différents d’une catégorie à l’autre, et ont une distribution similaire à la distribution des scores d’AVA. Pour essayer d’améliorer le comportement de NIMA, nous proposons d’augmenter la base de données d’entraînement en utilisant les images professionnelles collectées au préalable. Afin de conserver au maximum les performances du modèle sur AVA et de réduire le temps nécessaire à l’entraînement, nous avons décidé d’utiliser une technique de raffinement en entraînant le réseau une seconde fois. Nous montrons que notre technique améliore les performances de NIMA sur les images professionnelles tout en conservant les performances sur AVA.

Pour conclure, nous proposons deux différentes contributions dans cette thèse. Tout d’abord, nous souhaitons évaluer la pertinence d’AVA, la principale base de données annotée utilisée pour les modèles de prédiction d’esthétique. Pour cela, nous collectons plusieurs sortes d’images professionnelles afin d’évaluer des modèles entraînés avec AVA. Comme nous observons un manque de généralisation pour certains modèles, nous avons entraîné ces modèles avec de nouvelles photographies, et nous montrons que notre technique de raffinement est une solution pour améliorer la portée des modèles. Dans un second temps, nous proposons un nouvel algorithme d’augmentation de la gamme dynamique (iTMO) appelé HDR-LFNet. Notre nouvel iTMO contient beaucoup moins de paramètres, mais a des performances similaires aux méthodes de l’état de l’art.

Le travail présenté ici peut se poursuivre de plusieurs manières. Tout d’abord, notre

travail soutient la conception d'une nouvelle méthode d'évaluation automatique de l'esthétique. Les deux points à traiter en premier lieu sont la conception d'une architecture spécifique à l'évaluation d'esthétique, et la création d'un ensemble d'images plus riche en contenus qu'AVA. De plus, étant donné que l'esthétique est une notion complexe et subjective, il serait intéressant d'avoir une explication de la décision du modèle. Plusieurs travaux ont déjà exploré la notion d'explicabilité dans le cadre de l'évaluation d'esthétique [Wan+19b ; KVD20 ; Sch+23].

Enfin, il est envisageable de travailler sur d'autres modalités de contenu visuel. Nous pouvons en premier lieu citer la vidéo, dont la principale difficulté est l'aspect temporel. En effet, la méthode naïve qui consisterait à traiter chaque image de la vidéo indépendamment des autres ne fonctionne pas. Dans le cas d'un iTMO, cela peut faire apparaître des artefacts temporels. Il est alors nécessaire de concevoir des modèles adaptés à la vidéo. D'autres formats visuels tels que les images omnidirectionnelles ou bien les images à grand gamut de couleurs sont des formats riches en contenus, similaires aux images HDR ou aux vidéos par certains aspects, pour lesquels il serait possible d'adapter notre travail.

INTRODUCTION

The amount of visual content (images and videos) created has been steadily increasing for the past decade. For example, we can estimate that there are 300 hours of video uploaded to the website YouTube every minute in the world. The true amount of visual content in the entire world is then much larger than that, if we take into account images and offline content. This raises questions and problematics in many different fields: How to store a very large amount of data? Is it possible to design efficient queries in very large image databases? In this thesis, we consider the following question: *Are there automatic methods to improve the quality of visual content?* Indeed, more than creating content, there is a need to create good quality content. However, what does “good quality” mean for an image or a video? Depending on the answer to this question, there exist different tools to measure the quality (either objective metrics, that are based on numerical algorithms; or subjective studies, where the opinion of independent observers are compiled to evaluate the quality of an image) or to improve the quality (using hardware or software solutions).

In our work, we propose two different definitions of “image quality”. The first definition of quality is to consider images as two-dimensional signals. There are many ways to define what is a good signal. Consequently, there are many ways to improve a signal. For example, denoising methods allow to have a cleaner signal. In this thesis, good quality images refer to “precise” images. The precision can be considered in two different domains: in spatial domain (and so a precise image is an image with high resolution) or in frequency domain. We have focused our work on the latter domain. For numeric images, the frequency precision corresponds to the precision with which we store pixel values (number of bits per pixel per channel). Images with great frequency precision have better contrast, and higher brightness values. Such images are called High Dynamic Range (HDR) images. The main problematic is to generate such HDR images. Many methods developed in Chapter 3 have been devised to generate new HDR images. The method that we propose in our work is called inverse tone mapping (iTMO), and is based on expanding a given standard image to high dynamic range. The idea behind such methods

is to train a statistical model to efficiently map a standard image to its HDR counterpart. One of the most used statistical model in the literature is the neural network. It is the main model used in the state-of-the-art for iTMO. Neural networks contain a number of parameters that are modified during the training process. These parameters are used to compute several representations of the input image, that are in turn used to compute the final output. To get better results, recent models have increased the model size in terms of number of parameters (usually around 1 million parameters, but some models require more than 10 million parameters). Increasing the number of parameters allows to get finer representations of images, and so better results in the end. However, this increases the number of images needed to get good representations, as well as the general resources needed to set up and execute such models. We devised a specific, light architecture called HDR-LFNet (which is presented in details in Chapter 5). Our method requires around 400,000 parameters, which is a lower number of parameters than the state-of-the-art. Besides, we show that our method runs faster and needs fewer resources (training images and time, evaluation time) than the state-of-the-art.

Instead of considering images as simple 2D signals, a second solution is to consider them as external stimuli for human beings. In this case, our measure of quality cannot be purely objective and universal. The quality of an image heavily depends on the observer and their taste, the environmental conditions (the room lighting, the dimension of the screen for a digital image, etc...) and many different factors that cannot be perfectly reproduced to get a reproducible result. The main point of interest is the interaction between the observer and the object. From this interaction arises the notion of aesthetics, and a good quality image corresponds to an image with high aesthetic quality. For an image, the quality as a signal and the aesthetic quality are two different notions, and one cannot be entirely computed from the other (an example is shown on Figure 1.1). This explains why studying aesthetic quality independently from signal quality is important. Several methods for automatic aesthetic quality assessment have already been devised, and the large majority of those methods are based on supervised learning techniques. One of the most used dataset for training such methods is called Aesthetics Visual Analysis (AVA) [MMP12]. The large majority of the methods of the state-of-the-art use AVA as training dataset. Therefore, it is important to assess whether AVA is a good training dataset or not. In our work, presented in details in Chapter 4, we show that AVA lacks a diversity of images, and that these images can be included in the training data without impeding on the training performed on the original images. This shows that improving



Figure 1.1 – Image quality and Aesthetics quality are different. (left) *La Suerte de Capa*, Ernst Haas (1956): although the photograph is blurry, the quality degradation serves the artistic message, and so the image is considered of good aesthetic quality; (right) an image from the AVA dataset [MMP12] with an average aesthetic score of 4.95/10. Even though the image has a good focus and no particular technical errors, the aesthetic score is not high.

automatic aesthetics assessment models can be done by improving the training dataset.

To conclude, our first contribution is to assess the relevance of AVA as a training dataset for aesthetics assessment models. Our proposed approach consists in gathering different kinds of professional photographs to evaluate some aesthetics assessment models. As we observed a lack of generalization for some models, we have trained these latter with new photographs, and we have shown that this training effectively improves the range of action of the model. Our second contribution is the design of a new inverse tone mapping operator called HDR-LFNet. Our new iTMO contains fewer trainable parameters, but achieves results of similar quality than the state-of-the-art. Our model is proposed along with its training dataset, a new set of HDR images that may be used by the community.

The rest of this thesis is structured as follows.

Chapter 2 presents some general notions important to understand our contributions. We explain the notion of aesthetics in general and in the specific case of computer science, as well as the concept of High Dynamic Range imaging.

Chapter 3 provides a survey of techniques related to our contributions in High Dynamic Range and Automatic Aesthetics Assessment. We present the different methods that we based our work on, and the datasets and metrics available.

Chapter 4 explains the process that we have devised to evaluate some aesthetics assessment models. Using a kind of photograph different from the training photographs of

aesthetics models, we show that the widely used AVA dataset [MMP12] can be improved.

Chapter 5 presents a new inverse Tone Mapping Operator developed to be lighter and faster than the state-of-the-art methods. Along with the architecture and training process, we present an evaluation of our work using objective metrics and a user subjective study.

Chapter 6 is a conclusion that contains a summary of our work, as well as some perspective about image quality improvement.

GENERAL NOTIONS

In this chapter, we present general knowledge about some aspects of our work, namely aesthetics and inverse tone mapping.

2.1 Introduction

For human beings, one of the methods to capture information from their surroundings is by seeing. The visual system, composed of the eyes, the optical nerve and the parts of the brain that process visual information, is used not only as a sensor to gather factual data about the world, but also as a processing unit that uses the factual information to make decisions. To better understand how we make a given decision based on what we see, it is important to understand each of the parts at stake here: how is the light transported through the world? How do we see? What happens in our brain? And most importantly, is it possible to model the different factors (namely the light transportation and the human visual system), and the interaction between both of them?

2.2 Human Visual Perception

In this section, we present characteristics of the human visual system that guided some choices we have made and that are explained in the following chapters.

2.2.1 The eye

Human visual perception starts from the eye. This organ captures light coming into it, and transforms it into an electric signal later processed by the brain. Many processes related to how this signal is processed are still unknown nowadays.

The part that acts as the transducer is the retina. The surface of the retina is composed of two different types of cells, called cones and rods. The rods are sensitive to light

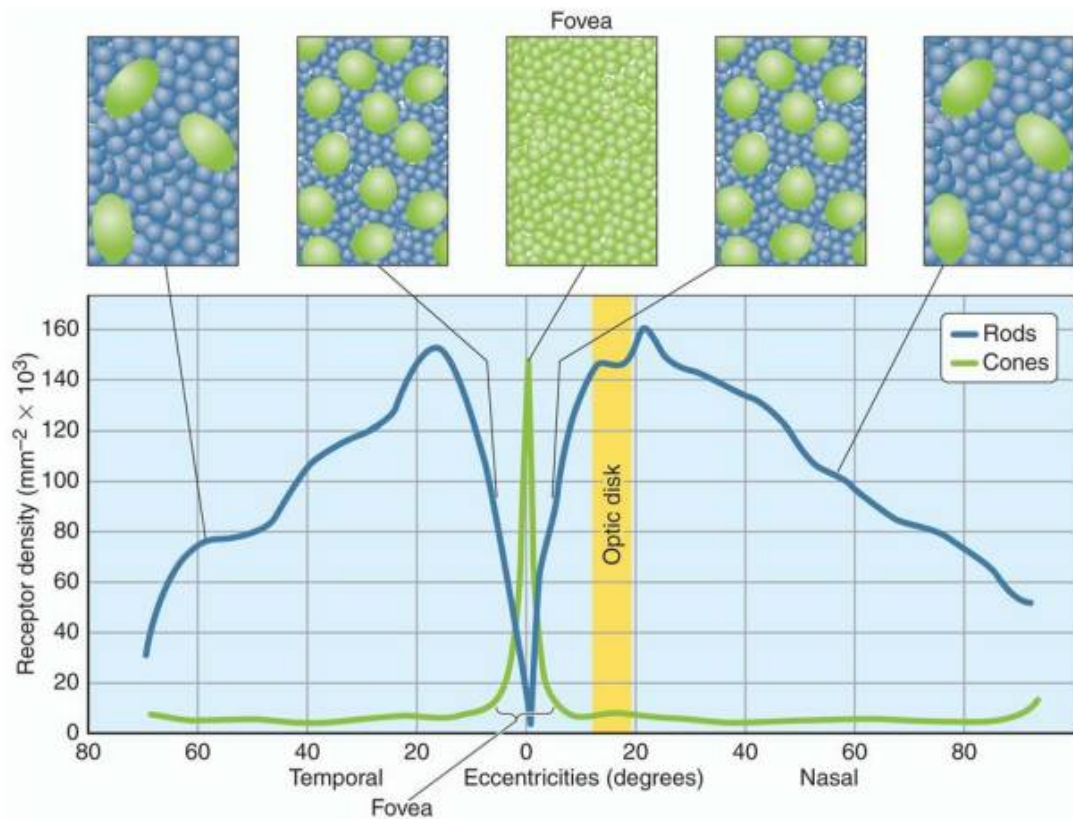


Figure 2.1 – Distribution of rods and cones on the retina (extracted from [MEP09]). The fovea (zone of maximal visual acuity) is represented.

intensity: they allow vision in low light conditions, and they are able to discriminate brightness values. On the other hand, the cones are more sensitive to color values: they need more light to operate than the rods, but they can discriminate color hues. The cones and rods transform the light wave into an electric signal, and sends it to the brain through the optical nerve. We show on Figure 2.1 a repartition of the cones and rods on the retina as a function of the distance to the optical nerve. We can see that the majority of the cones are concentrated around a small area. This zone is called the *fovea*, and is the zone of maximum visual acuity. It captures light in a cone of peak angle of 2° . This two-cells repartition of the retina lead to one of the representation of images, where the color value of one specific point is represented by two values: the color value, and the brightness value. As the rods are globally more present on the retina, human eye is more sensitive to light variations than to color variations. That is why in Chapter 5 we process only the luminance of images, and then recolor them.

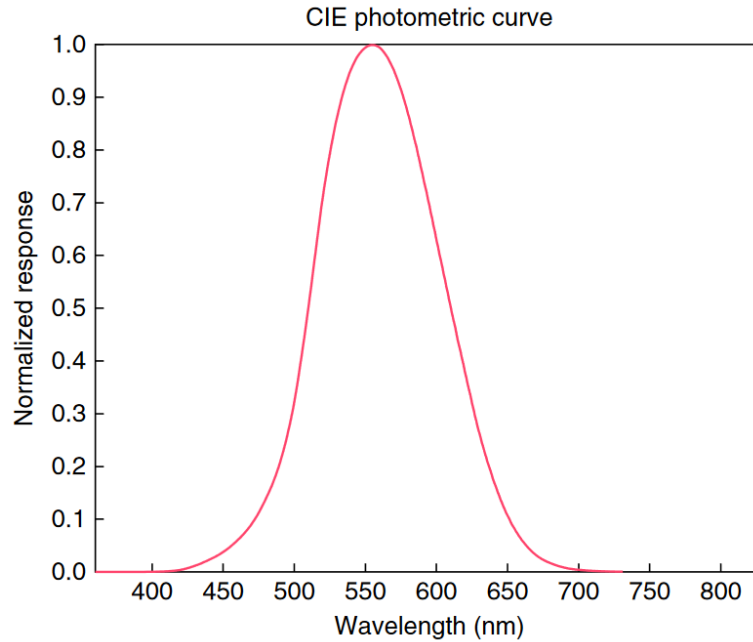


Figure 2.2 – The luminous efficiency function (extracted from [Rei+10]).

2.2.2 Radiometry and Photometry

In the world, light is transported by photons or waves. Light has several physical properties – such as color hue, energy, direction, polarity, etc...– that can be measured and expressed using different physical quantities. The human eye acts as a transducer and we are usually more interested in how we perceive things rather than the precise energy value. That is why each value has a measurable physical value, and an estimated perceived value. The study of physical values is called radiometry while the study of perceived values is called photometry. Every radiometric value has a photometric counterpart.

The photometric values are obtained by computing the average of the corresponding radiometric value, weighted by a function called the luminous efficiency function. This luminous efficiency function (represented on Figure 2.2) represents the eye sensitivity of a standard observer, which corresponds to a norm established by the CIE (*Commission Internationale de l'Éclairage*, International Commission on Illumination). In particular, the energy in frequencies outside the visible domain range is not taken into account, and the frequency with maximum weight corresponds to the frequency of maximum acuity of the human eye (around a wavelength $\lambda = 510$ nm for scotopic vision, which corresponds to green).

One important value we consider in this work is the radiometric radiance (in $\text{W}\cdot\text{m}^{-2}\cdot\text{sr}^{-1}$) and its photometric counterpart the luminance (in $\text{cd}\cdot\text{m}^{-2}$). The luminance quantifies the perceived brightness of a point. Besides, we base our approach on results from psychophysical experiments. In particular, Fechner’s law [NW97] states that the perceived brightness of a surface by a human is proportional to the logarithm of the radiance. In consequence, the majority of the operations we have to do on the luminance is performed in the log-domain.

2.2.3 High Dynamic Range imaging

The Dynamic Range (DR) is a measure of the range of luminance in a scene, in an image, or capable to be rendered by a display. It is computed as $DR = \frac{\max L}{\min L}$ where DR is the Dynamic Range and L is the luminance in $\text{cd}\cdot\text{m}^{-2}$. The DR is usually given with a reference of $\min L = 1 \text{ cd}\cdot\text{m}^{-2}$, that we note $10^n:1$ in the rest of this section. For example, a screen that can display luminances from $\min L = 0.1 \text{ cd}\cdot\text{m}^{-2}$ to $\max L = 100 \text{ cd}\cdot\text{m}^{-2}$ has a dynamic range of $DR = \frac{10^2}{0.1}$, noted here as $10^3:1$. We can compare the DR of a conventional screen, which is about $10^2:1$ to the DR of the real world, which can be up to $10^{32}:1$. As we can see, the displaying technologies of classic screen are nowhere near enough to display the full range of luminances of the real world. Besides, the DR of the eyes is about $10^{10}:1$, which means that we are able to perceive more details than what can be displayed on screens.

Due to that observation, researchers and industries have studied and devised new devices capable of displaying a broader range of luminances. Those devices and the content that they are able to display are called High Dynamic Range (HDR), in contrast to the Standard Dynamic Range (SDR) of conventional screens. Therefore, one operation, which is very important nowadays as the HDR displays are not as widespread as the HDR images themselves, is the transformation from an HDR image to an SDR one. This transformation is called *tone mapping* and is a key subject of many articles in computer vision.

There are several ways of creating HDR images. It is possible to synthesize HDR images using rendering techniques. Indeed, computer-generated images can have pixel values of arbitrary precision. Nevertheless, the main focus of our work is HDR photographs, so HDR rendering technique will not be detailed in this thesis.

There exists two main methods to capture photographs with HDR content. The first intuitive way consists in using specific sensors that are able to capture HDR images. For instance, we can use cameras adapted to HDR photography, or other devices added

to a camera to improve the captured DR, such as lenses, event cameras, etc... Taking HDR photographs is very easy with specific devices, however, such cameras are not easily available.

The second way to capture HDR photographs consists in merging together several pictures of the same scene, but containing information in different areas of the image. To do so, it is possible to take several photographs with different exposure values. The exposure value is controlled by three variables: the film sensitivity (ISO), the shutter speed and the shutter aperture. For this purpose, it is easier to change the exposure value by changing the shutter speed of the camera. The different shots of the scene then contain information for every range of luminance: the darker shot (with the lowest exposure time) contain information in bright areas, while the brighter shot (with the highest exposure time) contain information in dark areas. This method is easily achieved on a modern camera using the bracketing function. However, this method has many constraints, the main one being that everything must stand still for the duration of all the shots. We can act on the camera itself by using a tripod (to avoid camera shaking) and activating the camera from a distance (as pushing the button could move the camera), but sometimes, it is difficult to act on the scene itself, as objects can move (for example, clouds or birds). The presence of a moving object in the scene leads in the final reconstituted HDR image to ghosting artifacts, as shown on Figure 2.3.

2.2.4 Tone Mapping and Inverse Tone Mapping

The first algorithms needed to work with HDR content are tone mapping operators (TMO). HDR content is quite easy to create (exposure fusion using bracketed images, renderings from virtual scenes) compared to designing HDR-capable hardware. Therefore, we need to convert HDR content to SDR one, while keeping as much information as possible depending on the use case. Algorithms that transform HDR content to SDR are called TMO. The most simple and naive TMO would be the operation that consists in clamping the value of the image to the range of the display. More sophisticated methods were proposed to try to keep specific information (such as local contrast, more details and examples can be found in Chapter 3).

Generally, applying a TMO on an image leads to a loss of information when we try to compress values. However, some of the TMOs are functions that are completely reversible. This means that if the SDR image is stored correctly, then no information is lost due to the tone mapping operation. This leads to the creation of Inverse Tone Mapping Oper-



Figure 2.3 – An example of an HDR image with ghosting artifacts (the image has been tone-mapped with the Drago algorithm for ease of view).

ators (iTMOs) that process a single SDR image to yield an HDR image. By extension, researchers have devised new algorithms to expand the dynamic range of images that are called Expansion Operators (EOs), or abusively inverse Tone Mapping Operators. Expansion operators work using heuristics, as it is impossible to recreate HDR content from one unique source of SDR content.

One application of EOs is to easily create a large amount of HDR images. Since the emergence of machine learning, it is now possible to process HDR images, but to do so, a large amount of HDR content is needed. The collection of such images is tedious, so it is possible to rely on computational methods, such as EOs.

2.2.5 Metrics

To assess the performance of image processing algorithms, we need a way to measure the quality of images. To do so, it is possible to use metrics. There are two main kinds of metrics: with reference (that needs a reference image along with the tested image, which can be treated as a distance between images); and no reference metrics (which try to compute an intrinsic quality value of images). For SDR images, there are many metrics such as MSE, SSIM, PSNR (if we see the image as a 2D signal), etc... Some of them can be expanded to work with HDR content quite easily. Other metrics were especially designed for HDR content [Rou19], to compare two HDR images, or even an HDR to an SDR image (to be able to evaluate the performances of (i)TMOs).

HDRVDP is certainly the most used metric to compare HDR images. It is a metric with reference (meaning that we compare the original HDR image with a degraded version of it). This metric mimics the visualisation environment to predict the perceived differences between the two input images (for example, some differences are visible if you are close to the screen, but are unnoticed if you are farther away).

2.3 Aesthetics

Based on the information given by the visual system, one of the feeling that the brain can evoke is the aesthetic feeling. Aesthetics is commonly a measure of beauty. In this section, we propose a more precise definition of aesthetics applied in computer science.

2.3.1 Definition

It is quite difficult to give a general definition of what is aesthetics. The aesthetic feeling has been studied, quantified and defined for a very long time, by people of different fields. Moreover, depending on the field, the different problematics about aesthetics are different. This means that the actual concept of aesthetics is different, although the name is the same.

