Semi-Unbalanced Optimal Transport for Image Restoration and Synthesis

Simon Mignon, Bruno Galerne, Moncef Hidane, Cécile Louchet, Julien Mille

Semi-Unbalanced Optimal Transport for Image Restoration and Synthesis

Simon Mignon* Bruno Galerne*† Moncef Hidane‡
Cécile Louchet* Julien Mille‡

March 21, 2024

Abstract

In this paper, we build on optimal transport (OT) theory to present a novel asymmetrically unbalanced variant, the semi-unbalanced optimal transport (SUOT), specifically designed for imaging applications with the presence of a reference. SUOT addresses the lack of robustness of OT and the rigidity inherited from its formulation by taking inspiration from the unbalanced OT formulation. Rather than relaxing the constraints on both the source and the target measures, we relax only the marginal related to the reference. We consider both the unregularized and entropy-regularized versions, deriving dual formulations, corresponding minimization algorithms and formulas for the gradient. These derivations enable us to employ SUOT in variational inverse imaging and synthesis problems, as well as a loss for training a neural network. We evaluate the use of SUOT in a reference-driven super-resolution problem and show its benefits. We also incorporate it into a state-of-the-art single-image generation algorithm and show that it leads to increased diversity. Our results advocate for the adoption of SUOT as a general tool for variational and learning-based inverse imaging and synthesis problems with the presence of a reference.

1 Introduction

At the heart of numerous algorithms designed to tackle image processing problems is the modeling of image statistics. Two related yet distinct considerations are usually involved: how to model image statistics, and how to incorporate a known given model into a computational procedure. In this article, we consider inverse imaging and synthesis problems, presenting both a unified modeling tool and a corresponding algorithm. Our working hypothesis is that we have access, *at recovery or synthesis time*, to a *reference* image, related to (but different from) the *specific* unknown image one seeks to recover or to synthesize. We turn this reference into an image model and an associated computational

*Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS, Orléans, France (simon.mignon@univ-orleans.fr, bruno.galerne@univ-orleans.fr, cecile.louchet@univ-orleans.fr).

†Institut Universitaire de France (IUF).

‡Laboratoire d'Informatique Fondamentale et Appliquée de Tours, UR6300, INSA Centre Val de Loire, Université de Tours, France (moncef.hidane@insa-cvl.fr, julien.mille@insa-cvl.fr).

procedure by adopting a variational approach. In doing so, we leverage recent advances in computational optimal transport to *explicitly penalize deviation of the patch distribution of the solution from the one of the available reference*. Additionally, we illustrate the use of a neural network to *amortize* the cost involved in minimizing the proposed energy, that is, we use the proposed energy as a loss for training a neural network.

1.1 Related Work

Inverse Imaging Problems

In inverse imaging problems, one seeks to recover an image $x^\star \in \mathcal{X} := \mathbb{R}^N$ from noisy measurements $y \in \mathcal{Y} := \mathbb{R}^K$, $K \leq N$. The classical approach for solving such problems involves two modeling assumptions: a forward model that relates the sought image $x^\star$ to the obtained measurements y, and a prior model that encapsulates existing knowledge about the class of images one aims to recover. In the variational approach, the two previous modeling assumptions translate into an *energy function* $\mathcal{E}$ composed of a data fitting term $D(\cdot; y)$ and a regularizer R. An estimate $\hat{x}(y)$ of $x^\star$ is then obtained as a minimizer of $\mathcal{E}$:

$$\hat{x}(y) \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathcal{E}(x; y) := R(x) + D(x; y), \tag{1}$$

where $\operatorname{argmin} \mathcal{E}$ is the set of minimizers of $\mathcal{E}$.

Classical analytical regularizers that impose structural properties such a sparsity [59, 66] or patch redundancy [13] have now been largely replaced by *data-driven* ones. These come in two flavors: either explicit or implicit, the latter corresponding to the family of plug-n-play methods. Considering the substantial volume of relevant literature, we refer the interested readers to the survey papers [3, 35, 52].

The recent years have also seen the emergence of neural-network-based *end-to-end* approaches for solving inverse imaging problems [35, 49, 52]. These methods employ *paired examples* $(x_i^\star, y_i)$ to learn a mapping from $\mathcal{Y}$ to $\mathcal{X}$. Despite demonstrating favorable empirical performance, numerous issues have been reported in the literature. Notably, a significant challenge involves the need to retrain the network whenever the degradation model changes. Additional concerns include potential instability [31], a lack of interpretability, and how to incorporate the forward model into the network architecture.

Enforcing Statistical Constraints

A series of works [18,21,67,68] has proposed to complement established variational methods with penalties that enforce soft statistical constraints on the solution. These penalties usually seek to align the distribution of certain features of the solution with a reference empirical distribution. In [18], the authors consider penalizing the Kullback-Leibler (KL) divergence between the gradient distribution of the solution and a reference gradient distribution. The latter is estimated, locally, from the degraded image. Pixel values are used as features in [42, 67] with an optimal transport (OT) [72] cost. Closeness to reference histograms of high-pass filter responses is also enforced in [21, 68] through the use of an OT cost.