The questions we want to answer here are: Are there intrinsic features of objects (more specifically photographs) that make them beautiful? If so, can we model them to be able to automatically assess aesthetics quality? If no, then the aesthetic feeling is a pure subjective value: can we model the perception of human beings to be able to assess aesthetics at least for a given individual?

According to Shelley [She22], aesthetics is a by-product of the concept of taste. Intuitively, the aesthetic feeling comes from the interaction between the human and a given object. Aesthetics is strongly linked to emotions, as aesthetically pleasant objects are usually the objects that make people feel some emotion when they interact with it. Human beings interact with the world using their senses, and as such, aesthetics is a strongly subjective notion. Through the years, several scientific fields tried to explain and define the aesthetic feeling, such as philosophy, psychology, and even computer science. Even in each of these fields, aesthetics is defined differently depending on the use case and the historical context of such definition.

At first, the aesthetics was said to be "objectivist": this means that each object is defined by its own attributes, and its aesthetic quality is inherent in the object itself. However, this implies that only an expert with knowledge of aesthetics can truly appreciate an object and accurately rate its aesthetic value. With the observation that people have their own sense of taste and aesthetic feeling, this theory as a general explanation of aesthetics is not sufficient. As a consequence, the subjectivist theory had been studied: this theory claims that the aesthetic feeling only comes from the interaction between a sentient being and an object, and if this interaction evoke some emotions, then the object is considered aesthetically pleasing. The aesthetic quality of an object is then entirely defined by the perception of people interacting with it, and is considered entirely subjective. Different people may have different experience with the same object, meaning that it is impossible to assign an aesthetic quality to an object without taking into account the possible observers. However, this hypothesis failed as some objects (especially some work of art) were considered beautiful by a large majority of people. This implies that

objects themselves can have characteristics that impact the aesthetic perception of observers, whoever the observer may be. Therefore, the exclusive objectivist and exclusive subjectivist theories are both not enough to fully explain the aesthetic feeling.

These discussions led to the creation of a new domain called *neuroaesthetics*. The goal is to observe the brain signals of observers to try to explain how do human beings perceive an object as aesthetic or not. This field tries to understand which parts of the brain are activated when viewing a beautiful object, and what specific patterns activate or not these brain parts. The experimentations done in neuroaesthetics try to prove using the scientific methodology what is exactly aesthetics.

2.3.2 Application in computer science

In our work, we use the definition of aesthetics given by the field of computational aesthetics. One of the precursor of computational aesthetics is George David Birkhoff, with his book *Aesthetic Measure* [Bir33]. In his book, Birkhoff claims that the aesthetic quality of any object (paintings, vase patterns, polygons, but also poems, or music pieces) heavily depends on the ratio $\frac{O}{C}$, where C is the complexity of the object, and O a measure of its order. Of course, the definition of C and O depends on the type of the object we consider. More than giving some hints about what impacts aesthetics, Birkhoff provides the community with a mathematical formula which computes the aesthetic quality of an object as a real number. The aesthetic quality is then comparable between different objects and, based on this formula, we can claim that an object is "more beautiful" or "less beautiful" than another.

The aesthetic quality is in particular useful for images. In our work, we mainly study the aesthetics of photographs. Computing the complexity and order of an image is possible using the notion of entropy and of information theory [Mol66]. In the first few articles tackling the problem of aesthetics assessment [FB09], complexity and order were computed using image features: either specific features – such as rule of thirds compliance, color harmony or symmetry – or generic features such as SIFT. Then, learning algorithms such as SVM or simple regression were used to link the features with a level of aesthetics.

From then until today, the extraction of characteristic features from images has evolved, and aesthetic quality followed this trend. Deep convolutional networks are known to successfully extract patterns from images to be able to perform computations (generating new images for style transfer, super-resolution, re-colorization; or generating other types of data, such as text for captioning, numerical value for quality, and many

other applications) and this is also the case for aesthetic quality.

The goal is to design an algorithm to predict what would be the global aesthetic feeling of one photograph, by extrapolating from the data gathered from hundreds of observers, thanks to learning techniques.

The definition of aesthetics that we consider is then "liked by a majority of people". We gather the subjective feelings of a lot of people to get an objective value. This definition is questionable, and is one of the discussions of the conclusion.

2.4 Conclusion

In this chapter, we have presented some important notions (aesthetics, HDR imaging) and some operations (aesthetics assessment and inverse tone mapping). We want to address some of the problematics raised by those operations. In the next chapter, we present the state-of-the-art, and a general scientific outline of the different domains we tackle.

RELATED WORK

In this chapter, we present the most recent work about what is presented in the first chapter, that is to say aesthetics prediction, and HDR imaging.

3.1 Aesthetics assessment

Computational aesthetics involves supervised learning methods today. There are two main characteristics of supervised learning algorithms, which are the architecture (and its loss function, training algorithm, etc...) and the training datasets. In this section, we present an overview of the different learning-based aesthetics prediction methods. A more thorough study of such methods can be found in [VKD22].

3.1.1 Prediction models

The first learning-based methods were using handcrafted features, derived from photographic rules. Photographic rules are common rules in photography. These rules are not absolute, but compliance to these rules is seen as a good start for a beginner in photography. An overview of some photographic rules is presented in Table 3.1. It is important to note here that the term "rules" is misleading: a photograph that complies to many photographic rules is not necessarily beautiful. These photographic rules are closer to general guidelines that the photographer chooses to follow or not. The first aesthetics prediction model was proposed in 2006 by Datta et al. [Dat+06]. Their work is heavily based on photographic rules: they have chosen 47 different features easily computed for photographs and then feed a SVM classifier with those features for a given dataset to learn how each of those features impacts the aesthetic value of the photographs.

The development of generic feature extraction also contributed to the improvement of aesthetics prediction models. Indeed, as we explained in the current section, photographic rules are not absolute, and so are not directly linked with aesthetics. As such, it

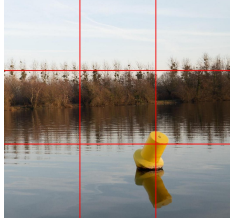
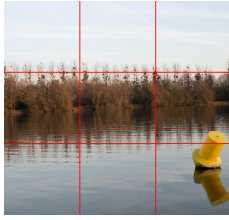



Name	Description	Positive example	Negative example
Rule of thirds	The object of interest in the photograph should be placed on one of the lines dividing the image in thirds		
Depth of field	The focus should be placed mainly on the object of interest, blurring the other parts of the photograph		
Color harmony	Opponent colors allow to highlight the main subject of the photograph		N/A

Table 3.1 – Presentation of some of the commonly accepted photographic rules. The example images all come from the AVA dataset [MMP12].

is important to propose models that do not rely solely on photographic rules, and generic features allows to represent images in a different way. For example, Marchesotti and Perronnin [MPM13] use SIFT [Low99] along with pattern-mined textual features to classify the aesthetic quality of photographs.

The emergence of deep learning lead to tremendous breakthrough in computer vision, and so in computational aesthetics. The first deep learning-based model for assessing the aesthetic quality is RAPID [Lu+14]. The idea behind RAPID is to process both large scale (or global) features, and detail (or local) features using a two-stream convolutional neural network.

Many other networks followed, and as such, it is quite difficult to offer an exhaustive overview of the different learning-based aesthetics assessment models since RAPID.

One way to classify the different aesthetics assessment models is by grouping together models with the same kind of output quality. RAPID [Lu+14] for example outputs a binary category: "high quality" or "low quality". This is one of the most basic output possible for aesthetics prediction.

To improve on these classification networks, it is possible to enrich the output infor-

mation. The next logical step is to output a numerical score.

Many networks offer to get more information from meaningful output. NIMA [TM18a] has an architecture derived from already-trained classification networks (such as Inception [Sze+16a] or MobileNet [How+17]) adapted to aesthetics assessment. The output is a full histogram of scores. Histograms allow to have information on the consensus. If the standard deviation of the histogram is very small, then the large majority of the scores are gathered around the mean, and so it means that the model predicts that people would agree on the given score (be it high or low). On the other hand, a high standard deviation means that people would not agree on the aesthetic quality of the image, and then this image is more prone to subjectivity than others.

The Ranking Network [Kon+16a] provides a single aesthetic score, but bases its calculations on several attributes that an image may or may not have that are deemed to have an impact on aesthetics by the authors. These attributes are based on photographic rules (compliance to the rule of thirds, symmetry, good field of view, etc...) but each of them corresponds to the output of a specific CNN, and the final aesthetics score is given by a final CNN that merges the outputs of all intermediate CNNs. Therefore, the output of the network is one general aesthetics score, with a rating of each intermediate attributes. This is one step closer to explainable aesthetics.

3.1.2 Aesthetics assessment datasets

The goal here is to propose aesthetics assessment methods that are as generic as possible. This means that a characteristic that we want to foster is the ability to be able to rate accurately any kind of images, that we call coverage in the rest of this thesis. As the training dataset delimits the theoretical boundaries of acceptable inputs of the network, it is important to have an effective training dataset to improve the coverage. However, the training dataset is not the only impactful factor on the coverage. Indeed, it is possible to act on the coverage by working on the architecture or other parts of the model. We present in this section the different datasets proposed for aesthetics assessment, as well as some techniques that were used to improve the coverage other than changing the dataset (that is to say, techniques to compensate for the lack of variety in the dataset).

Providing new datasets Recent models of aesthetics prediction are based on supervised machine learning algorithms. One key component of such algorithms is the training dataset. In this chapter and the following ones, we consider the datasets to be images

along with their respective ground truth value (which may be of different form depending on the dataset). In this section, we present the different training datasets that exist for aesthetics prediction.

There are two main ways to create aesthetics prediction datasets. The first method is to ask a population to rate a specific set of images according to the wanted criteria (in our case, aesthetics). This first method allows to have ground truth values which are very accurate and that can describe precisely the wanted feature in the images. However, to have enough information, it is important to have a sufficiently large population of observers, which is not easy to gather. Besides, it is not possible to ask people to rate thousands of images in a sensible time, so it is quite hard to gather enough scores for a number of images on the scale of what is needed to train neural networks.

The second method to create aesthetics prediction datasets is to crawl existing websites that propose a system of votes for images. The first datasets that were proposed were based on all-purpose photograph social networks, such as Flickr (used in [DOB11]). The ground truth scores have to be derived from values given by the website. In the case of Flickr, the interestingness value given by the website for each image may be used as an estimation of the aesthetic value. However, it is not suitable to train a robust network. It is not possible to precisely know how this value is calculated and Flickr do not claim that this value should be at least proportional to the aesthetic value of the image. It is better to take images and scores from more specialized websites, for example ones that tackle more specifically photography.

Datta et al. [Dat+06], in the first known aesthetics assessment model, used a dataset based on photo.net. On photo.net, the images are rated by the users of the website on two different scales: aesthetics and originality. Even though the website contains enough images to be able to train deep networks, each image is not rated by many users.

In 2012, Murray et al. [MMP12] proposed the first large scale training dataset for aesthetics prediction, called Aesthetics Visual Analysis (AVA). AVA contains more than 250000 images taken from dpchallenge.com, a social network for photographer. Every week, the organizers propose a new challenge with a given theme, and the users have a few days to upload a photograph corresponding to the theme. After the end of the time limit, the visitors can then grade the images according to two criteria: aesthetics and accordance to the theme. The website has been crawled to gather the images, the challenge theme and the repartition of scores for both scales. It is still, to this day, the largest training dataset for aesthetics prediction.

Name	Source of annotations	# of images	Information available
CUHKPQ [LWT11]	Manual	17,500	Aesthetics class
AVA [MMP12]	dpchallenge.com	220,000	Histogram
PCCD [CLC17]	gurushots.com	4,000	Aesthetics captions
EVA [KVD20]	Crowd-sourcing	4,000	Scores for different characteristics
RPCD [NCF22]	reddit.com	74,000	Aesthetics captions

Table 3.2 – Summary of aesthetics assessment datasets.

Many work have improved on AVA to create new, richer dataset. In some cases, AVA is improved as the designed method needs more information in the dataset. It is the case for NAIR [Wan+19a] that is an aesthetics assessment network that outputs an aesthetics score as well as a textual review of the image. The training of such a network calls for a specialized dataset that contains ground truth value for aesthetic score and a textual explanation of the aesthetic quality. Therefore, Wang et al. proposed their architecture NAIR along its training dataset called Ava-reviews, that contains a subset of AVA augmented with the comments given by the users of dpchallenge.

On the other hand, some improvements over AVA do not stem from a necessity of the application. For example, the Explainable Visual Analysis [KVD20] (EVA) dataset contains images from AVA, graded by additional people (via crowd-sourcing) on different attributes, that were chosen by the authors by thinking they were impactful for aesthetics. We present a summary of the different aesthetics datasets available in Table 3.2.

Using the architecture to improve the coverage Wang et al. [Wan+16] propose a new architecture for aesthetics models based on neurosciences. In parallel, the authors conduct some user experiments to assess how the image transformations used in data augmentation affect aesthetics. In their study, participants were presented with rotated, color-changed or cropped images and they were asked to rate the images according to their preference. They conclude that only three transformations do not severely impact the aesthetic quality of an image: horizontal flip, scaling with a small random factor, and adding some small gaussian noise. This conclusion gives some insights when using data augmentation for training aesthetic assessment models, but does not question the validity of currently available datasets.

Carballal et al. [Car+19b] recently expose the limits of existing training datasets, and create their own dataset composed of images from www.dpchallenge.com. Those images

are rated in three ways: the mean score given to the image by users of www.dpchallenge.com; an aesthetic score given by observers in an environment with controlled viewing conditions; and a preference score given by observers in the same controlled environment. This is the first and only dataset dedicated to aesthetics prediction with scores from several populations. However, this dataset has not been compared to the dataset most used today (AVA) and contains only 1,000 images.

Some models were proposed to counter the bias caused by the imbalance of AVA [MMP12]. In that optic, Jin et al. [JSS16] propose the weighted CNN architecture. This is an architecture based on the VGG-16 network, but differs from the VGG-16 network in the cost function used. Instead of using a classic mean square error function, they add weights corresponding to the frequency of apparition of the score of the image. To compute those weights, the authors use the histogram of scores of the training dataset (in this case, AVA). The weight $w(I)$ of an image I is proportional to the inverse of the number of images having the same average score as I . The cost function is then defined as

$$C(\hat{\mathbf{y}}) = \sum_{I \in \mathcal{I}} w(I) \|\mathbf{y}_0(I) - \hat{\mathbf{y}}(I)\|_2^2 \quad (3.1)$$

with \mathcal{I} the set of training images, $\hat{\mathbf{y}}$ the vector composed of the computed scores of each image in the dataset and \mathbf{y}_0 the vector composed of the ground truth score of the images. This method allows to assign greater weights to images which scores are less represented in the training dataset. This leads to a network having a better coverage and being able to rate accurately images with very high ($s > 6$) or very low ($s < 4$) scores.

Rather than assessing the aesthetics of images, a number of authors addressed the problem of Blind Image Quality Assessment (BIQA). These methods use new techniques based on machine learning, such as Continual Learning [Zha+22], meta-learning [Zhu+20] or datasets fusion [Zha+21; WM21]. Differently from aesthetic quality assessment methods, BIQA methods process images purely as signals. This implies that the quality of an image as measured by BIQA methods is an objective value, contrary to the aesthetic quality which may differ depending on the observer. Therefore, we think that these methods are not suited for general purpose aesthetics assessment models, but could be adapted to personalised aesthetics prediction [Ren+17b]. Indeed, when considering only one observer, an objective aesthetic quality score exists. That is why, in our approach presented in Chapter 4, we did not use these techniques and prefer an aesthetic-based one.

Dataset	Number of images	Characteristics
pfstools [Man+07b]	8	-
HDR PS [Fai07]	450	Some images come with photometric measurements
Raise [Dan+15]	8152	Only RAW images
HDR-Synth [Liu+20]	562	Aggregation of different datasets
HDR-Real [Liu+20]	480	-
Our dataset [Cha+23]	496	High resolution

Table 3.3 – Overview of different datasets with HDR content.

3.2 High Dynamic Range Imaging

In this Section, we present the recent work about HDR imaging.

3.2.1 HDR datasets

HDR image datasets serve several purposes, and depending on the use case, the best characteristic would change. For example, testing a new tone mapping operator by comparing the SDR output of one method to the SDR outputs of the state-of-the-art methods requires far fewer images than training a deep supervised neural network containing several millions of parameters. In this section, we present some of the HDR datasets used in the state-of-the-art depending on their application. Table 3.3 presents an overview of commonly used HDR datasets.

First, for testing purposes, HDR images need to be of high quality (resolution, artistic quality, dynamic range) and of different kinds to be as exhaustive as possible. Nevertheless, a few hundreds of images are usually enough. In this case, HDR photographs are a good source of test images.

However, to train neural networks, a large number is usually better than image of great quality. There are several methods to gather many HDR images of correct quality: transformations from fewer high quality HDR image (cropping); synthetic images from rendered scene (although the variety of depicted scenes in this case might not be much). Several techniques are used: smaller networks [Cha+23] (that need fewer training images), data augmentation of many kinds (cropping, exposure change, mirroring, etc...), transfer learning (first learn on many synthetic data, then finetune on few real data).

3.2.2 HDR Image Generation

Hardware-based HDR generation Intuitively, the best way to create HDR content is to devise specific devices. One method is to use other kinds of sensors along with classic CCD sensors to get more information. For example, Han et al. [Han+20] propose a new method to fuse an SDR image with an intensity map provided by an event camera (also called neuromorphic camera) using deep learning.

Instead of adding new information (such as intensity maps) of the same scene, another method is to modify an existing camera to better reconstruct the HDR afterwards. Some articles [Met+20; Sun+20] combine the design of a new lens – which point spread function is thus known and optimized for HDR content – and a neural network to reconstruct HDR content from the image taken by this modified lens.

Multi-exposure fusion Using only a classic camera, the most popular way to create HDR content is by fusing multiple images of the same scene with different exposure times. Several fusion algorithms exist [MKV08; DM08]. The fusion methods are known to yield HDR images of very good quality, but the process of taking the image is more difficult than other methods: both the camera and the scene must stand still for several seconds, the time to take several images with different exposure times. If the camera, or objects in the scene, move between the different shots, some ghosting artifacts will appear on the merged HDR image.

Single-exposure non-deep fusion The first iTMOs were based on TMOs: by looking at what was reduced during the tone mapping process, we can deduce where it would be good to expand the dynamic, hence the term "inverse" tone mapping. Some tone mapping operator are invertible, so we can consider their inverse function.

By using iTMOs to generate HDR images, information is lacking in several parts of the images and HDR generation becomes an ill-posed problem. In this situation we need to make some assumptions about the images we have. For example, we can assume that only the high luminance areas are lost. In that case, any pixel in the low or medium ranges would not be modified by the iTMO. Many algorithms were devised to expand the dynamic range from a single image. One idea is to use a non-linear function to modify the luminance of the image differently based on the pixel value of the SDR version. We can cite several methods, such as Akyuz [Aky+07], Kovaleski & Oliveira [KO14], or Landis [Lan02]. All of them differ from one another by the non-linear function they use

to improve the dynamic range. The modification of the luminance value is solely based on the SDR pixel value, and so the context is not taken into account. Later works try to consider the semantics of each pixel to better improve the output quality using neural networks.

Deep neural networks for single image HDR reconstruction The latest and more powerful algorithms for single image HDR reconstruction are based on deep-learning algorithms, and more specifically supervised learning methods. These algorithms allow for tuning several millions of parameters using big datasets of images as ground truth. The first widely recognized deep CNNs for single image HDR reconstruction are HDRCNN [Eil+17] and DrTMO [EKM17].

DrTMO is a collection of several CNNs. Each CNN computes a new exposure from a given SDR input (the output of each CNN is then an SDR image with an exposure value different from the input). These exposures are then merged together using the HDR generation techniques described at the begin of the current section.

HDRCNN uses a really deep network of about 30 million parameters to enhance the brightest part of the SDR picture in input. The output of the network is then combined with an augmented version of the SDR (obtained with an average inverse camera response function estimated over a dataset) using a mask to use the network output in bright zones and the augmented SDR in the other areas. The network has been pre-trained on simulated HDR data (using a simple iTMO on a large image dataset) and fine-tuned with true HDR images. Other papers [SRK20; Liu+20] improve on HDRCNN by using inpainting-like tasks in the network – either as pre-trained weights or as another module. The idea is that such a network must reconstruct the over-exposed areas of the images, as this information is lost in the SDR image. Usually, HDR imaging focuses on high lights, and therefore the proposed methods work on improving the over-exposed areas. However, the same tricks can be used to improve the quality of under-exposed areas as well. The network of Marnerides et al. [MBD21] uses generative adversarial networks and inpainting to improve the dynamic range in both lowly and highly lit parts of the image. Some authors carefully designed the architecture of their network to process HDR images more efficiently. For example, Yu et al. [Yu+21] designed a specific CNN module, called luminance attention module, to process luminance information. The authors proposed LANet, a deep learning-based iTMO composed of two different parts: a stream designed to perform a dynamic range expansion (the iTMO part) and a stream trained for luminance

segmentation. The luminance attention module is used to transfer needed information from the luminance segmentation stream to the iTMO stream.

To improve on classic CNN architectures, new methods have been developed, and some of them have been implemented to perform dynamic range expansion. For example, the Deep Recursive HDRI model [LAK18] presents a Generative Adversarial Network (GAN) architecture for iTMOs. GANs allow to have more realistic results. The main idea of GANs is to add a second network trained together with the generative CNN. The new network, called the discriminator, is trained to recognize, given two inputs, which one comes from the generative network, and which one does not. The training of the discriminator improves the results from the generator, and therefore the overall quality of the network.

During the past few years, transformers have acquired a large notoriety thanks to their improved performances over classic CNN. Many different computer vision task were improved thanks to transformers, and iTMOs are not an exception. Several new iTMOs that includes transformers were proposed. DenSE SwinHDR [BYB22] uses both CNN and transformers to produce an HDR image from an SDR input. By using transformers for global features and CNN for local ones, DenSE SwinHDR have great performances as an iTMO.

3.2.3 Our approach

The majority of the above presented methods are based on neural networks containing a large number of parameters (between 10^6 and 10^8 trainable parameters). A large number of parameters usually allows to have better performance (thanks to a finer latent representation of images in the network), however, the network then requires a large amount of resources (training time, training images, etc...). This is a problem in situations with limited resources (such as HDR processing, as there exists few high quality HDR images for training neural networks). That is why we decided to propose a new iTMO.

In our approach presented in chapter 5, we use preprocessing to reduce significantly the number of parameters of our network, and postprocessing to correct the network output as best as possible. This allows us to improve lowly and highly lit parts of images.

The method we propose tries to broaden the scope of the above presented models by modifying the used training dataset. We show that, although a smart data augmentation and the use of a weighted loss function can alleviate the effect of the specialisation of the training dataset, the fine-tuning method we propose is more effective to improve the

coverage of the aesthetics models.

ASSESSING AESTHETICS USING PROFESSIONAL PHOTOGRAPHS

4.1 Introduction

Assessing the aesthetic quality of images using computational models has been a problem in the computer vision field for many years. Aesthetic quality assessment has some applications in image sorting for databases management, or in aesthetics-driven image processing for example. Yet, automatically scoring the beauty of an image is still a difficult problem. Compared to the problem of image quality assessment [Wan11; WSB03], aesthetics assessment has to be measured by using high level features, which are hardly described by common low level features. Beyond this point, the problem is even trickier since the aesthetic quality of an image is a highly subjective quantity.

For predicting the aesthetic quality of an image, many computational models have been proposed. The first models are based on specific photographic rules, such as the rule of thirds related to image composition, or narrow depth of field [Jos+11a]. Performances of such models are however rather limited. We are currently witnessing a new breakthrough in this field thanks to the emergence of huge datasets (e.g. AVA [MMP12]) and new machine learning methods relying on deep learning algorithms. Thanks to deep networks, trained over millions of images [SZ14], it is now possible to get a large number of features able to describe complex and abstract patterns. A more thorough study on recent models can be found in [DLX17].

In 2012, Murray et al. [MMP12] proposed AVA, a dataset specifically designed for aesthetics assessment methods. This dataset, composed of more than 250,000 aesthetically annotated and scored images from the photography website www.dpchallenge.com, greatly helped research in this domain. However, this dataset is mainly composed of competitive photographs aimed to be shown in juried exhibitions, to be published in specialised photography magazines or web sites, and to compete for recognition and prizes,

as described in [TM18b]. For the specific case of AVA, the scores range from 1 (i.e. ugly) to 10 (i.e. beautiful). The mean score is 5.10; the maximum and minimum scores are 8.52 and 1.81, respectively. As current models are trained over the AVA dataset, we could say that current aesthetic models are mainly dedicated to scoring competitive photography.

Using such a dataset suggests that some assumptions are implicitly made, in particular the definition of aesthetic quality used. If we consider an image with a high score in AVA dataset as an image with high aesthetic quality, this means that we recognise the general agreement as an accurate measure of the aesthetic quality: if enough people find an image beautiful, then it must be objectively beautiful. While this assumption is quite restrictive, it certainly fits the competitive nature of the photographs of AVA. However, this definition of aesthetics is not applicable to all categories of photographs. Because most of the recent papers on aesthetic assessment do not make explicit this assumption, we want to show in our work the importance of content diversity in the training images, especially for aesthetics quality assessment.

It is common to consider three main usages of photography: competitive, vernacular and professional.

Competitive photographs, as mentioned previously, are beautiful images that should be liked by a very large audience [TM18b]. Competitive photography existed before Internet, but it becomes very popular with social networks specialised in photo sharing such as www.instagram.com, www.DPchallenge.com, www.flickr.com, etc. To be liked by a larger number of people, competitive photography usually follows classic aesthetics rules and is very aesthetically conservative.

Vernacular photographs are for a personal and family usage. They capture personal events in order to keep a souvenir or to share these personal events with family or a private community. The beauty is not the main objective in this context.

Professional photographs aim to be seen by a large number of people and aim to convey a message or an emotion. Since the beginning of professional image creation (including paintings and photographs), an image of high aesthetic quality is in most of the cases more efficient to convey the intent, message or emotion. In this professional case, aesthetics means well designed technically speaking, but does not necessarily mean that the photography is pleasant. Indeed, the photograph objective can be to shock in order to produce a reaction. Professional photography can further be classified into two main genres: photojournalism (war photography, sport photography) and product photography (fashion photography, real estate or architecture photography, etc.). The former endeavors

to capture an instant or its related emotion, while product photographs aim to promote a product to make it desirable.

The main difference between those categories is the aim of the photograph. Competitive photography aims at being pleasant for a majority of people, therefore high aesthetic quality is the end goal of such photography. On the other hand, professional photography aims at conveying the intent of the photographer or its commissioner. Complying with common aesthetic rules in this case is only one mean among others to achieve such a goal.

In this chapter, we present our two main contributions. Our first contribution is to test aesthetics prediction algorithms (which we call models in the rest of this chapter) to assess whether or not they perform well on different categories of photography. This test allowed us to quantify the coverage of the models – kinds of photographs accurately rated by the model –, to eventually improve it. As models are mainly trained using competitive photographs, we have collected several datasets of professional photographs to test the recent models NIMA [TM18a] as well as the Ranking Network [Kon+16a]. This first study shows that there exists a significant difference in the behaviour of the models on competitive and professional photographs, and between different categories of professional photographs. This observation led us to our second contribution: we propose a solution to reduce the impact of the training dataset on the coverage of the networks by fine-tuning the already trained models using other kinds of photographs (in our case, using professional photographs). We prove that this method effectively improves the coverage of the models as explained in the following sections.

These contributions led to an article [CCL22]. The rest of this chapter is structured as follows: Section 4.2 presents the models and the different professional datasets we used in the experiments and Section 4.3 presents the experiments themselves; Section 4.4 exposes the results of the experiment and finally, we conclude in Section 4.5 by summing up our findings and proposing some future work.

4.2 Testing models and datasets

The experiment we propose relies on testing two recent computational aesthetics models to measure their coverage, namely NIMA [TM18a] and the Ranking Network proposed by Kong [Kon+16a] with datasets of professional photography. We chose these two networks primarily because the source codes and final weights of the network can be found easily. We present the models and the datasets in this section, as well as our method to

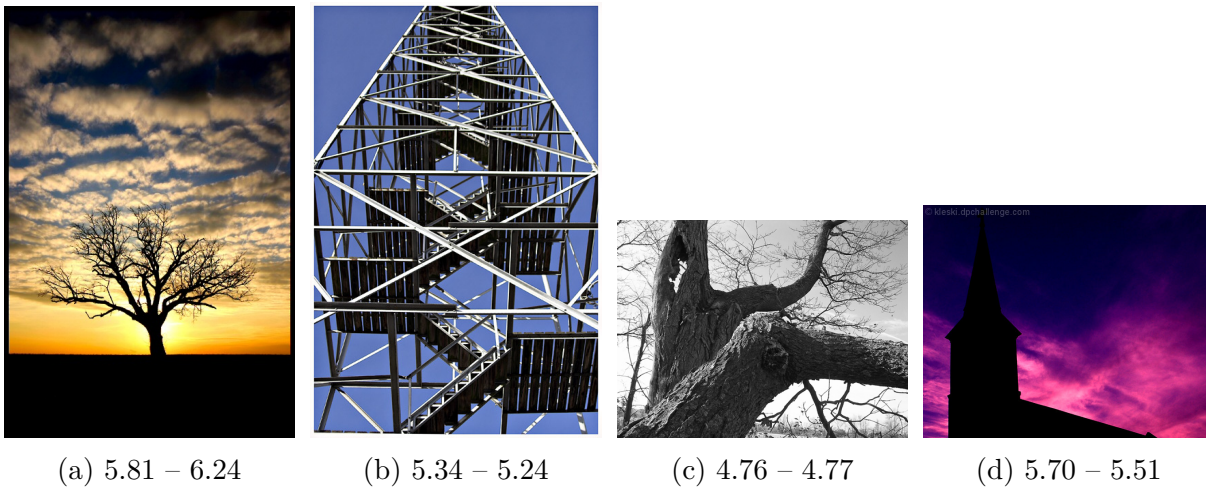


Figure 4.1 – Different images from the AVA dataset (competitive photographs) with scores from the third-party implementation (on the left) – quoted from the original article [TM18a] (on the right).

improve the coverage of the models.

4.2.1 NIMA model

In the following, we present the NIMA architecture we use and the training process. More details can be found in the original paper [TM18a]. The main feature of NIMA model is its ability to rely on existing pre-trained deep networks. It consists in replacing the last layers of an image classification network with a fully-connected layer, and then train only the final layer. The model is first completely trained on ImageNet, and then fine-tuned using a dataset specific to the final use of the network. Two different examples of uses are proposed: (1) prediction of the aesthetics score of an image and (2) prediction of the quality of an image. The former uses the AVA dataset whereas the latter uses the TID2013 dataset. In this study, we focus only on the aesthetics prediction models.

NIMA model relies on existing architectures, two of which are Inception [Sze+16a] and MobileNet [How+17]. Those are architectures for image classification. NIMA adapts those models to the problem of aesthetics assessment. According to the authors, the model based on Inception is more accurate, but slower than the model based on MobileNet.

A third-party implementation is available online¹. This implementation also proposes a NIMA model based on NasNet [Zop+18], another neural network for image classification.

1. <https://github.com/titu1994/neural-image-assessment>

NasNet-based NIMA model was not presented in the original article, but it performs better on AVA than the models based on Inception or MobileNet. As NIMA outputs a distribution of scores, the performance of these models is measured with an Earth Mover Distance (EMD). NasNet gets 0.067 EMD while Inception gets 0.070 EMD and MobileNet gets 0.080 EMD (lower is better). In the following, we perform our study on the most relevant architectures which are Inception (the best method presented in the original article) and NasNet (the best method available in our third-party implementation).

Before going further, we have checked that the behaviour of the third-party model and the original NIMA model are similar, although the scores of images are not exactly the same. Figure 4.1 illustrates some examples with both predicted scores. To see if the two implementations are similar, we compute the correlation coefficient between the scores given by the third-party implementation and the ground truth score of AVA, then we compare it to the given correlation coefficient in the original article [TM18a] between the ground truth scores and the scores given by the original network. On a sample of 4662 images from AVA, we achieve a correlation coefficient of 0.581 between the ground truth and the third-party implementation, which is close to the 0.636 announced in the original article.

4.2.2 Ranking Network model

The Ranking Network was proposed by Kong et al. in 2016 [Kon+16a]. The input of the model is two images passing through two identical networks. The output is a score for each image (on a scale from 0 to 1) and a ranking between the two images. Besides, the model outputs several characteristics of the images (compliance with the rule of thirds, presence of symmetry, of vivid colors, etc...) that were used to compute the final score. As the model needs this information for training, the authors also devised a new training database called AADB. The implementation was provided by the authors themselves on their GitHub². More details can be found in the original paper.

4.2.3 Datasets of professional photography

In this section, we present the photograph datasets we used to test the above presented models on professional photography. In order to cover a wide range of professional photograph, we used six datasets corresponding to different photography genres. Among the

2. <https://github.com/aimerykong/deepImageAestheticsAnalysis>

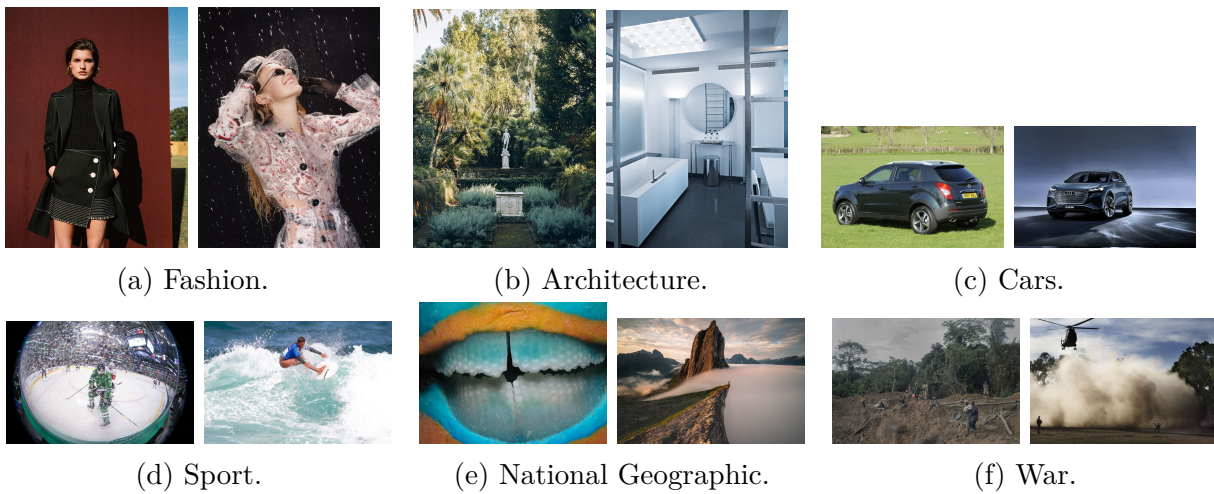


Figure 4.2 – Different sample images from professional datasets. We show here the worst (on left) and the best (on right) of each category according to NIMA.

six datasets, we have created five of them by collecting images by hand. As the process of collecting images one by one is very time consuming, we have collected around 100 images per category, as it is sufficient to test the different models. Figure 4.2 illustrates a sample of images of these photography genres.

Fashion category contains photographs coming from various editorial fashion photo-shoots and published in fashion magazines during the year 2018. These photos are captured by professional photographers in collaboration with magazine art director. The photographs are of different aesthetic styles (black and white, color, high key, low key, etc.) These images are not only of high aesthetic quality but also have an artistic dimension. These images have been collected for a long time and thus, the dataset collected is larger than the other categories. We have collected 1373 images.

Architecture category contains real estate, indoor design and architecture photographs published in *Architectural Digest Magazine*. These photographs have been captured by professional photographers and promote the beauty of architecture. They are of high aesthetics quality but in a more classic manner than fashion ones. We have collected 117 images.

Cars category contains photographs done by various car manufacturers to advertise their new cars. Due to marketing strategy, the aim of these photographs is to promote different values such as power, robustness and sometimes beauty. We have collected 109 images.

Sport category contains photographs of various sports from the French journal

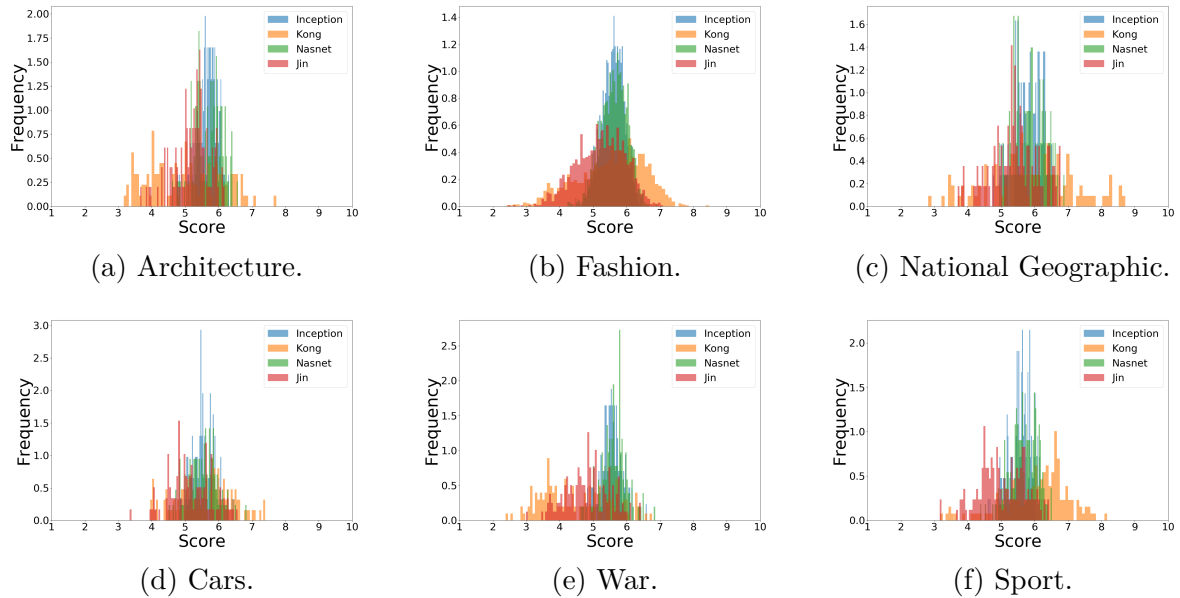


Figure 4.3 – Histograms of scores for the four models (Inception-based NIMA; Nasnet-based NIMA; the Ranking Network; and the Weighted CNN network) and for six different datasets.

L'équipe dedicated to sport news. These images have been captured by professional photographers and try to capture crucial moments. We have collected 155 images.

War category contains war photographs from the photography agency called *Agence VU*. Obviously the first aim of these photos is not to make aesthetically pleasant photos, but to tell the truth about war. We have collected 138 images.

National Geographic category contains wildlife and landscape photographs from the National Geographic website. Similarly to the previous datasets, the photographs have been captured by professional photographers. The usual objective of these photographs is to show the beauty of Earth and wild life. We have collected 110 images.

4.2.4 Overview of models

Figure 4.3 shows histograms of scores for images from the professional datasets we have built presented in section 4.2.3 computed by the different models we tested.

We can draw several observations from these distributions. First, we notice that the histograms for NIMA (green/blue on Figure 4.3) and for the Ranking Network (named "Kong", in orange on Figure 4.3) are extremely different from each other. Two histograms

are noteworthy: War and Sport from Kong model. The War histogram (Figure 4.3e in orange) is well below the average score while the Sport histogram (Figure 4.3f in orange) is well above the average. The scores from War were expected. While they come from a professional dataset, their main goal is not to have a high aesthetic value, but to show the truth of war. Therefore, they are bound to have low aesthetics values. On the other hand, the sport images are quite colourful and with a low depth of field. These are two qualities that the Ranking Network assesses to compute the final aesthetic score. Therefore, they get higher score than average.

Several observations can be made from these distributions. First, we observe that all histograms are located near the mid-point of the evaluation scale. The number of ugly and beautiful images is almost zero. This observation is not what was expected for the categories Fashion and National Geographic. Indeed, these categories gather beautiful images, for which photographers pay attention to composition, lighting, depth of field, etc. Therefore, those images should get, for most of them, a high aesthetic score. For other categories, such as war, it was also expected to get some bad aesthetic scores. The second observation is related to the rather small difference in terms of aesthetic scores between competitive and professional photographs. These two observations may suggest that the NIMA models do not generalise well because they specialise in images from AVA (this specialisation is what we call over-fitting). To make this point clearer, we have conducted a statistical analysis presented in Section 4.3.2.

Several differences exist in NIMA and the Ranking Network. First, the architectures of the associated networks are different. On the one hand, the authors of the Ranking Network carefully designed the architecture using several pipelines corresponding to different aesthetic attributes. On the other hand, NIMA uses a generic features extraction learned from classification tasks. While it has been shown that networks which use features learned from classification can perform well for various tasks in computer vision [Wen+19], features crafted for a specific task (in our case aesthetic quality prediction) are more relevant. Besides, aesthetics does not rely only on signal patterns [Jos+11b] but also on more subjective criteria. That is why generic features learned from classification (which is a problem with an objective answer) might not be optimal for aesthetics prediction. That could explain why NIMA does not perform as well as the Ranking Network on professional photographs.

4.3 Presentation of experiments

As the results of Section 4.2.4 show that NIMA does not discriminate between competitive and professional images, we have devised a method to improve the coverage of NIMA. We present in this section our method that consists in fine-tuning the trained network with new images.

4.3.1 Hypotheses

We conducted our experiments under some assumptions. These assumptions are presented and motivated in this section.

The images from the category Fashion are expected to have a high aesthetic quality, and therefore high scores. Indeed, the aim of such images is to promote the value of fashion products, so we can argue that the end goal specific to the Fashion category is to have high aesthetic quality. Besides, in the images we collected we can argue that the end goal was achieved as these photographs were taken by professional photographer, and they also were published, which proves that they are acknowledged to be efficient.

On the other hand, the images from the category War are expected to have a low aesthetic value (according to common aesthetic sense) and therefore would get rather low scores.

4.3.2 Fine-tuning of Nasnet-based model

We want to address the over-fitting problem. As we already have a trained model to start with, we do not use classic regularisation techniques. It is easier and faster to train again the network using the professional photographs. Indeed, using the already trained network as a basis requires less epochs and less training image for the second training. That is what we call fine-tuning. We fine-tune the Nasnet-based model using one of the professional datasets. We consider Fashion for two reasons: first, this category contains the most images among all professional datasets we have, second, based on the content and the aim of those photographs, they must have a high aesthetic score according to competitive photography aesthetic. We construct a training dataset using the 1373 images from Fashion and 1373 random images of AVA. Among these 2746 images, 300 were set aside and used as a validation set (150 from Fashion and 150 from AVA). As we use supervised learning methods, we need labels (in our case, the labels are numerical scores)

for our images. Several methods are possible, we chose to create artificial scores for the Fashion images. This method is faster, but less accurate than collecting new scores from observers. However, as the goal of this experiment is only to see whether a fine-tuning can effectively improve the coverage of the network, the absolute scores do not matter. We want to see if we can improve the networks coverage using some professional photographs (increasing scores in Fashion, decreasing scores in War) without impeding on the learning done on competitive photographs. Therefore, a full user study to collect new scores is not necessary here, but is mandatory if the network is designed to give accurate results.

As explained in Section 4.3.1, we make the assumption that the images from Fashion must have a high mean score. We also assume that although the actual scores for Nasnet-based NIMA on the Fashion dataset are not correct, the ranking of the images is. Then, to create artificial scores for the Fashion images, we shifted the average value of the histogram of scores given by Nasnet-based NIMA to a higher value. More specifically, if the current mean score for the dataset on Nasnet-based NIMA is μ , the ground truth score for each image of score s is computed as $\bar{s} = s - \mu + \bar{\mu}$ where $\bar{\mu} = \mu_{high} + 1$ is the new mean score. We considered the value $\mu_{high} = 6$ because it corresponds to high aesthetics value according to many previous work [Wan+18; MMP12]. Using this method, we ensured that the fashion dataset had a mean score of $\bar{\mu}$, and therefore, a majority of images from Fashion had a score higher than μ_{high} .