The OT-based methods mentioned earlier [21, 42, 67, 68] use one-dimensional features due to the availability of a closed-form expression for the OT cost in this case. In [37], Hertrich *et al.* enforce OT-closeness to the *patch* distribution of a reference image. This

so-called Wasserstein patch prior (WPP) is applied to a super-resolution (SR) problem where one assumes that a reference image is available.

Statistical losses between feature distributions have also been used for training end-to-end networks that perform image restoration. Mechrez *et al.* [47] use an approximation of the KL divergence between the distributions of the features of the predicted output and the ground truth. Interestingly, the features themselves are computed as the feature maps generated by a neural network, generally VGG [64]. This approach was introduced by Gatys *et al.* [27] in the context of texture synthesis and has since been employed to define what are now commonly referred to as perceptual losses [41]. In [20], a pixelwise loss is combined with an OT-based perceptual loss. The authors of [1] amortize the minimization of the WPP energy proposed in [37] by training a neural network called WPPNets. Additionally, they propose to learn a conditional normalizing flow [53] that samples from the distribution whose negative log density is the WPP energy.

Texture and Single-Image Synthesis

Synthesizing a texture from a single example is a long-standing image processing problem. In the variational formulation of this problem, an optimization problem is solved for each new synthesis. This optimization problem constrains the statistics of relevant descriptors of the synthesized image to be close to those of the example [55]. As mentioned earlier, Gatys *et al.* [27] pioneered the use of the feature space of a deep convolutional neural network as a descriptor. Synthesis is then performed by matching the Gram matrices of the feature maps. The link to the Maximum Mean Discrepancy distance [33] has been established in [45]. An OT cost between aggregated one-dimensional features is used in [68] and texture mixing has been conducted in [71] using OT distance between elliptical distributions in feature space. OT in patch space for texture synthesis has been introduced in [26, 34]. Extension to deep features is considered in [40]. A sliced-OT formulation has been adopted in [36], following the formulation initially proposed for texture mixing in [57].

The variational approach of Gatys *et al.* has been amortized using a neural network in [70]. Once such a *texture network* G is trained, one can sample a new texture $x = G(z)$ starting from a random vector z. A GAN approach [30] to texture networks has been considered in [10].

Closely related to texture synthesis is the single-image generation (SIG) problem popularized by the SinGAN approach [63]. The primary objective of SIG is to generate visually realistic and diverse images that bear a strong resemblance to a given reference image. This allows a diverse range of image editing applications [32]. A variational sliced-OT approach for SIG that uses the patch distribution of the reference as a prior is presented in [22]. An approach combining patch-nearest-neighbors search with an OT cost is proposed in [16].

Robust Optimal Transport

The previously mentioned works that employ OT to impose statistical constraints on the recovered or synthesized image are all based on the classical formulation of OT, where *strict marginal constraints* are enforced on the transport plan. This leads to two well-known issues: non-robustness to outliers, and preservation of the relative frequencies of the features. Both issues are illustrated in Section 2. The first issue is related to the fact that a small fraction of outlier mass can significantly influence the *value* of the OT

cost [4, 50, 65]. The second concern is related to the fact that the use of OT as a cost tends to replicate, in the computed image, the frequencies of the features of the reference. This behaviour is generally not the one sought: one is interested in finding an image that exhibits coherence with respect to the reference, rather than one that exactly matches the frequencies of its features.

To address the latter concern, the authors of [37] apply the WPP regularizer to an artificially padded version of the unknown image, keeping the fidelity term unaffected. This allows outlier patches of the reference to be aggregated into the artificial bounds of the output. Rather than matching the feature distribution of a single output image to a single reference, the approach taken by the authors of [22] involves generating N output images. These output images have their *aggregate* feature distribution matched to that of the reference. This strategy ensures strict preservation of the *total* feature distribution of the reference, while still allowing for variation *within* each generated image.

A more principled and systematic way of addressing both previous issues consists in adopting an OT formulation that allows for *partial displacement of mass*. This formulation has been introduced by Benamou in [7] under the name unbalanced optimal transport (UOT), initially for dealing with measures of different masses. A static formulation with (information) divergence-based approximate marginal constraints has been introduced in [46] and considered in [11, 14]. [17, 25] consider the same formulation but with an added entropy term, following [19]. Interested readers can refer to [62]. For completeness, we note that robust versions of OT have also been proposed in [4, 44, 50, 51]. The main difference between these versions and UOT is that, in the latter, the divergence-based approximate marginal constraints appear as regularizers added to the objective, while in the former, they are included as constraints.