4.4 Results

We present in this section the results of our two experiments: the behaviour of the different models when exposed to professional photographs, and the outcome of the fine-tuning process on NIMA.

4.4.1 Statistical analysis

Table 4.1 presents the mean scores and whether the paired t-test for all pairs of scores computed by the tested models are significant or not.

NIMA models. Results indicate that the mean score of ground truth scores (5.1) is significantly lower than the mean score of tested professional photographs (ranging from 5.52 to 5.84). This difference, although very small on a scale of 10 grades, was expected and statistically significant. From the t-test analysis, we can infer a clustering of datasets

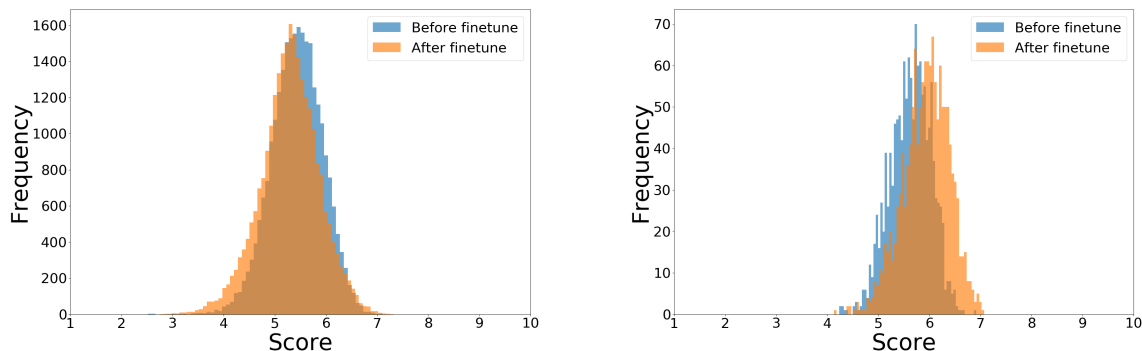


Figure 4.4 – Influence of the fine-tuning process using the Fashion dataset on Nasnet-based NIMA scores. Results are reported for AVA (left) and Fashion (right).

depending on the p-values of the t-test. However, these clusters are quite different for the different models. They are represented on Figure 4.5 (a) and (b), and given below with clusters between parentheses in ascending order of aesthetic scores:

- For NasNet-based NIMA model: (AVA), (AVA), (Cars), (War; Fashion; Architecture; Sport) and (National Geographic).
- For Inception-based NIMA model: (AVA), (AVA), (War; Fashion; Cars; Sport), (Architecture) and (National Geographic).

Ranking Network model. As the distributions are not normal, we did not use a t-test, but a Wilcoxon rank-sum test. We observe that the range of scores is greater than for NIMA; the dynamic of scores represents 54.4% of the whole scoring scale, whereas NIMA scores represent only 22.3% of the scoring scale in average. This observation would suggest that the ranking model is more selective for the professional categories. As previously, Figure 4.5 (c) represents the clustering inferred from statistical analysis. The clusters for the Ranking Network are: (War), (AVA; Architecture), (Cars; Fashion; AVA) and (National Geographic; Sport).

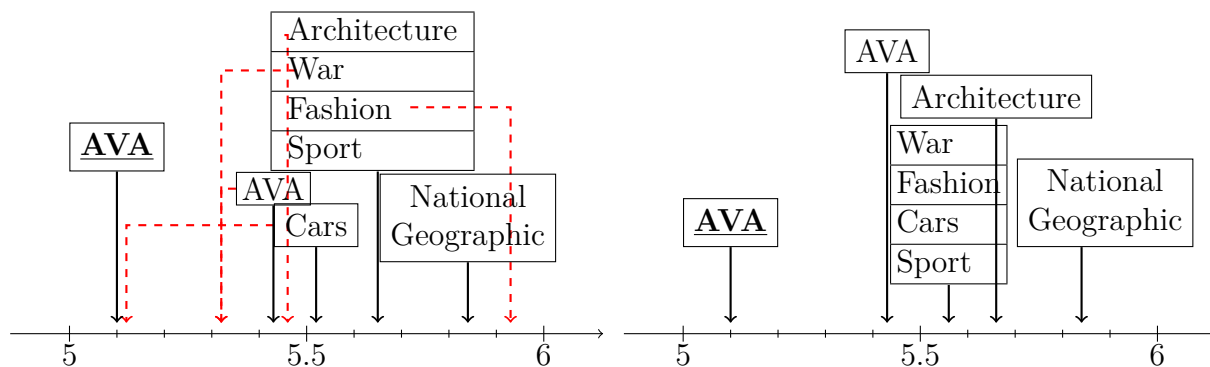
We notice that in both cases, AVA and National Geographic are separated from the others. The mean scores of datasets Cars and Architecture do not vary much between the models, so we can argue that their scores are more reliable than the score of the four other datasets. In any case, the p-values show that there are significant differences between some photography categories. This could come from differences in the aesthetics features of the images.

	AVA	NG	C	F	A	S	W
NIMA NasNet Mean Score	5.10	5.84	5.52	5.63	5.69	5.67	5.62
AVA		***	***	***	***	***	***
Nat.Geo.			***	***	*	**	***
Cars				*	**	*	ns
Fashion					ns	ns	ns
Archi.						ns	ns
Sport							ns
War							
NIMA Inception Mean Score	5.10	5.84	5.55	5.59	5.66	5.57	5.53
AVA		***	***	***	***	***	***
Nat.Geo.			***	***	**	***	***
Cars				ns	**	ns	ns
Fashion					**	ns	ns
Archi.						*	***
Sport							ns
War							
Ranking Net Mean Score	0.455	0.545	0.510	0.503	0.443	0.565	0.386
AVA		***	***	***	ns	***	***
Nat.Geo.			*	**	***	ns	***
Cars				ns	***	***	***
Fashion					***	***	***
Archi.						***	***
Sport							***
War							

Table 4.1 – Table of paired t-test (or Wilcoxon rank-sum) p-values for different datasets (AVA=Ground truth scores; NG=National Geography; C=Cars; F=Fashion; A=Architecture; S=Sport; W=War) on the three models. The stars are attributed using the p-values: * for $0.05 \geq p > 0.005$, ** for $0.005 \geq p > 0.0005$, *** for $0.0005 \geq p$; *ns* stands for non significant. For the Ranking Network, the scale is from 0 to 1 instead of 1 to 10. The scores given for each dataset correspond to the mean for this particular model.

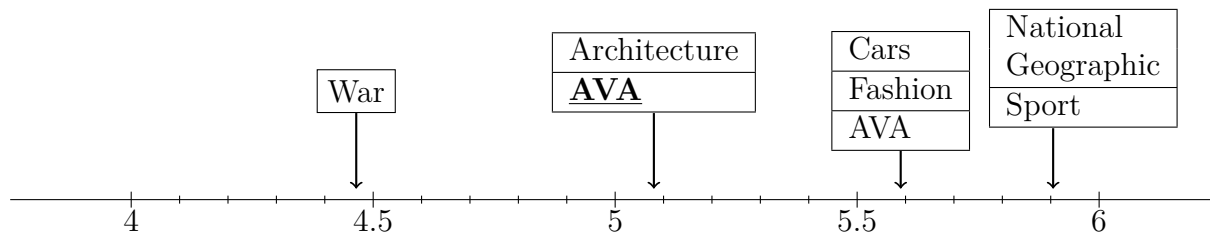
4.4.2 Does fine-tuning Nasnet-based NIMA model improve the overall prediction capabilities?

Figure 4.4 presents predicted scores when the Nasnet-based NIMA model is fine-tuned following the procedure described in Section 4.3.2. The blue histograms represent the score distribution of AVA and our Fashion dataset on the original networks, and the red histograms those after fine-tuning the model. As expected, we notice that the scores of Fashion have increased after fine-tuning while AVA scores slightly decreased. The fine-tuned model is then able to better discriminate Fashion photography from competitive photography. It may suggest that models trained over AVA are specialised for competitive photography. However, a simple fine-tuning process allows to make the model more generic and more relevant. The results of the statistical tests for the fine-tuning experiment are



(a) NasNet. Red arrows show the modifications of the mean scores brought by the fine-tuning: Sport and National Geographic are not modified whereas War and AVA are in the same cluster.

(b) Inception.



(c) Ranking Network.

Figure 4.5 – Clustering the datasets according to the p-value of the statistical tests. The arrows point to the mean aesthetic score given by the corresponding model. AVA corresponds to the average estimated score of 26122 images of AVA on the model and AVA corresponds to the ground truth mean.

not represented here, but a representation of the clusters can be found in Figure 4.5 (a). Results indicate that after fine-tuning, the mean scores of the datasets have slightly changed. As we used the Fashion dataset to train the model using high scores as ground truth, it was expected that Fashion photography would get a higher score after fine-tuning. We observe that War, Cars and Architecture have lower mean scores. This must come from the modification of the features used by the network.

4.4.3 Comparison with weighted CNN

As the proposed fine-tuning process and the weighted CNN [JSS16] have the same goal (reducing the bias caused by the imbalance of AVA), we compare the performance of both

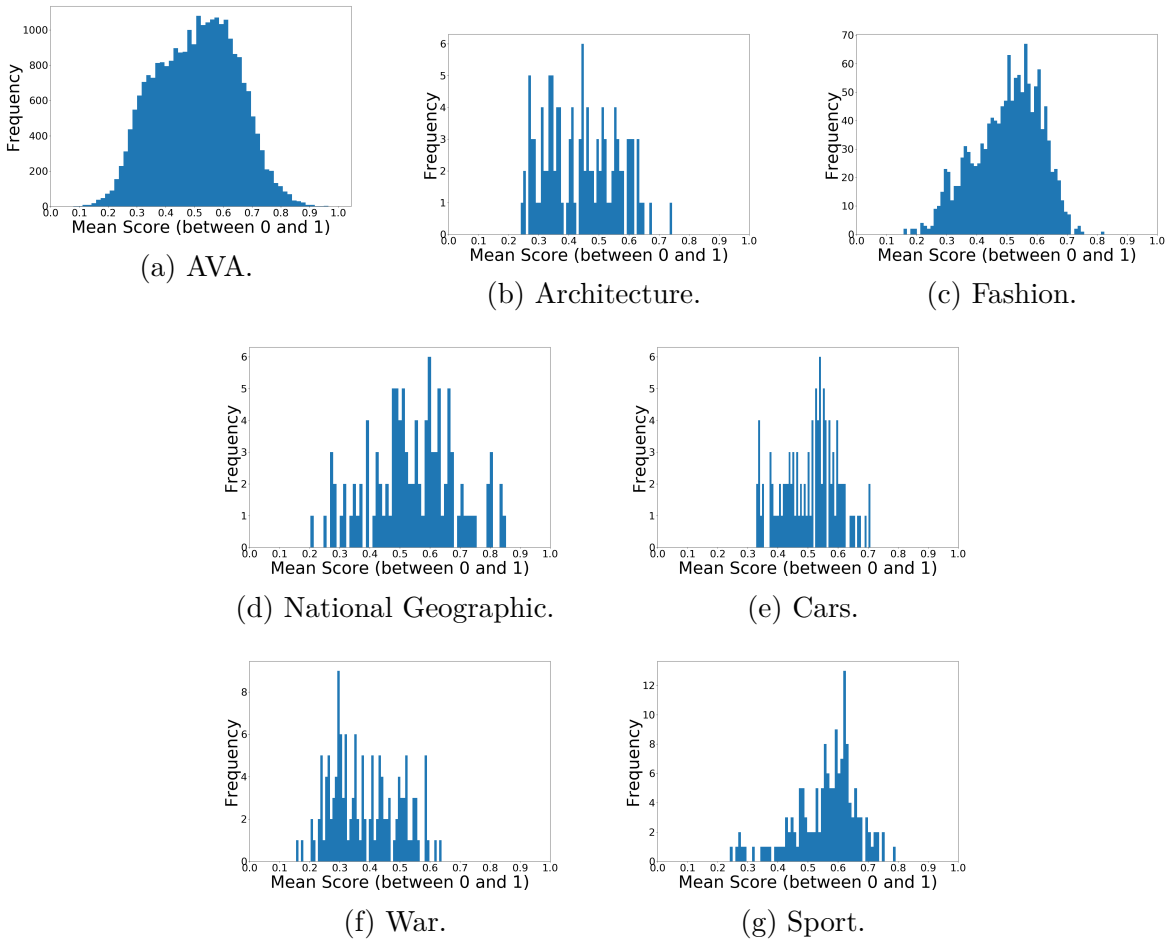


Figure 4.6 – Histograms of scores for different categories on the retrained Ranking Network.

networks. The model and weights of Jin et al. [JSS16] are available on their website³. We can then compare our Nasnet-based NIMA model fine-tuned with Fashion and the original weighted CNN model. Besides, to assess the relevance of their proposed novel loss function, we trained another model – namely the Ranking Network [Kon+16a] – on AVA using their loss function. Figure 4.4 presents our datasets outputs on Nasnet-based NIMA fine-tuned (in orange) and Figure 4.3 presents the results for the original weighted CNN model (in red) and Figure 4.6 presents the results for the Ranking Network retrained using the weighted loss function.

The histograms in Figure 4.6 and 4.3 (in orange) are very similar: the fine-tuning process has a very slight impact on the Ranking Network. We will therefore focus on the

3. https://ivrlwww.epfl.ch/bjin/project_aesthetics/Image_Aesthetics.html

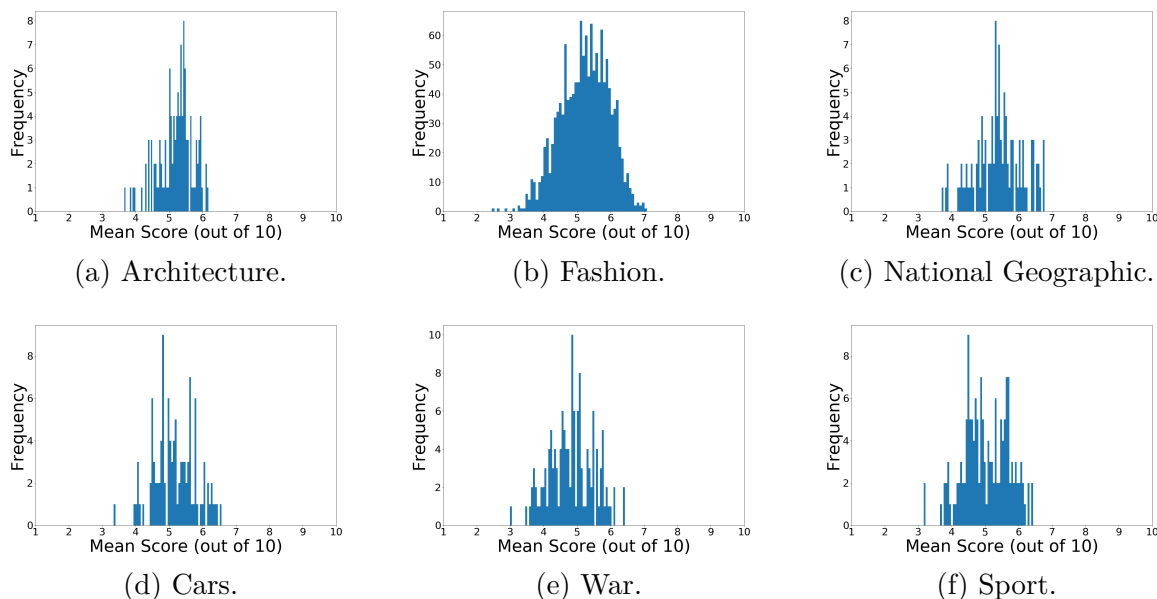


Figure 4.7 – Histograms of scores for different categories on the weighted CNN proposed by Jin et al. [JSS16].

comparison of the two other models. The histograms from Nasnet-based NIMA (Figure 4.4 in orange) have smaller dispersion than the ones from the weighted CNN (Figure 4.3 in red). Therefore, the model from Jin et al. seems to be more able to reduce disparity in high and low score value zones. However, we notice a significant difference in the average scores for the Fashion category in both models, Jin et al. being the lowest. It is thus possible that the weighted CNN significantly improve the dispersion of the score histogram, but is overall less accurate on the score themselves. We can verify this using a correlation metric.

Therefore, we compute the Pearson correlation metric, – as well as the Mean-Square Error (MSE) to measure the differences per image – with AVA for different models: our Nasnet-based NIMA fine-tuned with the Fashion dataset, the original Nasnet-based NIMA and the weighted CNN proposed by Jin et al. These values are reported on Table 4.2.

	MSE (\downarrow)	Pearson ρ (\uparrow)
Original NIMA	0.387	0.618
Fine-tuned NIMA	0.391	0.596
Jin et al.[JSS16]	0.500	0.585
Retrained Ranking Network	1.939	0.165

Table 4.2 – Mean square error and Pearson correlation metric for four different models.

First, we notice that the best model according to both metrics is the original Nasnet-based NIMA. Our fine-tuning process slightly degrades the performances of NIMA on AVA. However, as explained previously, it also significantly improves the performances on the Fashion dataset. The fine-tuning process is thus quite effective and allows for better results on Fashion, and good results on AVA. In terms of correlation, the weighted CNN is worse than NIMA and the fine-tuned NIMA. This shows that our method is more faithful to the ground truth scores. All of this proves that our fine-tuning process is a real improvement over the weighted loss function.

4.4.4 Discussion

Using the fine-tuning method, we managed to increase the score of the Fashion database without modifying too much the scores from other categories. If we assume that the Fashion dataset is mainly composed of high aesthetic quality images, we have effectively improved the model accuracy and coverage. However, we can discuss the relevance of our assumption.

The score threshold used as high aesthetic quality is used in previous work as a distinction between professional and amateur photographs. The professional photographs we used in our experiments were chosen because of their relevance. Indeed, these photographs were taken by professional photographers, furthermore they were published, which proves that they are acknowledged to be efficient. This shows that our assumptions on the scores (of images from War and Fashion) are reasonable.

4.5 Conclusion

In this chapter, we have presented a study based on the models NIMA and the Ranking Network. We wanted to understand how these models behave for other kinds of photography than their own training dataset. As the dataset AVA is composed of competitive photographs, we have chosen six datasets of professional photographs in order to test whether or not the models generalise well for these photographs.

We have observed that NIMA and the Ranking Network have different behaviours. NIMA gives scores with rather small deviation around the mean, whereas the Ranking Network scores are much more spread on the rating scale. We have also noticed that, for NIMA, there is a strong discrepancy between the scores of AVA and professional pho-

tographs. This is alleviated by the fine-tuning process, but raises other issues, such as the correct way to fine-tune and train the networks. Despite the fact that there is a significant difference of scores for the datasets, meaning that NIMA models actually recognise photographs from professional sources as more beautiful than images from AVA, this study raises a number of issues. First, we do not observe, but we were expected to observe, a strong discrepancy between the scores of AVA and professional photographs. Second, all aesthetic scores of professional photographs are in a very limited range, approximately from 5 to 6, whereas AVA scores span from 3 to 8.

These observations reflect how far we are from accurately predicting the aesthetics of an image. However, we have demonstrated that fine-tuning existing models with professional photography can reduce the specialisation of existing models to competitive photography. This work shows the importance of the training dataset, especially when dealing with such a difficult notion as aesthetics.

There are several future improvements of this work. First, to improve the training datasets of aesthetics assessment networks, we could define and provide the community with a new annotated image dataset of professional photographs. Then, by looking more in details into a network, we could find how a network discriminates between different kinds of photographs, and eventually better understand and model human aesthetics feeling. This could lead to the design of novel personalised aesthetics prediction architectures that are based on new generic techniques used for example in BIQA (as presented in Section 3.1). Finally, as our work does not take into account the standard deviation of scores, a more thorough study about means and standard deviations would also be interesting, because this could help characterise the categories of professional photography.

HDR-LFNET: INVERSE TONE MAPPING BY FUSION METHODS

5.1 Introduction

High Dynamic Range (HDR) images consist of 2D arrays which contain, for each pixel, the raw luminance captured by the sensor at this point. These values are, in theory, unbounded and not quantized. This is a much richer content than Standard Dynamic Range (SDR) images, which correspond to the common widespread images. HDR-compatible hardware, such as monitors or cameras, are more and more available to the general public. This creates a need for image processing algorithms adapted to HDR content, making the HDR imaging a trending research domain.

HDR photographs are usually obtained using the exposure fusion technique. We take several photographs of the same scene with different exposure times. The longer the exposure time, the more light is coming through the camera, and the more information we can obtain in dark parts. On the other hand, short exposure time pictures provide information in the brightest areas of the scene. This allows to gather information in every area of the scene. The different photographs are then merged together to yield a single HDR image. Modern cameras can take several photographs with different exposure times using a single press of the trigger. This function is called the bracketing function, and it allows to take the specific photographs that are needed for exposure fusion. That is why exposure fusion to generate HDR is so popular.

As most of the image datasets available are composed of SDR images, it is useful to devise algorithms to recover lost information in SDR images. From a single SDR image, we must use approximations if we want to extend the dynamic range. Algorithms that extend dynamic range of SDR images are called inverse tone mapping operators (iTMO). Contrary to the classic method of exposure fusion, iTMOs only take as input a single SDR image and yield an HDR image. The result is an approximation of the truth, as SDR

images do not contain as much information as HDR ones. However, iTMOs manage to avoid some artifacts caused by the exposure fusion, such as ghosting (due to movement in the scene), motion blur or Moiré effect (usually present in high exposure pictures). While at first, iTMOs were based on content-based assumptions and were using photographic rules to extend the dynamic range, neural networks are now used to this end.

Inspired by other computer vision domains (such as saliency [MNL13] or denoising [Ker14]), we devised the HDR Light Fusion Network (HDR-LFNet), a new iTMO aggregating several existing iTMOs that are not based on learning algorithms. Hopefully, the fusion of all iTMOs performs better than each input method individually. We work with standard gamut, HDR images. As such, the main component at stake here is the luminance. Besides, we know that processing possibly unbounded values in a neural network can lead to difficulties in the training. Therefore, instead of luminance, we process lightness values, which are bounded and closer to the human perception than luminance. That explains why we have decided to mainly process lightness, while using a generic recolourization algorithm (explained in Section 5.2.4) to generate coloured images.

Our approach uses a supervised neural network with several orders of magnitude fewer parameters to learn compared to existing networks, which is achieved thanks to some pre-processing. This pre-processing lowers as well the number of training images needed. This is a real improvement, as our network needs HDR training images and HDR images are not as easy to gather as SDR images.

However, as we need fewer images to train, each image is increasingly important and needs to carry a lot of information to effectively train the network. We consider that high resolution images have a higher probability to contain the qualities that we need (light sources, dark areas, smooth gradient and high frequency areas). The HDR datasets that already exist usually contain a few hundreds images at most, with size around (1920×1080) . To train our network, we then collected high resolution HDR images, and we compiled them in a new dataset.

We tested our method against state-of-the-art using four metrics on HDR images: HDR-VDP2, Harmonic HDR-IQA, PU-PSNR and PU-SSIM. Our method yields results similar to the state-of-the-art, but runs with fewer parameters. We also conducted a user experiment to compare HDR-LFNet to state-of-the-art methods. Results show that HDR-LFNet is preferred to others in the user study. Along with the short training time required by our network, this user study shows that our method is usable and effective in a wider range of applications.

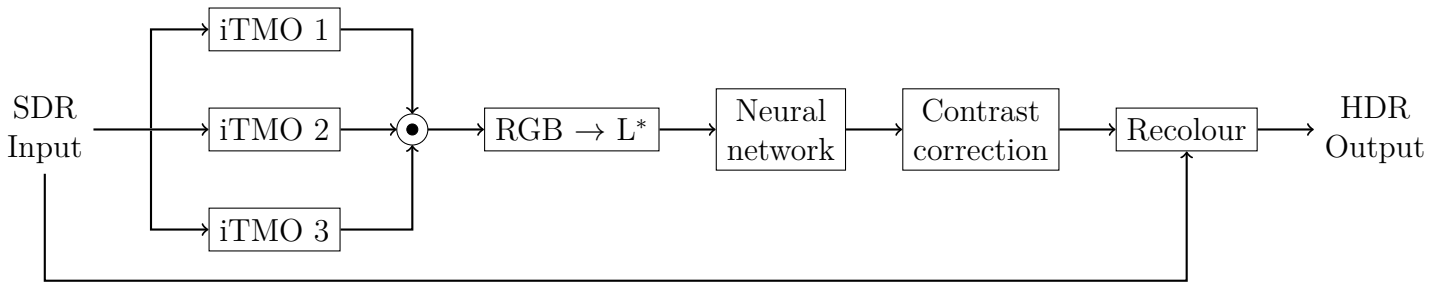


Figure 5.1 – Our method HDR-LFNet uses a neural network to fuse several expanded versions of the input. This allows for faster training and lighter network. We use a contrast correction and colourization as postprocessing. The operator \odot corresponds to the concatenation along the depth dimension: if I_1 and I_2 are two images of size (H, W, C) , then $I_1 \odot I_2$ is a volume of images of size $(H, W, 2, C)$.

Our contributions consist in:

1. devising a novel inverse tone mapping light architecture that merges several existing iTMOs to get a more powerful one;
2. proposing a new dataset of high definition HDR images composed of 496 pairs of middle exposure SDR and the corresponding HDR;
3. evaluating our work and other state-of-the-art methods using objective metrics and a user experiment.

These contributions led to a submission [Cha+23]. The rest of this chapter is divided as follows. We explain our approach in section 5.2 and then present the evaluation of our method in section 5.3. Finally, section 4.5 concludes the chapter.

5.2 Our fusion network

Our goal is to propose a new method to expand the dynamic range of images that is faster and requires fewer training images. To achieve this goal, we use supervised learning algorithms to train a neural network. In this section we present the network architecture and the training process along with the training dataset. As we decided to use expanded version of images through iTMOs as input of our algorithm, we also present how we choose those operators. Our method is represented in Figure 5.1.

5.2.1 Overview of 3D convolutions

In this section we present some general characteristics of 3D convolutions, that we extensively use in our network.

An image can be represented as a tensor of size (H, W, C) with H the height, W the width and C the number of channels. A channel is a scalar array of size (H, W) containing information about a specific characteristic. For example, in traditional coloured images, $C = 3$ (each channel contain information about a specific colour component for example in *RGB* values), or if we only use gray level images, $C = 1$. Convolutional neural networks work by learning the weights of convolution filters. At each level (that we call layer) in the encoder section, the number of channel increases, to learn more and more structured information. For example in our proposed architecture, the deepest layer contains 64 channels. Tensors which are not the input or the output tensors of a neural network are called feature maps.

A classic 2D convolution in a CNN layer is characterized by its kernel size, denoted by (k_x, k_y) . A 2D convolution between a layer with C channels and another one with C' channels will have $k_x \times k_y \times C \times C'$ parameters to learn. All scalar values in a (k_x, k_y, C) voxel of the input tensor are multiplied term by term with the kernel weights, and then added together to yield a single scalar value. This process is repeated C' times with C' different sets of weights, to get C' values. Starting with a tensor of size (H, W, C) , this convolution yields a tensor of size $(H - k_x + 1, W - k_y + 1, C')$.

In our case, we use 3D convolutions, which work on volumes of images of size (H, W, D, C) . The convolution kernels also have one more dimension (k_x, k_y, k_z) , but they are similar to 2D convolution: they have $k_x \times k_y \times k_z \times C \times C'$ parameters to learn, and starting with a tensor of size (H, W, D, C) , this convolution yields a tensor of size $(H - k_x + 1, W - k_y + 1, D - k_z + 1, C')$. Besides, as 2D convolutions, they consider all channels when computing the convolution sum. The advantage of 3D convolutions in our case is to specify the depth dimension: this allows for more control over the training of the network, and its behaviour.

2D convolutions and 3D convolutions are related: if we consider tensors with size (H, W, C) as $(H, W, 1, C)$, 2D convolutions (k_x, k_y) are the same as 3D convolutions $(k_x, k_y, 1)$. On the other hand, we can view tensors (H, W, D, C) as (H, W, DC) and convolutions (k_x, k_y, D) are the same as (k_x, k_y) , but convolutions (k_x, k_y, k_z) with $k_z < D$ are not translatable to 2D convolutions, so 3D convolutions are strictly more expressive than 2D ones.

We represent on Figure 5.2 a visualisation of the 3D convolution that we use in our work.

5.2.2 Architecture

As represented on Figure 5.1, our method uses a neural network at its core. The architecture of our network is represented on Figure 5.3. We adopt an encoder-decoder shape to reduce the size of the images during the forward pass, and thus reducing the time and memory needed for training. Many different characteristics of our network are explained in this section.

Our network only processes the lightness of images – the L^* component in the $L^*u^*v^*$ colour space –, and thus the input has one channel. We apply a contrast correction to the network output, as well as adding back colour to yield a final HDR output. These postprocessing operations are presented in Section 5.2.4. The activation function is based on ReLU. As HDR images usually do not contain values of 0, we define a new activation function called Nonzero-Relu as

$$ReLU_a(x) = \begin{cases} a & \text{if } x \leq a \\ x & \text{else} \end{cases} \quad (5.1)$$

with a value of $a = 10^{-12}$.

While designing our architecture, we follow some common guidelines to avoid well known problems. To avoid artifacts usually caused by deconvolutions, we instead upsample the feature maps, and then apply a classic convolution in the decoder part. Besides, along with maxpooling to reduce the dimension of our network, we use dropout layers to stabilize the training and dodge local minima of the loss function.

Inspired by other neural networks [Liu+20; Kon+16b], we decided to lighten our network — in order to improve its performances — by using an architecture specific to our problem. For this purpose we have provided the network with two new characteristics: (1) the input of our network are images expanded with existing iTMOs, (2) 3D convolution layers are used to force the network to learn the added value of each pair of expanded images. These two characteristics are detailed in this section.

By using already inverse tone-mapped images as input of the network, this latter must learn an easier transformation from HDR to HDR rather than from SDR to HDR. These images are concatenated on the depth dimension of the tensor.

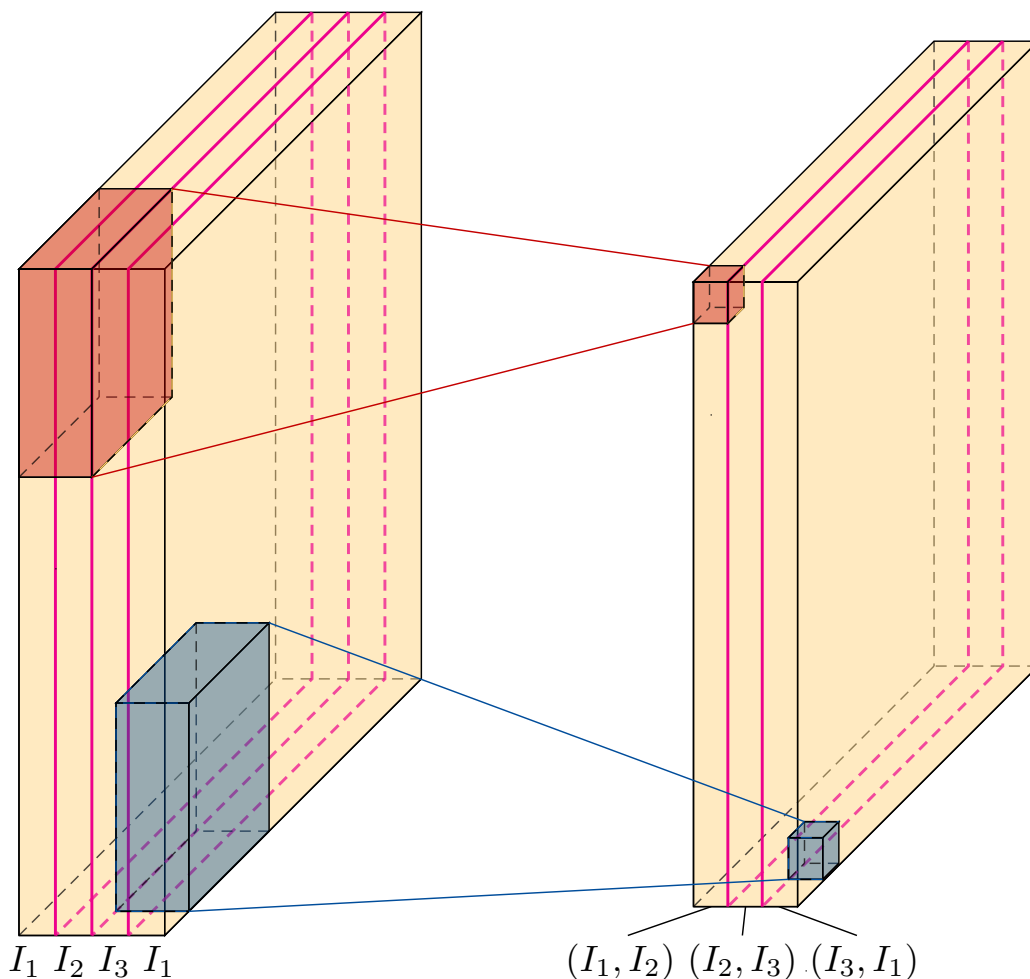


Figure 5.2 – Diagram representing the 3D Convolution used in our architecture. The kernel size of this convolution is $(k, k, 2)$. The input of the convolution (on the left) is a volume of images $(H, W, D = 4, C)$ where the depth D is represented by magenta lines. The output of the convolution (on the right) is then a volume $(H', W', D' = 3, C)$. The red and blue blocks represent two different steps of the convolution.

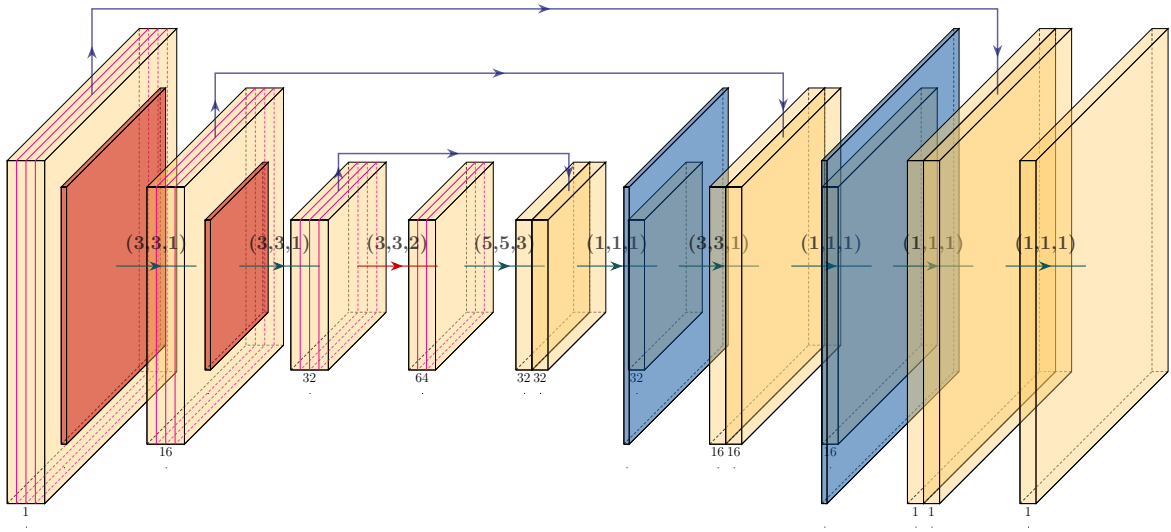


Figure 5.3 – Architecture of our network, which uses 3D-convolutions. The network processes volumes of images (H, W, D, C) (represented in yellow), where depth D is represented by magenta lines. The number below each feature map corresponds to the number of channels C (not graphically represented on the figure). Green arrows represent 3D convolution of kernel size (k_x, k_y, k_z) followed by the activation function $ReLU_a$, blue arrows represent skip connections (which are detailed in Section 5.2.2). The red arrow indicates the 3D convolution that effectively fuse the input images. Downsampling and maxpooling are represented by red-orange layers, and upsampling are represented by blue layers.

We manage to drastically reduce the number of parameters of our network. By using three expanded versions of the same image (obtained with three different iTMOs) as input,

we are able to reduce the number of parameters to approximately 2×10^5 , against the 10^6 to 10^8 parameters of state-of-the-art networks. The choice of iTMOs we use as input of our network is discussed in Section 5.2.6. As the number of trainable parameters is quite low compared to other networks, the function learned by our network should be simpler than the functions learned by state-of-the-art networks, but as the inputs are more complex, the output quality should be at least similar that the quality of state-of-the-art methods. Besides, we use full resolution images during training instead of random crops, this should not impact the performance of the network while reducing the training time. This assumption is verified in Section 5.3.2 by training our architecture with the HDR-Real dataset.

As we have several versions of the same image with only one channel, we can induce the network to learn how each iTMO interacts with the others. To this end, we use 3D convolutions instead of the classic 2D ones to guide the training.

To ensure that each pair of iTMO is considered, we need to input redundant information in our network. Our input tensor is the concatenation (I_1, I_2, I_3, I_1) in the depth dimension with I_i the expanded image generated by the i -th iTMO algorithm. This allows 3D-convolution with depth $k_z = 2$ to process all pairs (I_1, I_2) ; (I_1, I_3) and (I_2, I_3) as explained in Section 5.2.1. This convolution is done at the heart of the network (represented by the red arrow on Figure 5.3). After this convolution, we just need to upsample the feature maps to get back the original resolution. The final convolutions are done using maximum depth 3D convolutions to simulate 2D convolutions, as the depth dimension is not useful anymore. We compare this method to the classic 2D Convolution in the ablation study (Section 5.3.3). Therefore, the input tensor of the network is of size (Height, Width, Depth = 4, Channels = 1). These considerations determine the depths of all feature maps of the networks, and therefore the depth values k_z of all 3D convolutions. In the following, the depth values of 3D convolutions which are irrelevant (because they are already fixed by our previous construction choices) are denoted by a question mark ?.

Finally, we know that using convolution filters usually averages neighbour pixels, and therefore degrades the edges. First, to keep high frequencies as much as possible, we use skip connections. However, the feature maps before and after the fusion convolution have different depths, so we compute the average values of the first feature maps in the depth

dimension before merging them with the second ones. The aggregation of the feature maps is done by concatenating both feature maps, and then running a convolution with a kernel of $(1, 1, ?)$. Moreover, we modify the kernel sizes of the convolutions of the decoder part in a coarse-to-fine manner. Indeed, we decrease the kernel size from $(5, 5, ?)$ to $(1, 1, ?)$. This allows the final convolution to be a $(1, 1, ?)$ convolution, which better preserves the edges in the images. The impact of the kernel size decrease in the decoder part is discussed in the ablation study (Section 5.3.3).

5.2.3 Loss function

Our loss function is designed for comparing HDR images. We denote by $L(I_{HDR})$ the lightness of the ground truth HDR image, and by \hat{L} the output of the network. The loss function used for training consists of four parts: (i) a mean absolute error (MAE) $Y_c = \mathcal{L}_1(L(I_{HDR}), \hat{L})$ for the actual values, (ii) a gradient-based error (gMAE) $Y_g = \mathcal{L}_1(g(L(I_{HDR})), g(\hat{L}))$ – with g the gradient computation using Scharr filters – to emphasize the shapes, (iii) a perceptual loss Y_p to be more accurate on areas that are sensitive, and (iv) a dynamic range error Y_d .

The perceptual loss is based on VGG16 [SZ15]. The idea is to compute visible errors at different scales, and to that end we use activation maps from an already trained VGG network. We compute the mean absolute error between deep features of the target and the output images at the first four layers. The computation is done using a \mathcal{L}_1 difference.

Finally, we add a dynamic range error Y_d :

$$Y_d = \left| D(L(I_{HDR})) - D(\hat{L}) \right| \quad \text{with} \quad D(X) = \max(\log X) - \min(\log X)$$

This dynamic range error is to ensure the network produces an output with a dynamic range similar to that of the ground truth. The actual loss function is then a combination of those four components:

$$Y = \alpha Y_c + \beta Y_g + \delta Y_p + \varepsilon Y_d \quad (5.2)$$

Using the validation set, we found that the values $\alpha = 1$, $\beta = 0.3$, $\delta = 0.15$ and $\varepsilon = 1$ work best.

5.2.4 Postprocessing

As our network only processes the lightness of images, we need some postprocessing to at least add colour to the output image. Besides, we add a contrast correction to the output of the network to better match the ground truth image. We explain this process in this section.

In the following, we denote by I the coloured SDR input and I_X its colour component X (the red, green, or blue channel in our case); I_Y the luminance channel of the image I ; \hat{L} the output lightness of the network; and \hat{I} the HDR output recoloured in RGB with our method. To recolor our HDR images, we use the luminance preserving formula proposed by Mantiuk et al. [Man+09]:

$$\hat{I}(\gamma, s)_X = Y_{exp}(\hat{L}, \gamma) \left(\left(\frac{I_X}{I_Y + 10^{-5}} - 1 \right) \times s + 1 \right) \quad (5.3)$$

with γ the contrast factor, $s \in [0; 1]$ the saturation factor and $Y_{exp}(\hat{L}, \gamma)$ the expected HDR luminance. Because this formula was designed to preserve luminance, we ensure that $\hat{I}(\gamma, s)_Y = Y_{exp}(\hat{L}, \gamma)$. We add the value 10^{-5} to avoid problems with luminances of zero. We convert the output lightness computed by the network to luminance using a power function $Y_{exp}(\hat{L}, \gamma) = \hat{L}^\gamma$. The work of Mantiuk et al. also contains a method for automatic colour correction, however, their studies show that this method is not applicable to HDR images.

The saturation factor s and the contrast factor γ can be modified to yield different results. To get the maximum of correlation between the ground truth and our modified output, we choose both factors by minimizing the mean square error between the images of our training dataset and the computed image from our network. As human beings are more sensitive to order of magnitude of light values rather than absolute values, we compute those differences in the log-domain. These considerations yield the equations 5.4 and 5.5, where T is the set of training images.

$$\gamma^* = \arg \min_{\gamma} \sum_{I \in T} \left\| \log_{10}(L(I_{HDR})) - \gamma \log_{10}(\hat{L}) \right\|_2^2 \quad (5.4)$$

$$s^* = \arg \min_s \sum_{I \in T} \left\| I_{HDR} - \hat{I}(\gamma^*, s) \right\|_2^2 \quad (5.5)$$

After optimizing these formulas, we consider the parameters $\gamma^* = 2.659$ and $s^* = 1$.

Dataset	Nb. of HDR images	Images size
HDR-Eye [Nem+15]	46	(1920 × 1080)
DEIMOS [Klí+11]	79	(4300 × 2900)
pfstools [Man+07a]	8	Variable
HDRPS [Fai07]	105	(4300 × 2900)
HDR-Real [Liu+20]	480*	Variable
Our dataset	496	~ (6000 × 4000)

Table 5.1 – Characteristics of different HDR datasets. (*This is the number of original HDR images in the training set, but this dataset is composed of more than 19,000 crops of size (512 × 512))

5.2.5 Training dataset

Because we use supervised learning methods, we need an image dataset with HDR ground truth and SDR input images to train our network. We present our new training dataset in this section.

The training dataset is an essential component of a neural network. It is mandatory that we carefully select the right images regarding the architecture and our needs. The network we devised includes fewer parameters than the state-of-the-art networks. To train such a network, we need a sufficient number of images – which is less than other state-of-the-art networks –, but each image must contain as much information as possible. To do so, we need very high resolution images. These images may be found in currently available datasets, but they are not easily recoverable (see Table 5.1 for a comparison of the different existing HDR datasets). Therefore, we collect a new HDR image dataset.

This dataset is mandatory to train our network, but can be added to other datasets: the same data augmentation techniques can be applied to yield several thousands of smaller images. As such, our dataset can be used for the same purposes as other datasets.