Let us finally remark that a semi-relaxed version of OT, where one relaxes only one marginal constraint, has been considered in [56] in the context of color image transfer. A similar formulation has since been mentioned in [11, 44].

1.2 Contributions and Plan of the Paper

We consider in this paper both inverse imaging and synthesis problems with the presence of a reference image. In the case of inverse problems, the reference corresponds to a high-quality image related to, but different from the one we want to recover. This setting is relevant each time one aims at recovering an image from a narrow family of images, e.g., the same texture or the same material [37] for which reference samples are available. In the case of image synthesis, the reference corresponds to an example from which we want to synthesize similarly looking samples but with a high degree of diversity [16, 22, 63]. Following [26, 37], we tackle both problems through a variational approach where we explicitly penalize deviation of the *multiscale patch distribution* of the solution from the one of the reference. We also consider the use of the introduced penalty as a loss for training a neural network, as done in [1].

A principled and well understood way to penalize such a discrepancy is to use OT as a cost [22, 37, 40]. As discussed earlier, this approach is not robust to outliers present in the reference and it also tends to replicate the frequencies of the features of the reference into the solution. Rather than relying on heuristic solutions, we present a systematic approach for dealing with these issues. Taking inspiration from the static UOT formulation of [17, 25, 46], we propose an asymmetric formulation, hereafter called *semi-unbalanced* OT (SUOT) where we *relax the constraint related to the preservation of the mass of the*

reference, while enforcing strict preservation of the mass of the unknown image. We study both the unregularized and entropy-regularized versions of SUOT, deriving dual formulations, corresponding minimization algorithms and formulas for the gradient of the cost with respect to the support of one of the measures. The latter allows to use a gradient descent scheme either to minimize an energy, or to learn a neural network, both involving a SUOT cost.

For the numerical evaluation of our method in the context of inverse imaging problems, we concentrate on the SR problem described in [1, 37] that already produces state-of-the-art results, improving upon variational methods that use learned deep regularizers. We observe that our proposed method improves robustness when the reference presents outlier values. Even without outlier values, the ability of our method to explicitly modify the frequency of the patches between the reference and the computed solution leads to improved results with respect to the original OT-based works [1, 37].

In the context of single image synthesis, the PSinOT method [16] exhibits superior performance compared to SinGAN [63], and its results are on par with those achieved by the state-of-the-art methods [22, 32]. We show that incorporating SUOT into PSinOT leads to increased diversity, as measured by the Single Image Fréchet Inception Distance (SIFID) [63] and by the per-pixel standard deviation calculated from 50 generated images.

The organization of the paper is as follows: In Section 2, we provide a brief review of optimal transport, illustrate the issue related to its lack of robustness and explain how we adapt it to our semi-unbalanced formulation. We also provide all relevant details for its use as a loss in inverse imaging and synthesis problems. We finally provide exact formulas and asymptotics for the case involving one Dirac and two Dirac measures, showcasing quantitatively the difference between OT and SUOT. Section 3 studies the entropy-regularized SUOT formulation. In Sections 4 and 5, we demonstrate the effectiveness of our approach in the context of inverse imaging problems and SIG, respectively. Conclusions and perspectives are drawn in Section 6.

A preliminary version of this work appeared in [48] wherein only the entropy-regularized version was proposed and the only application considered was related to inverse problems. Our code is available at: https://github.com/SimonMignon/SUOT-for-reference-based-image-restoration-and-synthesis.

2 Semi-Unbalanced Optimal Transport

We start this section by describing the notations that we use in the rest of the paper. For an integer $N \geq 1$, let $[\![N]\!] = \{1, \dots, N\}$. We consider discrete probability measures with support in $\mathbb{R}^n$, $n \geq 1$. Throughout the rest of the paper, $\alpha = \sum_{i=1}^{N} a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{M} b_j \delta_{y_j}$ will denote two such measures, with $x_i, y_j \in \mathbb{R}^n$, $a = (a_i)_{i \in [\![N]\!]} \in \Sigma_N$, $b = (b_j)_{j \in [\![M]\!]} \in \Sigma_M$, Σ_P being the probability simplex in $\mathbb{R}^P$. Without loss of generality, we assume throughout the paper that $a_i > 0$ and $b_j > 0$ for all i and j. The product measure of α and β will be denoted $\alpha \otimes \beta$. Given a continuous *ground cost* function $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, we consider the cost matrix $C \in \mathbb{R}^{N \times M}$ such that $c_{i,j} = c(x_i, y_j)$. For any matrix $\pi \in \mathbb{R}^{N \times M}$, let $\pi_1 = \pi \mathbf{1}_M$ and $\pi_2 = \pi^T \mathbf{1}_N$. The set of matrices with nonnegative entries is denoted $\mathbb{R}_+^{N \times M}$.