We have taken photographs with a Sony Alpha 7 III camera. We use the bracketting mode to take 3 exposures at -3, 0 and +3 exposure values. The ground truth HDR image is obtained through exposure fusion using Photoshop algorithm, as it provides images with less artifacts. For the SDR input, we choose the middle-exposed photograph as it contains information balanced between dark areas and bright areas. Besides, colours are usually more saturated in low light environments, and less saturated in high light environments. The middle-exposed shot provides the best colour quality among the three exposure photographs. Each element of our dataset is then composed of the fused HDR image and the middle-exposed image as the SDR image input. Some examples of images

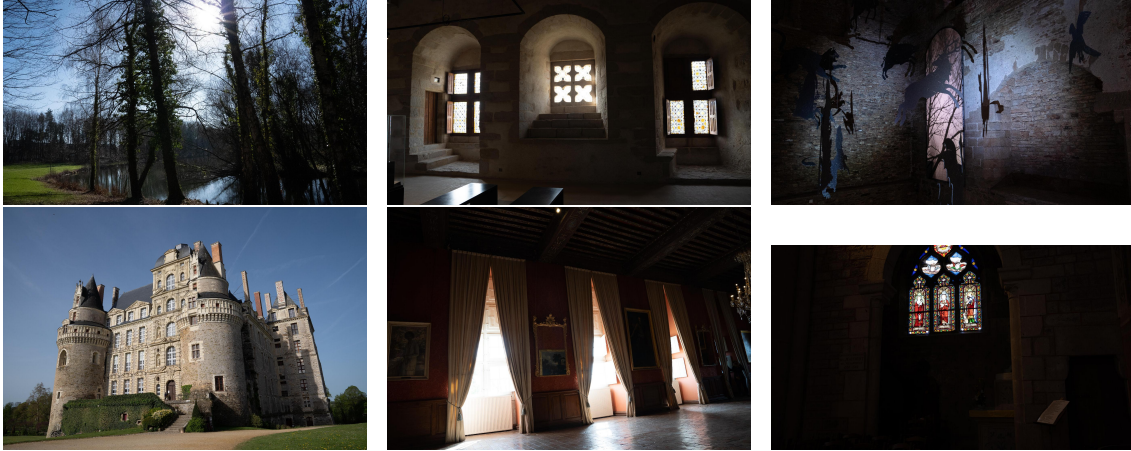


Figure 5.4 – Examples of SDR images from the proposed dataset.

are shown on Figure 5.4. We have managed to take 496 HDR photographs. We have done no photometric calibration, so all the images are provided in relative luminance values. We then have to normalize all images to use them as training images: we divide all image RGB values by the maximal RGB value in the dataset. As all images were taken with the same camera, this ensures that every HDR image has values in $[0; 1]$ while maintaining homogeneity.

To train our model, we split this dataset in three parts: 80 images for testing the model; 56 images for validation; and the remaining 360 images for training. The 360 training images are then flipped horizontally and vertically to yield an effective training database of 1,440 images. To allow the model and the images to fit in the memory, the training is performed using downscaled versions of the training images. We use images of dimension around $(2000, 1300)$ by resizing the gathered HDR and SDR photographs presented in Section 5.2.5. This dataset is available online¹.

For testing purposes and comparison with other methods, we use the HDR-Real test dataset proposed by Liu et al. [Liu+20]. It contains more than 8,000 pairs of SDR/HDR images and is widely used in the state-of-the-art as a test set thanks to its number of images. These images were obtained from 480 original HDR images using augmentation techniques, namely cropping (with a crop window of 512×512) and varying exposure times and CRFs to create SDR from HDR.

1. <ftp://ftp.irisa.fr/local/percept/public/hdrlnet/>

5.2.6 Choice of input iTMO

As we want to use already inverse tone-mapped images as input of our network, we now need to select the iTMOs that we can use.

Method

We start from a set of inverse tone mapping operators: the five operators implemented in the Matlab HDR toolbox [Ban+17] and the style-aware tone expansion [Bis+16]. Those iTMOs are not based on learning algorithms, but rather handcrafted using photographic rules and common assumptions. Setting up the input tensor so that each pair of images is processed by the 3D convolution is feasible only with three operators (as shown in Section 5.2.2). Among the six available iTMOs, we want to select three operators that are quite different, such that we have maximum performances with minimal network input size. The selection method is presented in this section.

First, we suppose that all of our operators are of similar output quality (with different strengths and weaknesses), and we use quality metrics to differentiate the operators on a chosen set of N images. We use our training dataset to do this, so we have $N = 360$. We process each inverse tone-mapped images of our set with six different metrics: HDR-VDP2, FSIM, MCS5, SI, PU-PSNR, PU-SSIM. These metrics are the ones used by Harmonic HDR-IQA [Rou+19], and PU-PSNR and PU-SSIM [AMS08]. Each of the N images is then represented by a 6D vector of metrics scores. Then, we project those vectors in a 2D space using the t-SNE visualization algorithm, to make the analysis easier. From this, we compute the coverage of the space of each set of 3 iTMOs among the 6. To do so, we compute the convex hull which encloses all points from 3 of the iTMOs, and we finally choose the 3 iTMOs which have a convex hull of maximum area. The t-SNE plot, along with the convex hull with largest area, is represented in Figure 5.5. The three chosen iTMOs are Akyuz [Aky+07], Landis [Lan02] and Kovaleski and Oliveira [KO14].

Discussion

All of these operators have different strengths and weaknesses. Akyuz operator is a simple algorithm that sets the maximum luminance value to a constant. While it usually burns the high luminance areas, low- and middle-exposed areas of the HDR output are quite faithful to the SDR version.

Landis operator uses a power function to improve the luminance values of pixels above

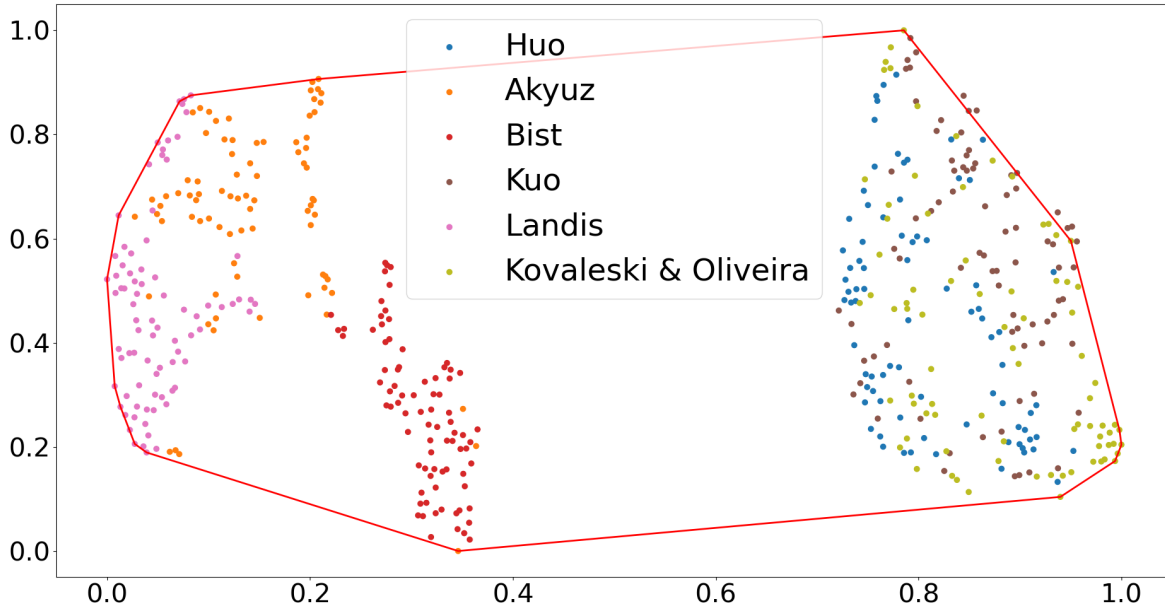


Figure 5.5 – 2D projection of the scores of images obtained with the t-SNE algorithm. Each colour corresponds to a different iTMO. The red line is the maximum area convex hull, which encompasses points from Akyuz, Landis and Kovaleski & Oliveira iTMOs.

a certain threshold. This yields high-exposed pixels with appropriate value, but may introduce artifacts: as the SDR image is quantized, some blocks or bands may appear on smooth gradient areas.

Finally, the operator from Kovaleski & Oliveira uses joint bilateral filters to smooth out the areas to expand. This method reduces the pixel values, but produces HDR images with less artifacts than Landis.

Note that the t-SNE algorithm which projects high dimension data points in smaller spaces is not very stable: small variations in the input data could modify the projection, and thus the chosen operators. We can however see on the Figure 5.5 that several iTMOs are very close to each other. As the areas of the bounding boxes are not very different from each others, this choice do not have a huge impact on the performances of our model.

5.3 Results

5.3.1 Implementation details

The network is written using the PyTorch framework and is available online². Using the dataset presented in Section 5.2.5, we train the network for 15 epochs, while reducing the learning rate each time the validation error increases. The number of epoch is quite low compared to other networks due to the small size of the network. Due to memory restrictions, we use a batchsize of 1.

Besides, each epoch runs for about 40 minutes, for a total training time of approximately 10 hours. This is a much faster training than state-of-the-art training, which ranges from a few days to a full week.

5.3.2 State-of-the-art comparison

In this section, we compare our method to other state-of-the-art ones using objective metrics. The iTMOs we consider are HDRCNN [Eil+17], ExpandNet [Mar+18], the Single Image Network [Liu+20], HDRUNet [Che+21] LANet [Yu+21], DrTMO [EKM17] and the model from Santos et al. [SRK20] (called HDR-masked in the rest of this article).

HDRCNN, DrTMO, LANet and the Single Image Network are presented in Section 3.2.2. We remind that HDRCNN contains about 30×10^6 trainable parameters while HDR-LFNet contains about 2×10^5 trainable parameters. Besides, the Single Image Network improves on HDRCNN by using inpainting-like tasks during the training. It contains about 2×10^6 parameters. HDRUNet contains 1.6×10^6 parameters. The main idea behind HDRUNet is to split the network into three modules: a base network that performs most of the work, a condition network that computes spatially-variant transformations used to modify the deep features of the base network, and a weighting network that detects over-exposed areas to improve the reconstruction in those areas. All these networks are trained using their novel $\tanh_{\mathcal{L}_1}$ loss function. ExpandNet is a much lighter network with around 5×10^5 parameters. It also proposes a network with different modules, with each module working on a different scale: a global branch, a local branch and a dilation branch for mid-scale features. Each of those branches adds new information, which allows for a more faithful HDR reconstruction. HDR-masked [SRK20] is a CNN-based iTMO. Inspired by HDRCNN, the model is composed of a neural network that estimates the over-exposed

2. <ftp://ftp.irisa.fr/local/percept/public/hdrlnet/>

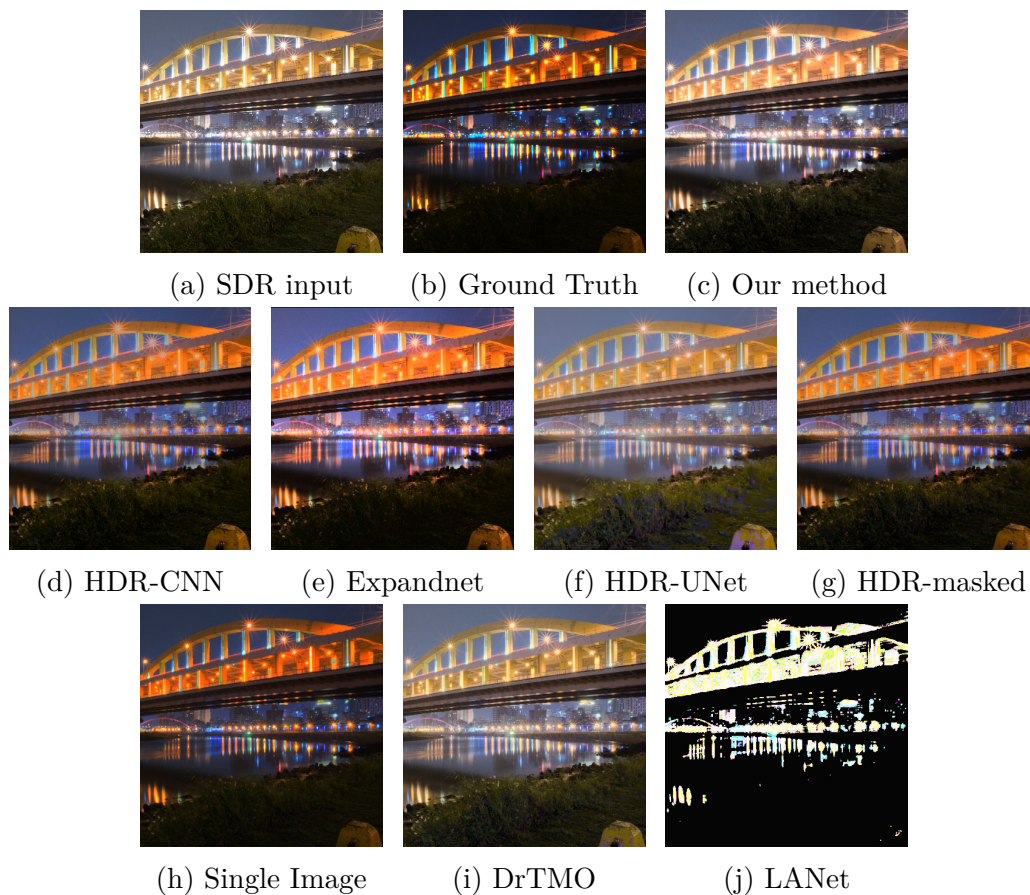


Figure 5.6 – Example of an image from HDR-Real dataset processed by different models. For ease of view, HDR images have been tone-mapped using the Drago algorithm [Dra+03], which impairs the colours. To view HDR images in full quality, an HDR display with HDR images is necessary.

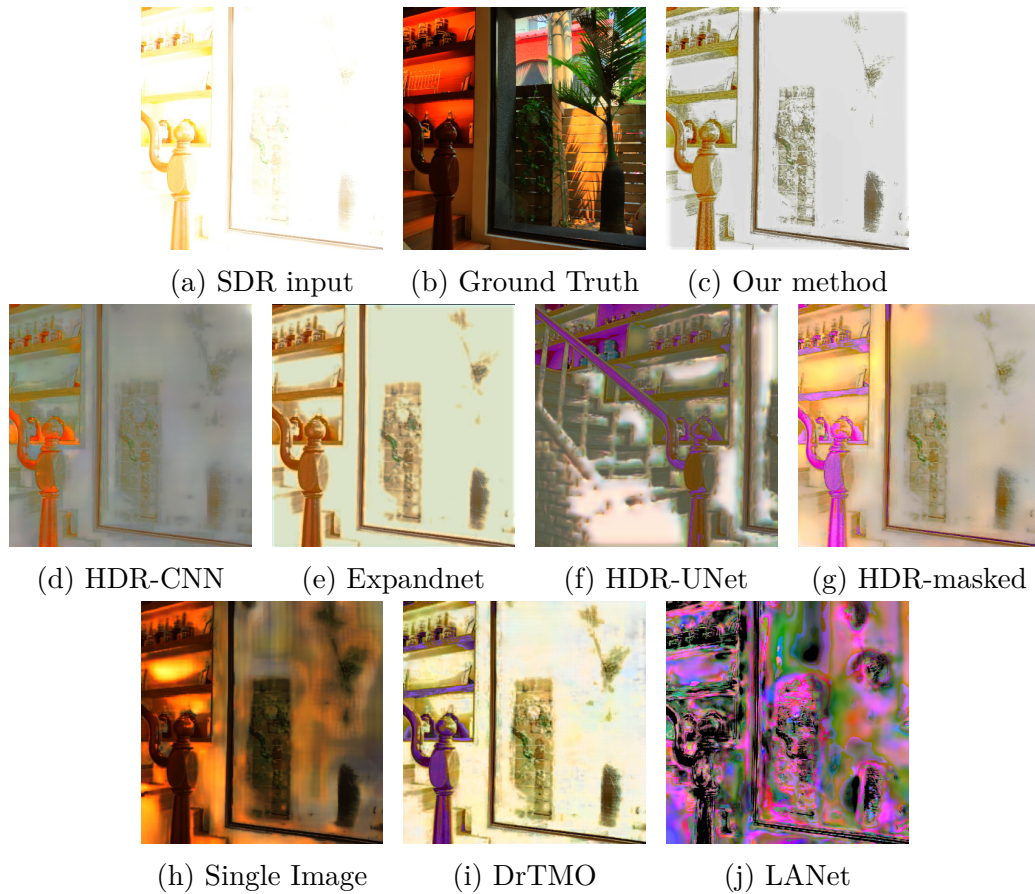


Figure 5.7 – Example of an over-exposed image from HDR-Real dataset processed by different models. Due to burned areas in the image, the colors have not been processed correctly. For ease of view, HDR images have been tone-mapped using the Drago algorithm [Dra+03].

areas in the SDR. As over-exposed areas are completely white in the SDR, the information is completely lost. Therefore, to improve the HDR reconstruction, the authors pre-trained their network for an inpainting task, before fine-tuning it for HDR reconstruction. Besides, to focus the reconstruction on over-exposed areas in the network processing, a mask is applied on the features to focus the modifications on the over-exposed areas. Their network contains more than 50×10^6 parameters.

All of the networks are fully convolutional neural networks, meaning that the input can theoretically be of any size. However, for HDRCNN, due to how the deconvolutions are used, the input must have its height and width multiple of 32. This means that some of the images are either cropped or resized to fit this requirement.

We show some examples of images obtained with our method and with existing models on Figures 5.6 and 5.7. We notice that LANet generates images with many artifacts. In our experiment, LANet uses the weights provided by the authors. Therefore, this model, trained with the dataset of the authors, does not perform well on the images from the HDR-Real dataset, but should benefit from a re-training using the HDR-Real training set. Figure 5.7 presents an over-exposed image. The reconstruction of burned areas in images is one of the main difficulties of iTMOs. As such, few methods manage to generate a convincing HDR image. This is one of the limits of most of the methods, including our proposed method.

The metric used in the state of the art for comparing models is HDR-VDP2 [Man+11]. It is a metric with reference, and it takes as supplementary argument the angular resolution (measured in pixel per degree (ppd)). The angular resolution depends on the distance between the observer and the screen, and the resolution of the screen. This metric allows for a faithful simulation of the viewing experience on a specific screen. We represent on Figure 5.8 the impact of the angular resolution on the score. We notice that, although the actual scores change, the ranking of the different models do not. For the rest of the evaluation, we use an angular resolution of 30 ppd, and we set the maximum luminance value of images to 1000.

As HDR-VDP2 only works on luminances, we use Harmonic HDR-IQA [Rou+19], which is sensitive to colours, as well as PU-PSNR and PU-SSIM [AMS08].

For this experiment, we test our network using the HDR-Real test dataset proposed by Liu et al. [Liu+20]. We train three versions of our network: (i) using the HDR-Real train, (ii) using our training dataset, and (iii) using our training dataset fine-tuned on HDR-Real train dataset. For fairness of comparison, we train these three versions for

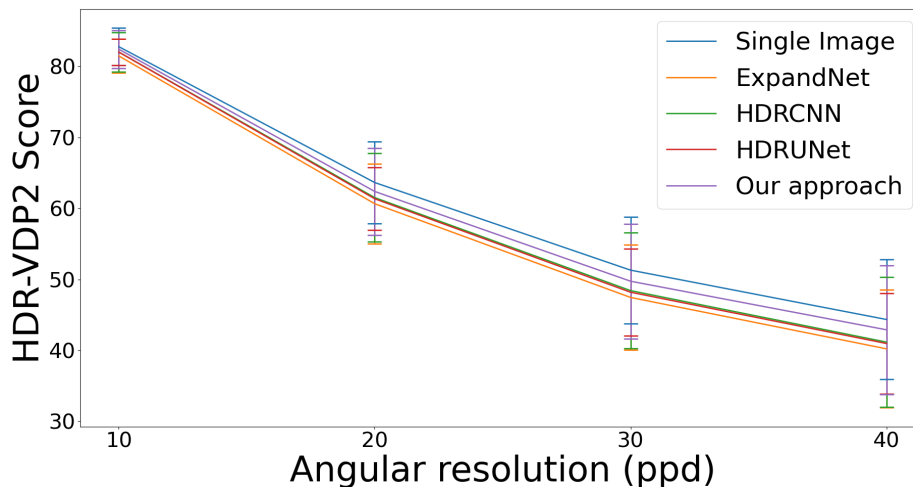


Figure 5.8 – HDR-VDP2 score variation depending on the angular resolution on the HDR-Real dataset for four models.

the same amount of time (10 hours). HDRCNN, HDRUNet, the Single Image Network and DrTMO are trained with the HDR-Real train set, while the other networks use the weights provided by their respective authors. The different images were provided by Liu et al. [Liu+20]. We present the results of our evaluation in Table 5.2. We find out that, although our method do not perform the best, we manage to get second best on most of the metrics, except on Harmonic IQA. As Harmonic IQA assess the differences in colour between HDR images, this shows that our method does not reproduce colours as well as the other methods. However, the user study presented in Section 5.3.4 reveals that our method is preferred by observers. These results mean that although we are less faithful to the colours of the original image, our method produces a more appealing picture than the state-of-the-art methods. Furthermore, our method is faster than all tested methods.