Given a set E, ι_E denotes its characteristic function: $\iota_E(x) = 0$ if $x \in E$ and $\iota_E(x) = +\infty$ if $x \in E^c$. When $E = \{v\}$, we write ι_v instead of $\iota_{\{v\}}$. Given two vectors $f \in \mathbb{R}^N$ and $g \in \mathbb{R}^M$, $f \oplus g \in \mathbb{R}^{N \times M}$ is defined by $(f \oplus g)_{ij} = f_i + g_j$, and $f \otimes g \in \mathbb{R}^{N \times M}$ is

defined by $(f \otimes g)_{ij} = f_i g_j$. When $N = M$, element-wise multiplication is denoted $f \odot g$. The order relations between matrices are to be understood element-wise. We define $\mathbf{0}_N$ as the zero vector of length N.

2.1 Background on Optimal Transport

The OT cost between α and β is defined as

$$\mathrm{OT}(\alpha, \beta) = \min_{\pi \in \mathbb{R}_+^{N \times M}} \langle C, \pi \rangle + \iota_a(\pi_1) + \iota_b(\pi_2), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product. Strong duality of linear programs allows to express (2) in a dual form [54, 60]:

$$\mathrm{OT}(\alpha, \beta) = \max_{(f,g) \in \Gamma(C)} \langle a, f \rangle + \langle b, g \rangle \,, \tag{3}$$

with $\Gamma(C) = \{(f, g) \in \mathbb{R}^N \times \mathbb{R}^M \mid f \oplus g \leq C\}$.

Considering the function to be maximized in the dual Problem (3), and keeping the value of g constant, it appears that a solution for maximizing with respect to (wrt) f is given by the c-transform $g^c \in \mathbb{R}^N$ of g defined by

$$\forall i \in [\![N]\!], \quad g_i^c = \min_{j \in [\![M]\!]} c_{i,j} - g_j. \tag{4}$$

This leads to a *concave unconstrained* maximization formulation of the OT cost, referred to as the *semi-dual* problem [54, 60]:

$$\mathrm{OT}(\alpha, \beta) = \max_{g \in \mathbb{R}^M} \langle a, g^c \rangle + \langle b, g \rangle \,. \tag{5}$$

The function

$$SG_F(g) = b - \sum_{i=1}^{N} a_i e_{j(i,g)} \tag{6}$$

is a super-gradient of the concave function $F(g) = \langle a, g^c \rangle + \langle b, g \rangle$, where $(e_j)_{j \in [\![M]\!]}$ denotes the canonical basis of $\mathbb{R}^M$ and

$$j(i, g) \in \operatorname*{argmin}_{j \in [\![M]\!]} c(x_i, y_j) - g_j \tag{7}$$

is a minimal index for the c-transform g^c [39]. This allows to solve (5) using a standard averaged (super-)gradient ascent algorithm (AGAA) [29, 37].

When $c(x, y) = \|x - y\|^p$ for some $p \geq 1$, $\mathrm{OT}^{1/p}$ corresponds to the p-Wasserstein distance [72]. This distance has been extensively used as a loss function in parameter estimation problems [6], usually with $\alpha = \alpha_\theta$ being a model distribution, parameterized by an unknown vector θ, and β an empirical measure associated with training data [54, Chap. 9]. In this paper, we are interested in the setting where $\theta = (x_i)_{i \in [\![N]\!]}$, that is, *the parameters to infer are the support of* α. Evaluating $\nabla_{(x_i)_{i \in [\![N]\!]}} \mathrm{OT}(\alpha, \beta)$ is thus necessary for gradient-based learning. It is known that the differentiability of $\mathrm{OT}(\alpha, \beta)$ wrt to a mass location x_i is related to the Laguerre tesselation associated with the solution $g^\star$ of the maximization Problem (5) [39]: If x_i belongs to the open

Laguerre cell $\mathcal{L}_j(g^\star) = \{x \in \mathbb{R}^d \mid \forall j' \neq j, c(x, y_j) - g_j^\star < c(x, y_{j'}) - g_{j'}^\star\}$ for some (unique) $j = j(i, g^\star) \in [\![M]\!]$ then $\mathrm{OT}(\alpha, \beta)$ is differentiable wrt x_i and

$$\nabla_{x_i} \mathrm{OT}(\alpha, \beta) = \nabla_{x_i} \mathrm{OT}\left(\sum_{k=1}^N a_k \delta_{x_k}, \beta\right) = a_i \nabla_{x_i} c(x_i, y_{j(i,g^\star)}). \tag{8}$$

However, this formula is only valid for points x_i whose mass is fully transported to a single target y_j. In general, mass splittings necessarily happen and the formula is not valid for all x_i. Still, in practice, one computes the gradient using this formula by randomly selecting one of the points y_j for which the c-transform is minimal. As shown by previous work, this is sufficient for image processing applications [1, 37, 39].

2.2 Motivation

As explained in the introduction, our use of an OT-type cost in this paper aims at enforcing statistical coherence between the multiscale patch distribution of a restored or synthesized image, and the one of a reference image. Before presenting both applications, and the solution we propose, we would like to illustrate, through 2-D examples, the shortcomings mentioned in the introduction, related to the use of the standard, *balanced*, version of the OT cost (2).

In Figure 1, we show the evolution of an initial point cloud $x = (x_i)_{i \in [\![N]\!]} \in \mathbb{R}^{N \times 2}$ under the gradient flow

$$\dot{z}(t) = -\nabla_z \mathrm{OT}(\alpha_{z(t)}, \beta), \tag{9}$$

where $\beta = \frac{1}{M} \sum_{j=1}^M \delta_{y_j}$ is fixed and $\alpha_z = \frac{1}{N} \sum_{i=1}^N \delta_{z_i}$, $z_i \in \mathbb{R}^2$ for all i. The discrete measure α in the figure corresponds to $\alpha_{z(0)} = \alpha_x = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$. The notation $\nabla_z \mathrm{OT}(\alpha_z, \beta)$ in (9) refers to the gradient wrt the support of α_z. The flow is implemented through a forward Euler scheme, the gradients $\nabla_z \mathrm{OT}(\alpha_{z(t)}, \beta)$ being computed with Equation (8). We stress that each evaluation of such gradients involves approximating an OT cost using an AGAA.

The first column of Figure 1 displays the initial distribution $\alpha = \alpha_{z(0)}$ and the target distribution β. The second column displays the evolution of α to the distribution $\alpha_\infty := \alpha_{z_\infty}$ associated with the steady state solution z_∞ of the ODE (9), as implemented by the forward Euler scheme. The trajectories of the points $(z_i)_{i \in \{1,\dots,N\}}$ during the gradient descent process are depicted with green lines, illustrating the transportation of points from the support of α to their destination within the support of α_∞. The first row corresponds to a configuration where the target β has two clusters, while in the second row β has a unique cluster but with a single outlier point. It is worth noticing that in both rows, α and β have the same number of atoms.

As can be seen in Figure 1, throughout the OT-based gradient flow, the mass of α gets redistributed across the *entire support* of the target β. However, this characteristic can pose challenges, particularly in imaging applications, where the distribution of the features of the computed image is not expected to precisely match the one of the reference, because of either a difference in the proportion of the atoms, or the existence of outliers in the reference.

To address this issue, we take inspiration from the static formulation of unbalanced OT [17,46], and decide to relax the constraint $\pi_2 = b$ into a penalty $\rho \mathcal{D}(\pi_2 | b)$, where $\rho > 0$, and $\mathcal{D}$ is a given divergence, while strictly enforcing the constraint $\pi_1 = a$. Columns 3 to 7 of Figure 1 show the steady state distribution α_∞^ρ obtained by replacing in (9) OT by

our semi-unbalanced formulation SUOT, for different values of ρ (with $\mathcal{D} = \mathrm{KL}$). One can see that, by adjusting the parameter ρ, SUOT can selectively ignore distant data points in β. As ρ approaches infinity, SUOT reverts to OT, and the support of α_∞ is again spread across the one of β. The precise definitions and computational tools are given next.

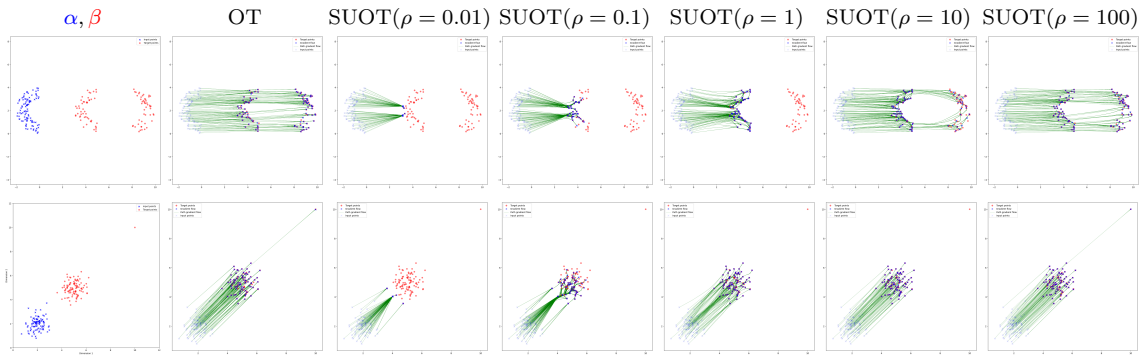

Figure 1: **Two instances of 2D gradient flows achieved by minimizing OT and SUOT between the distributions represented by α and β.** These gradient flows are realized through gradient descent concerning the data points of α, with the trajectory of each point during descent illustrated by green lines. By fine-tuning the parameter ρ (values: 0.01, 0.1, 1, 10,100), SUOT can selectively exclude distant data points in β, whereas OT takes into account all data points within β. In the first example, it becomes feasible to specifically target the first "croissant" in β that closely resembles α. In the second example, we have the capability to exclude anomalous data points from β.