Note that our method performs best on HDR-Real test set when it is trained on our dataset, and not on HDR-Real train set. This comes from the nature of the datasets and our architecture: we managed to drastically reduce the number of trainable parameters. As explained in Section 5.2.5, the low number of parameters calls for fewer train images, but with richer content. HDR-Real images being only 512×512 , our dataset is better suited to train our network than HDR-Real. We present on Figure 5.9 the mean HDR-VDP2 score as a function of the number of trainable parameters for our network, and networks of the state of the art. From this point of view, our model actually performs better than the state of the art. The state-of-the-art models would have better performances if they were trained with our training dataset. As the training of the state-of-the-art methods

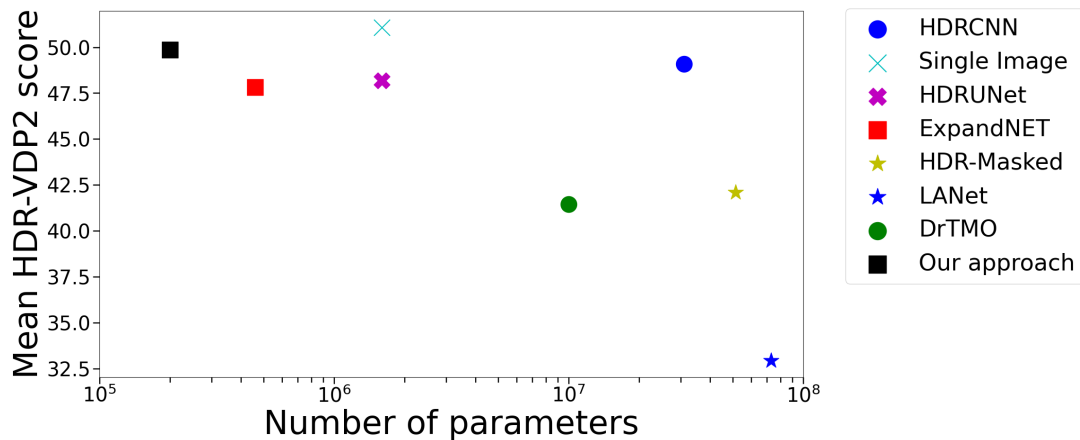


Figure 5.9 – Mean HDR-VDP2 score against the number of trainable parameters of the tested models.

takes a long time, the full study of the different architectures trained using all the different training dataset is left as future work.

Model	Time (min)	Mean HDR-VDP2 (std) †	Mean Harm. IQA (std) †	PU-PSNR (std) †	PU-SSIM (std) †
HDRCNN [Eil+17]	135	49.0745 (7.6484)	0.3554 (0.0997)	20.1330 (8.7622)	0.4325 (0.4499)
ExpandNet [Mar+18]	110	47.8268 (6.8777)	0.3685 (0.1003)	19.9602 (8.1942)	0.4043 (0.4643)
Single Image [Liu+20]	1700	*51.0739 (6.9557)	0.3798 (0.1136)	*26.4531 (8.8608)	*0.5861 (0.4353)
HDRUNet [Che+21]	850	48.1709 (6.1163)	*0.4174 (0.0480)	18.0796 (7.3235)	0.3270 (0.4764)
HDR-Masked [SRK20] †	180	42.0945 (8.7544)	0.3536 (0.0955)	20.3317 (8.6756)	0.4389 (0.4478)
LANet [Yu+21] †	90	32.9406 (7.0315)	0.3540 (0.0657)	17.8393 (7.1079)	0.3552 (0.4066)
DrTMO [EKM17]	-	41.4390 (7.9000)	0.3342 (0.0745)	18.3249 (8.1199)	0.2961 (0.4751)
Our Method (Real-1)	35	28.3378 (5.8721)	0.2764 (0.1386)	12.4031 (6.4205)	0.0771 (0.3641)
Our Method (Real-2)	35	41.3083 (8.7949)	0.3555 (0.1002)	19.7721 (8.9024)	0.3749 (0.4699)
HDR-LFNet †	35	<u>49.8686</u> (8.4689)	0.3597 (0.1053)	<u>21.0415</u> (8.8968)	<u>0.4647</u> (0.4407)

Table 5.2 – Mean and standard deviation of HDR-VDP2 and Harmonic IQA for several models on the HDR-Real [Liu+20] test set. The line "Our Method (Real-1)" corresponds to our network trained on HDR-Real; The line "Our Method (Real-2)" corresponds to our network trained on our proposed dataset, fine-tuned on HDR-Real. The symbol † denotes models that were not trained with the HDR-Real training dataset. Scores with **a star (*)** are the best; Scores underlined are the second best. The "Time" column corresponds to the approximate duration of the full evaluation of the 1837 images from the HDR-Real set on CPU.

Model	Score	Input iTMO	Score
With 2DConv	42.6345	Landis	32.4351
Without kernel size change	38.9088	Kovaleski & Oliveira (KO)	34.5834
Only MAE	36.3503	Akyuz	26.2426
MAE + gMAE	34.9649	KO + Landis	26.4762
MAE + dynLoss	39.9745	KO + Akyuz	32.1067
MAE + VGGLoss	35.4325	Landis + Akyuz	28.4841
Final model	49.8686		

Table 5.3 – Mean HDR-VDP2 score for different variants of the model. On the left, variations based on the architecture or the loss function. On the right, variations based on the input of the fusion network.

5.3.3 Ablation study

In this section, we present the study conducted to assess the quality of the different components: the composed loss function and the usage of 3D convolutions, skip connections, and convolution kernel with varying sizes.

Each subsequent section presents a modified variant of our network. It is trained on our dataset and tested on HDR-Real. We did not train our network using the HDR-Real training set as, to the best of our knowledge, the full resolution images are not available. The average HDR-VDP2 score is presented on Table 5.3 for all variants, as well as for our proposed method (called final model in this section).

Composed loss function

Our loss function is composed of four different components. We train the same architecture with loss functions composed of only some of the original components detailed in Section 5.2.3. The mean absolute error (MAE) keeps the overall structure better among the four components, so we train networks with MAE and gradient loss (gMAE); MAE and dynamic range loss; and MAE and perceptual VGG loss. We also train a network with only MAE to compute the added value of each component. We use the same weighting as given in Equation (5.2) when training the different networks: for example, the version $MAE + gMAE$ was trained using the loss function $Y = Y_c + 0.3Y_g$.

Surprisingly, our method performs better when trained with only MAE rather than MAE + gMAE or MAE + Perceptual loss. This can be explained as follows. In the final model, each weighting coefficient assigned to each component of the loss function has been carefully tuned with regards to the others. When using only two components of the loss

function, the weighting coefficients should be different.

Therefore, we can assume that the scores given in Table 5.3 are not optimal, except when training only with MAE (as no tuning is necessary because there is only one component). However, due to the large difference in score between the final model and the different loss function versions, we assume that our full loss function improves the output quality compared to the other tested loss functions. Besides, we notice that the dynamic loss is the most important component of the loss, as it effectively improves the quality according to the HDR-VDP2 scores. This is reflected in the relative weights of the loss: the weighting coefficient of the dynamic range loss component is much higher than the other weighting coefficients.

2-by-2 processing

As explained in section 5.2.2, we use 3D convolution layers in our network. To assess the usefulness of the 3D convolution layers, we train the same architecture with 2D convolution layers only. The related HDR-VDP2 scores in Table 5.3 show that 3D convolutions contribute positively to the quality of the result. Visually, the output of the 2D Convolution network is similar to the output of our network, but we notice some discrepancies, especially in lowly lit areas. This may come from the fact that highly lit areas are represented by high values in the tensor, and those values make low light levels not significant during the backpropagation. On the other hand, by using 3D convolutions, low light levels – represented by low tensor values – are not always processed together with high tensor values. If at least two input images present low light values, there exists one channel that processes only those two input images and the low light values are correctly passed on to the next layer. Therefore, low light values are more accurately described in the different layers, and thus also in the final output image.

However, we have to note that the total number of trainable parameters was changed by this simple modification of the network. Therefore, the observed performance drop may be partly caused by the lower complexity of the network with 2D convolutions.

Varying kernel size

To further improve the reconstruction of details, we use, in the second half of the network, convolutions with decreasing size from $(5, 5, ?)$ to $(1, 1, ?)$. We train the same architecture with fixed-sized convolution kernel of $(3, 3, ?)$. We notice some blur on those images, that comes from the convolution. Indeed, $(3, 3, ?)$ averages the values of the pixels

in the neighbourhood, which leads to faded edges on the image, and to a worse HDR-VDP2 score.

Changing the network input

We assume that the fusion of three different versions of the image can only improve the quality of the reconstruction. We verify this assumption by training networks with different inputs from our proposed method: either one reconstructed HDR image generated from one iTMO, or two reconstructed images generated from two different iTMOs. We notice some strong artifacts, which are the same as the ones produced by the chosen input iTMO. In the proposed version, with three inputs, the artifacts are present only on one of the versions, and therefore these artifacts are attenuated through the convolutions.

We notice that the scores for the different variants of the network with different inputs are quite low. This may come from the loss function. The relative weights of each component of the loss function were carefully tuned for our proposed dataset, and we used those weights to train the different versions of our network. As the input images are not the same, the relative weights should be different too.

5.3.4 Subjective user study

We present in this section the user study we have conducted using our HDR SIM2 screen, to effectively compare the performance of our proposed method with state-of-the-art ones.

We want to compare the quality of the HDR images produced by different methods, including ours. We decided to perform a Two-Alternative Forced Choice (2AFC) experiment setting. To do so, we need to define the images pairs to be displayed on the HDR screen. For each participant, we randomly chose 10 images among our 80 test images and 5 versions of each of them: our method, HDRCNN, HDRUNet, ExpandNet and Single Image Network. The participants were presented with all possible image pairs created from the 10 test images and the 5 methods, for a total of 100 image pairs observed per session. In the following, we denote by $\langle I_M, I_{M'} \rangle$ or $\langle I_{M'}, I_M \rangle$ an image pair (with I_X half the image generated by the method X): one half is generated by the method M from the SDR image I_{SDR} and the other half by the method M' from the same SDR image I_{SDR} (see Figure 5.10). We then asked the participants to choose their preferred method among the two methods shown in the image pair displayed on the HDR screen. To respect the



Figure 5.10 – Example of an image pair shown to one participant of the user study. Here, the left half was generated with HDRUNet and the right half with HDRCNN (all images have been tone-mapped using the Drago algorithm).

aspect ratio of the images, we randomly chose if we display the left sides or the right sides of each image in the pair. The displayed image pair is then composed of twice the same side of one image, as shown in Figure 5.10.

As it is difficult to train state-of-the-art networks with our dataset, we decided to use the pre-trained networks provided by the authors of the state-of-the-art methods. Therefore, our study compares not only the architecture of these networks, but also their respective training datasets.

Each participant can attend multiple sessions (a session corresponds to the evaluation of 100 images pairs). If a participant attends another session, we use 10 images that the participant has not seen yet. The user study involved 29 participants (23M, 6F; age: $\text{avg}=36.3 \pm 12.7$, $\text{min} = 23$, $\text{max} = 71$), and 30 sessions, for a total of 3,000 image pairs observed. Among the 29 participants, all of them reported to have normal or corrected-to-normal vision and 10 of them reported to have experienced HDR imaging before. During the experiment, each participant was asked to choose the method they preferred in the image pair. We have collected for each participant 100 answers (left or right of the image pair), each of them corresponds to a tested iTMO. For each participant p , we compute

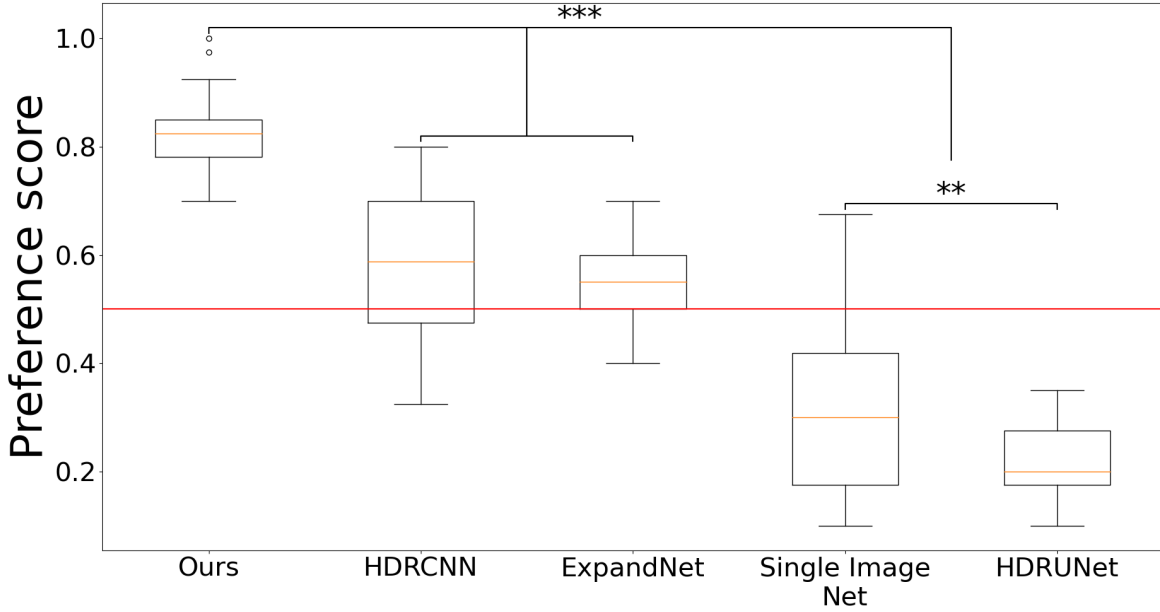


Figure 5.11 – Preference of all participants for each method. The red line corresponds to a preference of 50%. The stars are attributed according to the p-values of the Tukey HSD test: nothing for $p > 0.05$, * for $0.05 \geq p > 0.005$, ** for $0.005 \geq p > 0.0005$, *** for $0.0005 \geq p$.

the preference scores $x_p(M_1), \dots, x_p(M_5)$ for each method M_i . If we denote by $IP(M)$ the set of image pairs that contains an image generated with the method M , we can compute $x_p(M)$ with

$$x_p(M) = \frac{\text{card}(\{\langle I_M, I_{M'} \rangle \in IP(M) \text{ s.t. } p \text{ preferred } I_M \text{ over } I_{M'}\})}{\text{card}(IP(M))}.$$

The method M performs well according to the participant p if $x_p(M) \geq 0.5$. These preference scores are represented in Figure 5.11.

We can see on the Figure 5.11 that our method is largely preferred on average to all other tested methods. To further study these preferences, we performed a one-way ANOVA test after asserting that our data (the computed $x_p(M)$) come from a normal distribution (using a Shapiro-Wilk test). We obtain a p-value $p \ll 0.05$, meaning that the average values are significantly different. To further discriminate the methods, we perform a post-hoc test using a Tukey HSD test. This statistical test allows to compare the mean of every group (in our case, the groups are the different iTMOs) two-by-two. The results of the Tukey HSD test provides, for each pair of methods, the probability p

that the mean preference scores of the two considered methods are the same. We present the results of the Tukey HSD test in Figure 5.11 by grouping together the methods with close probabilities p .

Using the post hoc test, we notice that the participants found on average no significant differences between the images processed by HDRCNN and by ExpandNet (Average value of preference score of $\bar{x} = 0.57$). Using the same test, our method is preferred on average to all other methods (Average value of preference score of $\bar{x} = 0.83$). Our method, HDRCNN and ExpandNet all have an average preference score of above 0.5, meaning that those three methods are most of the time preferred by the participants. This study shows that our method performs well for human observers on our test dataset.

Surprisingly, we notice that HDRUNet performed the best regarding the Harmonic HDR IQA metric, and the worst in the subjective study. This may come from the artifacts on the images generated by HDRUNet. While HDRUNet provides photographs with general colours more faithful to the original ground truth HDR image, our method has fewer artifacts and more saturated colours. This explains why the proposed method is preferred in the subjective study.

5.4 Conclusion

We have presented a new inverse tone mapping operator, called HDR-LFNet, along with its training dataset. Our architecture is lighter than the networks of the state-of-the-art thanks to methods aggregation, a technique inspired by other computer vision domains. To be fully effective, this lightweight architecture requires high resolution images to train. As there is no existing training HDR dataset with sufficient resolution, we also release a new dataset of high resolution HDR images that can be used by the community in complement or in place of existing datasets.

Objective metrics showed that our method is on-par with other methods, but the conducted user study showed that our method is preferred by observers. Along with the lower number of parameters, our method is a real improvement over the state-of-the-art. Our HDR-LFNet can be used in several applications, where resources for training and storage of the model are limited. This also should allow to transform an SDR image dataset with annotations (quality score, aesthetics score, saliency data) to an HDR image dataset with corrected annotations.

CONCLUSION

In this thesis, we have tackled two main problems concerning image quality improvement: high dynamic range (HDR) image generation and aesthetics assessment.

In Chapter 4, we have presented the notion of aesthetics in photography. Aesthetics is essentially a subjective concept that may differ from one person to another, but thanks to statistical learning it is possible to aggregate many different opinions. We can then compute an average aesthetic value for photographs using this aggregation of opinions. However, as the statistical learning methods used in the state-of-the-art are based on supervised learning algorithms, it is mandatory to use an annotated image dataset. Our contribution is to highlight the biases that come from AVA, one of the largest and most used dataset for aesthetics prediction. We have shown that recent models do not behave correctly on professional photographs – which are not present in the training dataset. However, by fine-tuning the models it is possible to change the scores given to professional photographs without altering the scores from AVA. This proves that professional and competitive photographs have aesthetic features different enough such that a training on competitive photographs is not sufficient to effectively predict aesthetics for professional photographs. However, it is possible to train a network with both kinds of photographs, meaning that a general purpose aesthetics assessment model should benefit from training with as many kinds of photographs as possible.

In Chapter 5, we have devised a new inverse tone mapping operator (iTMO) that merges several outputs from existing iTMOs. The neural network that we propose takes as input three different HDR images to yield a single HDR image, as opposed to state-of-the-art methods that devise neural networks aiming at transforming SDR to HDR. Our transformation is easier, and therefore requires fewer parameters to learn. We have shown that our method achieves similar performances than the state-of-the-art using less resources. We evaluated our work using objective metrics such as HDR-VDP to compare the output of our network to the ground-truth HDR image built using the exposure fusion method. Besides, we have performed a user study that have shown that our method is

preferred by human observers on an HDR screen.

Our work lays the groundwork for future projects. As we have shown that the training dataset is very important for supervised learning methods, especially for the methods trained for aesthetics assessment, two main options are viable. The first one is to propose a new annotated dataset for image aesthetics assessment. This new dataset should be as large as AVA (same number of images, same number of votes, same number of voters), with a higher diversity of images. A second option would be to remove the dependency on the training dataset. This would have two effects. First, an annotated dataset would not be needed anymore. This would make the design of new aesthetics assessment models easier. Second, models could be trained with other kinds of aesthetic qualities as a goal than what is used today – namely “liked by the majority of people”. New aesthetics with no relation with the community of voters of the training dataset could be considered. Training with no annotated dataset can be achieved thanks to unsupervised learning algorithms. A ranking of aesthetic quality would be harder with an unsupervised model, but such a model should be able to discriminate images by their aesthetic features, and so this model would be a good recommender system. Half-way between supervised and unsupervised methods lie semi-supervised methods. In the case of aesthetics assessment models, there are several interesting semi-supervised techniques to explore. We can either use partially annotated data (we choose a subset of our training images to be rated by observers, while some of the training images have no ground truth: when we consider only one observer, this is the basis of personalized aesthetics assessment [Ren+17b]), or fuzzy annotations (we allow the observers and the trained network to give a range of aesthetic scores, instead of a unique ground truth). Finally, it is useful to process images containing as much information as possible to accurately assess their aesthetic quality. That was one of the reasons leading to the design of HDR-LFNet, but other rich image formats exist. We can mention wide color gamut (WCG) images and panoramic and 360° images. WCG images are images with a better color representation than standard images. They are similar to HDR images, as such, the same processing techniques may be used. To assess the aesthetic quality of WCG or HDR images, techniques used in HDR-LFNet (Convolutions for processing, reduction of dimension thanks to an operation before the neural network) may be tested. Panoramic images are images with a larger field of view than standard images, while 360° images are a particular kind of panoramic images (with a field of view of 360°). Unlike HDR images, the added information in panoramic images is in the spatial domain. As such, it may be preferable to use another course of action,

such as processing different levels of details using different sizes of convolution kernels. The ultimate goal of tackling these different kinds of images would be to design a method to assess the aesthetic quality of an image consisting in a combination of these different characteristics (HDR, WCG, panoramic). This would lead to a way to assess the quality of real world scenes, and could be used to help photographers choosing their preferred settings (camera settings and viewpoint) when taking a photograph.

Furthermore, one of our motivations to devise our inverse tone mapping operator was to be able to generate a large number of HDR images. This stems from the observation that deep learning methods were successful for many different tasks in SDR image processing. As there are more and more HDR images, it is important to be able to perform the same tasks on HDR images, and the first option is to try using the same methods. Therefore, it would be interesting to implement some tasks for HDR images (such as compression, denoising, inpainting, etc...) One method is to use generic features, but generic feature extractors do not exist for HDR imaging. Now that iTMOs allow to have as many HDR training images as possible, the next step would be to devise an architecture to perform generic feature extraction for HDR images. Finally, we can wonder about the processing of content richer than images. The most interesting visual content to tackle would be video. The main problematic of video content is the time component: the processing of a video as a collection of frames do not usually produce good results. For example, for video inverse tone mapping, the naive iTMO that consists in applying an iTMO to every frame of the video introduces temporal artifacts, with some frames brighter than others. It is then mandatory to take the time aspect into account. In our situation, we can discuss the use of 4D convolution operators in a neural network. Indeed, the extra dimension would help process the last few frames of the video so that the operation do not only depend on the current frame. Thanks to the 4D convolutions and the pre- and post-processing performed in our HDR-LFNet, it may be possible to propose a lightweight video inverse tone mapping operation.

BIBLIOGRAPHY

- [Aky+07] Ahmet Oguz Akyuz, Roland Fleming, Bernhard E Riecke, Erik Reinhard, and Heinrich H Bulthoff, « Do HDR displays support LDR content? A psychophysical evaluation », *in: ACM Transactions on Graphics (TOG) 26.3* (2007), 38–es.
- [AM19] Konstantinos A. and Vasileios M., « Image Aesthetics Assessment using Fully Convolutional Neural Networks », *in: International Conference on Multimedia Modeling*, Springer, 2019, pp. 361–373.
- [AMS08] TunÇ Ozan Aydin, Rafal Mantiuk, and Hans-Peter Seidel, « Extending Quality Metrics to Full Dynamic Range Images », *in: Human Vision and Electronic Imaging XIII*, Proceedings of SPIE, San Jose, USA, Jan. 2008, pp. 6806–10.
- [Ban+17] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers, *Advanced High Dynamic Range Imaging (2nd Edition)*, Natick, MA, USA: AK Peters (CRC Press), July 2017, ISBN: 9781498706940.
- [Bir33] George David Birkhoff, « Aesthetic measure », *in: Aesthetic Measure*, Harvard University Press, 1933.
- [Bis+16] Cambodge Bist, Rémi Cozot, Gérard Madec, and Xavier Ducloux, « Style Aware Tone Expansion for HDR Displays. », *in: Graphics Interface*, 2016, pp. 57–63.
- [BM21] Jakub Boksansky and Adam Marrs, « The Reference Path Tracer », *in: Ray Tracing Gems II: Next Generation Real-Time Rendering with DXR, Vulkan, and OptiX* (2021), pp. 161–187.
- [BYB22] Joon-ki Bae, Subin Yang, and Sung-Ho Bae, « DenSE SwinHDR: SDRTV to HDRTV Conversion using Densely Connected Swin Transformer with Squeeze and Excitation Module », *in: IEEE Access* (2022).