2.3 Semi-Unbalanced Optimal Transport

2.3.1 Primal, Dual, Semi-Dual Formulations and Gradient

Let us now precisely define the problem of semi-unbalanced optimal transport.

Definition 1 (Semi-Unbalanced Optimal Transport)**.** *For $\rho > 0$, the SUOT cost of level ρ between α and β is defined by*

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \min_{\pi \in \mathbb{R}_+^{N \times M}} \langle C, \pi \rangle + \iota_a(\pi_1) + \rho \mathrm{KL}(\pi_2 | b), \tag{10}$$

where

$$\mathrm{KL}(u|v) = \sum_{j=1}^K u_j \log\left(\frac{u_j}{v_j}\right) - \sum_{j=1}^K u_j + \sum_{j=1}^K v_j$$

is the KL divergence between two vectors $u \geq 0$ and $v > 0$ of size K, with the convention $0 \log(0) = 0$.

Remark 2. *When dealing with probability vectors u and v, KL divergence can be reduced to $\sum_{j=1}^K u_j \log(\frac{u_j}{v_j})$. However, we choose to use the general expression as it simplifies the calculation of the gradient wrt u.*

As discussed earlier, the difference between SUOT (10) and OT (2) is that we have replaced the constraint $\pi_2 = b$ with the penalty $\rho \mathrm{KL}(\pi_2|b)$. It is worth noticing that due

to the constraint on π_1, π_2 is still in Σ_M, though now some of its entries might be equal to zero. The choice of the parameter ρ plays a crucial role in determining the balance between the source and target distributions. A very low value of ρ will result in a strong imbalance, where the transportation plan might be heavily biased towards regions that are close to the support of α. Conversely, a high value of ρ will lead to a classical OT plan.

Theorem 3 (Dual formulation)**.** *The SUOT cost can be expressed in the following dual form:*

$$\mathrm{SUOT}^{\rho}(\alpha, \beta) = \max_{(f,g)\in\Gamma(C)} \langle a, f\rangle - \langle b, \phi^*(-g)\rangle\,, \tag{11}$$

where $\Gamma(C) = \{(f, g) \in \mathbb{R}^N \times \mathbb{R}^M; f \oplus g \leq C\}$ *and* $\phi^*(q) = \rho\left(\exp\left(\frac{q}{\rho}\right) - 1\right)$*, the* exp *function being applied component-wise.*

The proof given in Appendix A.1 relies on applying the Fenchel-Rockafellar theorem, which is recalled in the appendix (see Theorem 12).

Remark 4. *In a broader context, as outlined in [17],* ϕ^* *denotes the Legendre conjugate of an entropy function* $\phi : \mathbb{R}_+ \to \mathbb{R}_+$*, which is convex, positive, lower-semi-continuous, and satisfies* $\phi(1) = 0$*. The domain of* ϕ *is extended to* $\mathbb{R}$ *by setting* $\phi(x) = +\infty$ *for* $x < 0$*. This function* ϕ *is associated with a* ϕ*-divergence, defined in our case (i.e., for discrete measures with the same support) as* $\mathrm{D}_\phi(\alpha|\beta) = \sum_{i=1}^M b_i \phi\left(\frac{a_i}{b_i}\right)$*. We have selected* $\phi(q) = \rho(q\log(q) - q + 1)$*, which results in* $\mathrm{D}_\phi = \rho KL$*.*

Similar to OT, SUOT can be expressed as an unconstrained concave problem.

Proposition 5 (Semi-dual formulation for SUOT)**.** *The SUOT cost can be expressed as follows*

$$\mathrm{SUOT}^{\rho}(\alpha, \beta) = \max_{g\in\mathbb{R}^M} \langle a, g^c\rangle - \langle b, \phi^*(-g)\rangle\,. \tag{12}$$

The proof is given in Appendix A.2.

Now we can establish the differentiability of $\mathrm{SUOT}^{\rho}(\alpha, \beta)$ with respect to $(x_i)_{i\in[\![N]\!]} \in \mathbb{R}^n$.

Proposition 6 (Gradient with respect to $(x_i)_{i\in[\![N]\!]}$)**.** *Let* $g^\star$ *be a solution to Problem* (12) *and assume that* x_i *belongs to a Laguerre cell* $\mathcal{L}_j(g^\star)$ *for some unique* $j = j(i, g^\star)$*. Then,* $\mathrm{SUOT}^{\rho}(\alpha, \beta)$ *is differentiable wrt* x_i*, and we have*

$$\nabla_{x_i}\mathrm{SUOT}^{\rho}\left(\sum_{k=1}^N a_k\delta_{x_k}, \beta\right) = a_i\nabla_{x_i}c(x_i, y_{j(i,g^\star)}). \tag{13}$$

The proof is adapted from the similar result for OT [39] and is given in Appendix A.3.