-
- [Car+19a] A. Carballal, C. Fernandez-Lozano, J. Heras, and J. Romero, « Transfer learning features for predicting aesthetics through a novel hybrid machine learning method », *in: Neural Computing and Applications* (Feb. 2019), ISSN: 1433-3058, DOI: 10.1007/s00521-019-04065-4, URL: <https://doi.org/10.1007/s00521-019-04065-4>.
- [Car+19b] A. Carballal, C. Fernandez-Lozano, N. Rodriguez-Fernandez, L. Castro, and A. Santos, « Avoiding the Inherent Limitations in Datasets Used for Measuring Aesthetics When Using a Machine Learning Approach », *in: Complexity* 2019 (Jan. 2019), p. 12, DOI: 10.1155/2019/4659809.
- [CCL22] Mathieu Chambe, Rémi Cozot, and Olivier Le Meur, « Deep learning for assessing the aesthetics of professional photographs », *in: Computer Animation and Virtual Worlds* 33.6 (2022), e2105.
- [Cha+23] Mathieu Chambe, Ewa Kijak, Zoltan Miklos, Rémi Cozot, Olivier Le Meur, and Kadi Bouatouch, « HDR-LFNet: Inverse Tone Mapping using Fusion Network », *in: Computer & Graphics* (2023).
- [Che+17] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma, « Learning to compose with professional photographs on the web », *in: Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 37–45.
- [Che+21] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong, « HDRUnet: Single image hdr reconstruction with denoising and dequantization », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 354–363.
- [CLC17] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen, « Aesthetic critiques generation for photos », *in: Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3514–3523.
- [CMY19] C. Chen, S. McCloskey, and J. Yu, « Analyzing Modern Camera Response Functions », *in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2019, pp. 1961–1969, DOI: 10.1109/WACV.2019.00213.

-
- [Cui+18] C. Cui, H. Liu, T. Lian, L. Nie, L. Zhu, and Y. Yin, « Distribution-oriented Aesthetics Assessment with Semantic-Aware Hybrid Network », *in: IEEE Transactions on Multimedia* early access (2018), early access, DOI: 10.1109/TMM.2018.2875357.
- [Dan+15] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato, « Raise: A raw images dataset for digital image forensics », *in: Proceedings of the 6th ACM multimedia systems conference*, 2015, pp. 219–224.
- [Dat+06] R. Datta, D. Joshi, J. Li, and J. Z. Wang, « Studying aesthetics in photographic images using a computational approach », *in: European Conference on Computer Vision*, Springer, 2006, pp. 288–301.
- [DLX17] Y. Deng, C. C. Loy, and X. Tang, « Image Aesthetic Assessment: an Experimental Survey », *in: IEEE Signal Processing Magazine* 34.4 (2017), pp. 80–106, DOI: 10.1109/MSP.2017.2696576.
- [DM08] Paul E Debevec and Jitendra Malik, « Recovering high dynamic range radiance maps from photographs », *in: ACM SIGGRAPH 2008 classes*, 2008, pp. 1–10.
- [DOB11] S. Dhar, V. Ordonez, and T. L. Berg, « High level describable attributes for predicting aesthetics and interestingness », *in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 2011, pp. 1657–1664.
- [Dra+03] Frédéric Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba, « Adaptive logarithmic mapping for displaying high contrast scenes », *in: Computer graphics forum*, vol. 22, 3, Wiley Online Library, 2003, pp. 419–426.
- [Eil+17] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, and Jonas Mantiuk Rafałand Unger, « HDR image reconstruction from a single exposure using deep CNNs », *in: ACM Transactions on Graphics (TOG)* 36.6 (2017).
- [EKM17] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani, « Deep reverse tone mapping. », *in: ACM Trans. Graph.* 36.6 (2017), pp. 177–1.
- [Fai07] Mark D Fairchild, « The HDR photographic survey », *in: Color and imaging conference*, vol. 2007, 1, Society for Imaging Science and Technology, 2007, pp. 233–238.

-
- [FB09] Daniel Filonik and Dominikus Baur, « Measuring Aesthetics for Information Visualization », *in: 2009 13th International Conference Information Visualisation*, 2009, pp. 579–584, DOI: 10.1109/IV.2009.94.
- [FF21] Corneliu Florea and Laura Florea, « Multitask Regularization for Image Aesthetic Evaluation », *in: 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, IEEE, 2021, pp. 1–4.
- [Han+20] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi, « Neuromorphic camera guided high dynamic range imaging », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1730–1739.
- [He+16] K. He, X. Zhang, S. Ren, and J. Sun, « Deep residual learning for image recognition », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [How+17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, « Mobilenets: Efficient convolutional neural networks for mobile vision applications », *in: arXiv preprint arXiv:1704.04861* (2017).
- [Jos+11a] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J. Z. Wang, J. Li, and J. Luo, « Aesthetics and Emotions in Images », *in: IEEE Signal Processing Magazine* 28.5 (Sept. 2011), pp. 94–115, ISSN: 1053-5888, DOI: 10.1109/MSP.2011.941851.
- [Jos+11b] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo, « Aesthetics and emotions in images », *in: IEEE Signal Processing Magazine* 28.5 (2011), pp. 94–115.
- [JSS16] B. Jin, M. V. O. Segovia, and S. Süsstrunk, « Image aesthetic predictors based on weighted CNNs », *in: 2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 2291–2295.
- [Ker14] Charles Kervrann, « PEWA: Patch-based exponentially weighted aggregation for image denoising », *in: Advances in Neural Information Processing Systems* 27 (2014), pp. 2150–2158.

-
- [Klí+11] Miloš Klíma, Karel Fliegel, Petr Páta, Stanislav Vítek, Martin Blažek, Petr Dostal, Lukáš Krasula, Tomáš Kratochvíl, Václav Ríčný, Martin Slanina, et al., « DEIMOS—An Open Source Image Database. », *in: Radioengineering* 20.4 (2011).
- [KLM18] M. Kucer, A. C. Loui, and D. W. Messinger, « Leveraging Expert Feature Knowledge for Predicting Image Aesthetics », *in: IEEE Transactions on Image Processing* 27 (Oct. 2018), pp. 5100–5112, DOI: 10.1109/TIP.2018.2845100.
- [KO14] Rafael P Kovalski and Manuel M Oliveira, « High-quality reverse tone mapping for a wide range of exposures », *in: 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, IEEE, 2014, pp. 49–56.
- [Kon+16a] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, « Photo aesthetics ranking network with attributes and content adaptation », *in: European Conference on Computer Vision*, Springer, 2016, pp. 662–679.
- [Kon+16b] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes, « Photo aesthetics ranking network with attributes and content adaptation », *in: European Conference on Computer Vision*, Springer, 2016, pp. 662–679.
- [KVD20] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux, « Eva: An explainable visual aesthetics dataset », *in: Joint workshop on aesthetic and technical quality assessment of multimedia and media analytics for societal trends*, 2020, pp. 5–13.
- [LAK18] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang, « Deep recursive hdri: Inverse tone mapping using generative adversarial networks », *in: proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 596–611.
- [Lan02] Hayden Landis, « Production-ready global illumination », *in: Siggraph course notes 16.2002* (2002), p. 11.
- [Liu+20] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang, « Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline », *in: Proceedings of*

-
- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1651–1660.
- [Low99] David G Lowe, « Object recognition from local scale-invariant features », *in: Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, Ieee, 1999, pp. 1150–1157.
- [Lu+14] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, « Rapid: Rating pictorial aesthetics using deep learning », *in: Proceedings of the 22nd ACM international conference on Multimedia*, ACM, 2014, pp. 457–466.
- [Lu+15] X. Lu, Z. Lin, X. Shen, R. Shen, and J. Z. Wang, « Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation », *in: 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 990–998, DOI: 10.1109/ICCV.2015.119.
- [Lu+19] P. Lu, X. Peng, J. Yu, and X. Peng, « Gated CNN for visual quality assessment based on color perception », *in: Signal Processing: Image Communication* 72 (2019), pp. 105–112, DOI: 10.1016/j.image.2018.12.007.
- [LWT11] Wei Luo, Xiaogang Wang, and Xiaoou Tang, « Content-based photo quality assessment », *in: 2011 international conference on computer vision*, IEEE, 2011, pp. 2206–2213.
- [Man+07a] Rafal Mantiuk, Grzegorz Krawczyk, Radoslaw Mantiuk, and Hans-Peter Seidel, « High-dynamic range imaging pipeline: perception-motivated representation of visual content », *in: Human Vision and Electronic Imaging XII*, vol. 6492, International Society for Optics and Photonics, 2007, p. 649212.
- [Man+07b] Rafał Mantiuk, Grzegorz Krawczyk, Radosław Mantiuk, and Hans-Peter Seidel, « High Dynamic Range Imaging Pipeline: Perception-motivated Representation of Visual Content », *in: Human Vision and Electronic Imaging XII*, ed. by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, vol. 6492, Proceedings of SPIE 649212, San Jose, USA: SPIE, Feb. 2007.
- [Man+09] Radoslaw Mantiuk, Rafal Mantiuk, Anna Tomaszewska, and Wolfgang Heidrich, « Color correction for tone mapping », *in: Computer Graphics Forum*, vol. 28, 2, Wiley Online Library, 2009, pp. 193–202.

-
- [Man+11] Rafal Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich, « HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions », *in: ACM Transactions on graphics (TOG)* 30.4 (2011), pp. 1–14.
- [Mar+18] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista, « Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content », *in: Computer Graphics Forum*, vol. 37, 2, Wiley Online Library, 2018, pp. 37–49.
- [MBD21] Demetris Marnerides, Thomas Bashford-Rogers, and Kurt Debattista, « Deep HDR Hallucination for Inverse Tone Mapping », *in: Sensors* 21.12 (2021), p. 4032.
- [MEP09] Debarshi Mustafi, Andreas H Engel, and Krzysztof Palczewski, « Structure of cone photoreceptors », *in: Progress in retinal and eye research* 28.4 (2009), pp. 289–302.
- [Met+20] Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein, « Deep Optics for Single-shot High-dynamic-range Imaging », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1375–1385.
- [MJL16] L. Mai, H. Jin, and F. Liu, « Composition-Preserving Deep Photo Aesthetics Assessment », *in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, DOI: 10.1109/CVPR.2016.60.
- [MKV08] Tom Mertens, Jan Kautz, and Frank Van Reeth, « Exposure Fusion: A Simple and Practical Alternative to High Dynamic Range Photography », *in: Computer Graphics Forum* 28 (Sept. 2008), pp. 161–171, DOI: 10.1111/j.1467-8659.2008.01171.x.
- [MLC17] S. Ma, J. Liu, and C. W. Chen, « A-Lamp: Adaptive Layout-Aware Multi-Patch Deep Convolution Neural Network for Photo Aesthetic Assessment », *in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, DOI: 10.1109/CVPR.2017.84.
- [MMP12] N. Murray, L. Marchesotti, and F. Perronnin, « AVA: A Large-Scale Database for Aesthetic Visual Analysis », *in: 2012 IEEE Confer-*

-
- ence on Computer Vision and Pattern Recognition*, June 2012, DOI: 10.1109/CVPR.2012.6247954.
- [MNL13] Long Mai, Yuzhen Niu, and Feng Liu, « Saliency aggregation: A data-driven approach », *in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1131–1138.
- [Mol66] Abraham A. Moles, *Information Theory and Esthetic Perception*, Urbana, University of Illinois Press, 1966.
- [MPM13] L. Marchesotti, F. Perronnin, and F. Meylan, « Learning beautiful (and ugly) attributes », *in: BMVC*, vol. 7, 2013, pp. 1–11.
- [NCF22] Daniel Vera Nieto, Luigi Celona, and Clara Fernandez-Labrador, « Understanding Aesthetics with Language: A Photo Critique Dataset for Aesthetic Assessment », *in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022.
- [Nem+15] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi, « Visual attention in LDR and HDR images », *in: 9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, CONF, 2015.
- [NW97] Kenneth H Norwich and Willy Wong, « Unification of psychophysical phenomena: The complete form of Fechner’s law », *in: Perception & Psychophysics* 59.6 (1997), pp. 929–940.
- [Pra+19] K Ram Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, and R Venkatesh Babu, « A fast, scalable, and reliable deghosting method for extreme exposure fusion », *in: 2019 IEEE International Conference on Computational Photography (ICCP)*, IEEE, 2019, pp. 1–8.
- [Qui+19] Facundo Quiroga, Jordina Torrents-Barrena, Laura Lanzarini, and Domenec Puig, « Measuring (in) variances in Convolutional Networks », *in: Conference on Cloud Computing and Big Data*, Springer, 2019, pp. 98–109.
- [Ran+19] Aakanksha A Rana, Praveer Singh, Giuseppe Valenzise, Frédéric Dufaux, Nikos Komodakis, and Aljosa Smolic, « Deep Tone Mapping Operator for High Dynamic Range Images », *in: IEEE Transactions on Image Processing* 29.1 (Dec. 2019), pp. 1285–1298, DOI: 10.1109/TIP.2019.2936649.

-
- [Rei+10] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting*, Morgan Kaufmann, 2010.
- [Ren+17a] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, « Personalized Image Aesthetics », *in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, DOI: 10.1109/ICCV.2017.76.
- [Ren+17b] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran, « Personalized image aesthetics », *in: Proceedings of the IEEE international conference on computer vision*, 2017, pp. 638–647.
- [RMT20] Gajjala Viswanatha Reddy, Snehasis Mukherjee, and Mainak Thakur, « Measuring photography aesthetics with deep CNNs », *in: IET Image Processing* 14.8 (2020), pp. 1561–1570.
- [Rou+19] Maxime Rousselot, Xavier Ducloux, Olivier Le Meur, and Rémi Cozot, « Quality metric aggregation for HDR/WCG images », *in: 2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 3786–3790.
- [Rou19] Maxime Rousselot, « Image quality assessment of High Dynamic Range and Wide Color Gamut images », PhD thesis, Université Rennes 1, 2019.
- [Sch+23] Sven Schultze, Ani Withöft, Larbi Abdenebaoui, and Susanne Boll, « Explaining Image Aesthetics Assessment: An Interactive Approach », *in: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR '23*, Thessaloniki, Greece: Association for Computing Machinery, 2023, pp. 20–28, ISBN: 9798400701788, DOI: 10.1145/3591106.3592217, URL: <https://doi.org/10.1145/3591106.3592217>.
- [She22] James Shelley, « The Concept of the Aesthetic », *in: The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Spring 2022, Metaphysics Research Lab, Stanford University, 2022.
- [SRK20] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari, « Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss », *in: ACM Transactions on graphics*, vol. 39, 4, July 2020.

-
- [Sun+20] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide, « Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1386–1396.
- [SZ14] K. Simonyan and A. Zisserman, « Very deep convolutional networks for large-scale image recognition », *in: arXiv preprint arXiv:1409.1556* (2014).
- [SZ15] Karen Simonyan and Andrew Zisserman, « Very Deep Convolutional Networks for Large-Scale Image Recognition », *in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, ed. by Yoshua Bengio and Yann LeCun, 2015.
- [Sze+16a] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, « Rethinking the inception architecture for computer vision », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [Sze+16b] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, « Rethinking the inception architecture for computer vision », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [TM18a] H. Talebi and P. Milanfar, « NIMA: Neural Image Assessment », *in: IEEE Transactions on Image Processing* 27.8 (2018), pp. 3998–4011, DOI: 10.1109/TIP.2018.2831899.
- [TM18b] A. Tifentale and L. Manovich, « Competitive Photography and the Presentation of the Self », *in: Exploring the Selfie*, Springer, 2018, pp. 167–187.
- [VKD22] Giuseppe Valenzise, Chen Kang, and Frédéric Dufaux, « Advances and challenges in computational image aesthetics », *in: Human Perception of Visual Information* (2022), pp. 133–181.
- [Wan+16] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang, « Brain-inspired deep networks for image aesthetics assessment », *in: arXiv preprint arXiv:1601.04155* (2016).
- [Wan+18] Wenshan Wang, Su Yang, Weishan Zhang, and Jiulong Zhang, « Neural Aesthetic Image Reviewer », *in: IET Computer Vision* 13 (Feb. 2018), DOI: 10.1049/iet-cvi.2019.0361.

-
- [Wan+19a] Wenshan Wang, Su Yang, Weishan Zhang, and Jiulong Zhang, « Neural aesthetic image reviewer », *in: IET Computer Vision* 13.8 (2019), pp. 749–758.
- [Wan+19b] Wenshan Wang, Su Yang, Weishan Zhang, and Jiulong Zhang, « Neural aesthetic image reviewer », *in: IET Computer Vision* 13.8 (2019), pp. 749–758.
- [Wan11] Z. Wang, « Applications of objective image quality assessment methods [applications corner] », *in: IEEE Signal Processing Magazine* 28.6 (2011), pp. 137–142.
- [Wen+19] Long Wen, X Li, Xinyu Li, and Liang Gao, « A new transfer learning based on VGG-19 network for fault diagnosis », *in: 2019 IEEE 23rd international conference on computer supported cooperative work in design (CSCWD)*, IEEE, 2019, pp. 205–209.
- [WL09] L.-K. Wong and K. L. Low, « Saliency-enhanced image aesthetics class prediction », *in: 2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009.
- [WM21] Zhihua Wang and Kede Ma, « Active fine-tuning from gMAD examples improves blind image quality assessment », *in: IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [WSB03] Z. Wang, E. P. Simoncelli, and A. C. Bovik, « Multiscale structural similarity for image quality assessment », *in: The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, IEEE, 2003, pp. 1398–1402.
- [Yu+21] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao, « Luminance attentive networks for hdr image and panorama reconstruction », *in: Computer Graphics Forum*, vol. 40, 7, Wiley Online Library, 2021, pp. 181–192.
- [Zen+15] Kun Zeng, Jun Yu, Ruxin Wang, Cuihua Li, and Dacheng Tao, « Coupled deep autoencoder for single image super-resolution », *in: IEEE transactions on cybernetics* 47.1 (2015), pp. 27–37.
- [Zha+21] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang, « Uncertainty-aware blind image quality assessment in the laboratory and wild », *in: IEEE Transactions on Image Processing* 30 (2021), pp. 3474–3486.

-
- [Zha+22] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma, « Continual learning for blind image quality assessment », *in: IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [Zhu+20] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi, « MetaIQA: Deep meta-learning for no-reference image quality assessment », *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14143–14152.
- [Zhu+22] Hancheng Zhu, Yong Zhou, Rui Yao, Guangcheng Wang, and Yuzhe Yang, « Learning image aesthetic subjectivity from attribute-aware relational reasoning network », *in: Pattern Recognition Letters* 155 (2022), pp. 84–91.
- [Zop+18] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, « Learning transferable architectures for scalable image recognition », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.



Titre : Améliorer les images grâce à l'imagerie HDR et à l'évaluation automatique d'esthétique

Mot clés : Intelligence Artificielle ; Évaluation automatique d'esthétique ; Imagerie HDR

Résumé : Pour traiter la grande quantité de données visuelles disponible, il est important de concevoir des algorithmes qui peuvent trier, améliorer, compresser ou stocker des images et des vidéos. Dans cette thèse, nous proposons deux approches différentes pour améliorer la qualité d'images.

Tout d'abord, nous proposons une étude des méthodes d'évaluation automatique de l'esthétique. Ces algorithmes sont basés sur des réseaux de neurones supervisés. Nous avons récolté des images de différents types, puis nous avons utilisé ces images pour tester des modèles. Notre étude montre que les caractéristiques nécessaires pour évaluer précisément les esthétiques de photographies professionnelles ou compétitives sont différentes,

mais qu'elles peuvent être apprises par un seul et unique réseau.

Enfin, nous proposons de travailler sur les images à grande gamme dynamique (*High Dynamic Range*, HDR en anglais). Nous présentons ici un nouvel opérateur pour augmenter la gamme dynamique d'images standards, appelé HDR-LFNet. Cet opérateur fusionne la sortie de plusieurs algorithmes pré-existants, ce qui permet d'avoir un réseau plus léger et plus rapide. Nous évaluons les performances de la méthode proposée grâce à des métriques objectives, ainsi qu'une évaluation subjective. Nous prouvons que notre méthode atteint des résultats similaires à l'état de l'art en utilisant moins de ressources.

Title: Improving Image Quality using High Dynamic Range and Aesthetics Assessment

Keywords: Artificial Intelligence; Aesthetics Assessment; HDR imaging

Abstract:

To cope with the increasing amount of visual content available, it is important to devise automatic processes that can sort, improve, compress or store images and videos. In this thesis, we propose two different approaches to software-based image improvement.

First, we propose a study on existing aesthetics assessment algorithms. These algorithms are based on supervised neural networks. We have collected several datasets of images, and we have tested different models using these images. We report here the performances of such networks, as well as an idea to improve the already trained networks. Our study shows that the features needed to accu-

rately predict the aesthetics of competitive and professional are different but can be learned simultaneously by a single network.

In a second time, we propose to work with High Dynamic Range (HDR) images. We present here a new operator to increase the dynamic range of images called HDR-LFNet, that merges the output of existing operators and therefore, consists in far fewer parameters. Besides, we evaluate our method through objective metrics and a user study. We show that our method is on-par with the state-of-the-art according to objective metrics, but is preferred by observers during the user study, while using less resources overall.