2.3.2 Computing the SUOT Cost

We propose to adapt the approach of [40] to solve Problem (12) using a standard AGAA.

Proposition 7 (Concavity and Super-gradient of the semi-dual SUOT Functional)**.** *The functional* $F(g) = \langle a, g^c\rangle - \langle b, \phi^\star(-g)\rangle$ *associated with Problem* (12) *has the following properties:*

(i) F is concave.

(ii) For any $g \in \mathbb{R}^M$ and $j(i,g) \in \arg\min_{j\in\{1,\cdots,M\}} c(x_i, y_j) - g_j$, the expression

$$SG_F(g) = b \odot \exp\left(\frac{-g}{\rho}\right) - \sum_{i=1}^{N} a_i e_{j(i,g)}, \tag{14}$$

is a super-gradient of $F(g)$.

The proof is given in Appendix A.4

2.3.3 Exact Computation of SUOT in a Simple Case

We consider the simple scenario where $\alpha = \delta_x$ and $\beta = b_1\delta_{y_1} + b_2\delta_{y_2}$, with $b_1 + b_2 = 1$ and x, y_1, $y_2 \in \mathbb{R}^n$. Accordingly, in our notations, $N = 1$, $M = 2$, $a = 1$, and $b = (b_1, b_2)$.

To compute (balanced) OT, we need to optimize $\pi_{11}|y_1 - x|^2 + \pi_{12}|y_2 - x|^2$ with respect to the transport map $\pi = (\pi_{11}, \pi_{12})$, subject to the constraints $\pi_{11} = b_1$ and $\pi_{12} = b_2$. Therefore, there is nothing to optimize, and we have

$$\mathrm{OT}(\alpha, \beta) = b_1\|y_1 - x\|^2 + b_2\|y_2 - x\|^2,$$

achieved for $\pi^*_{\mathrm{OT}} = (b_1, b_2)$. The computation of the SUOT cost between α and β is detailed in Appendix B and proves that, for $d = \|y_2 - x\|^2 - \|y_1 - x\|^2$, the optimal transport plan is $\pi^*_{\mathrm{SUOT}} = (1 - \eta^*, \eta^*)$ with

$$\eta^* = \frac{b_2 e^{-\frac{d}{\rho}}}{b_1 + b_2 e^{-\frac{d}{\rho}}} \tag{15}$$

and that

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \|y_1 - x\|^2 - \rho \log(b_1 + b_2 e^{-\frac{d}{\rho}}). \tag{16}$$

In Figure 2, we illustrate this result with a 2-D example. The left plot illustrates the two distributions with the red point symbolizing α and the two blue points representing β, where the area of each point indicates its mass: 1 for the red point and 0.1 and 0.9 for the two blue points, respectively. On the right side of the figure, we plot the OT and SUOT transport plans as a function of ρ, represented by dashed and solid lines, respectively. The OT transport plan, which is independent of ρ, redistributes the mass of α to exactly match the distribution β. In contrast, with SUOT, a wider range of possibilities emerges. It is observed that selecting a small value of ρ allows redistributing most of the mass of the red point to the nearest light blue point. Conversely, choosing a large value of ρ forces π_{12} to tend towards b_2, pushing SUOT towards OT.

Now let us provide some more thorough insight into the transport plan $\pi^*_{\mathrm{SUOT}} = (1 - \eta^*, \eta^*)$ and $\mathrm{SUOT}^\rho(\alpha, \beta)$. First, let us consider η^* from (15) as a mere function of d/ρ. As soon as $\|y_1 - x\| < \|y_2 - x\|$, we have $\eta^* < b_2$. This means that in SUOT, the mass allocated to the more distant point y_2 is necessarily reduced compared to OT. Additionally, η^* is a decreasing function of d/ρ. It goes to b_2 when $d/\rho \to 0$ with an affine behavior given by the first-order expansion:

$$\eta^* \underset{\frac{d}{\rho}\to 0}{\approx} b_2 - b_1 b_2 \frac{d}{\rho}.$$

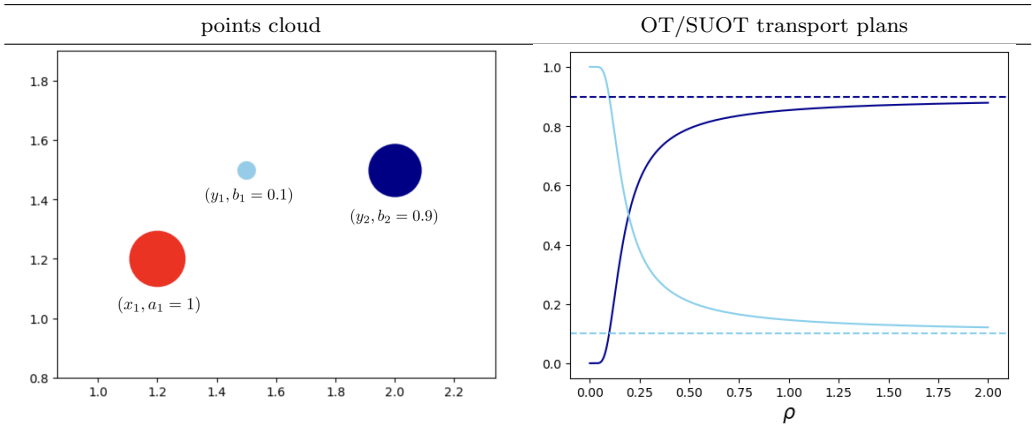

Figure 2: From left to right: the first plot presents the considered 2D example, followed by the plot of the corresponding OT and SUOT transport plans as a function of ρ. The OT transport plan is represented with dashed lines, while the SUOT transport plan is depicted with solid lines. OT is observed to split the mass of the red point to precisely match β. SUOT, in contrast, offers greater flexibility: with a smaller value of ρ, it can prioritize the closer light blue point over the farther dark blue one. Increasing ρ gradually aligns the SUOT transport plan with that of OT, showcasing SUOT's adaptability and advantages over OT.

Considering the only dependence in ρ, this is a rather slow convergence, which shows that the choice of the unbalance parameter ρ is not very sensitive. When $d/\rho \to \infty$, η^* converges to 0 with the asymptotic behavior

$$\eta^* \underset{d/\rho\to\infty}{\approx} \frac{b_2}{b_1} e^{-\frac{d}{\rho}},$$

demonstrating that the mass quickly vanishes when sent too far.

Now let us give some comments on $\mathrm{SUOT}^\rho(\alpha, \beta)$. A first remark, stemming from the very definition of SUOT, is that regardless of the values of the variables, the inequality $\mathrm{SUOT}^\rho(\alpha, \beta) < \mathrm{OT}(\alpha, \beta)$ holds true. Contrary to η^*, $\mathrm{SUOT}^\rho(\alpha, \beta)$ cannot be written as a function of d/ρ, which explains the following analysis wrt d and ρ taken separately. Let us start by noting that $\mathrm{SUOT}^\rho(\alpha, \beta)$ is not bounded, but considering only its dependence wrt ρ (resp. d, contrary to $\mathrm{OT}(\alpha, \beta)$) makes it bounded. Thanks to (16), it is immediate to see that $\mathrm{SUOT}^\rho(\alpha, \beta)$ is an increasing function of d and that when $d \to \infty$,

$$\mathrm{SUOT}^\rho(\alpha, \beta) \xrightarrow[d\to\infty]{} \|y_1 - x\|^2 - \rho \log(b_1),$$

Moreover, as $\rho \to \infty$, a first-order Taylor expansion in (16) confirms that $\mathrm{SUOT}^\rho(\alpha, \beta)$ approaches $\mathrm{OT}(\alpha, \beta)$, with a second-order expansion revealing

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \mathrm{OT}(\alpha, \beta) - \frac{b_1 b_2 d^2}{2\rho} + o(\frac{1}{\rho}).$$

Lastly, as $\rho \to 0$, $\mathrm{SUOT}^\rho(\alpha, \beta)$ converges to $\|y_1 - x\|^2$ which means that y_2 is regarded as an outlier, all the mass from x being transported to its nearest neighbor y_1.

We now come back to the general case and consider the SUOT formulation in the context of entropic regularization [19].

3 Regularized Semi-Unbalanced Optimal Transport

3.1 Background on Entropic Optimal Transport

To speed up the computation of the OT cost, a now common strategy [19] is to add a (negative) entropy term to the objective function in Problem (2), leading to a regularized OT (ROT) formulation. Apart from the computational benefit, the empirical version of the entropic ROT is known to converge to its population version at a rate independent of the dimension of the space on which the measures are defined [28]. However, a well-known problem with the entropic ROT is that it is biased [24]. This is made precise by Rigollet and Weed [58] who show that the ε-ROT-projection of a discrete measure on a class of measures satisfying a so-called closure under dominance hypothesis corresponds to a maximum likelihood estimator in a Gaussian deconvolution model whose standard deviation is precisely ε. Nevertheless, for small values of ε, ROT allows to compare distributions. It is computable, using Sinkhorn's algorithm, and differentiable.

Definition 8 (Regularized Optimal Transport [24, 54])**.** *Let $\varepsilon > 0$ be fixed. The ε-ROT cost between α and β is defined as:*

$$\mathrm{ROT}_\varepsilon(\alpha, \beta) = \min_{\pi \in \mathbb{R}_+^{N \times M}} \langle C, \pi \rangle + \varepsilon \, \mathrm{KL}(\pi | a \otimes b) + \iota_\alpha(\pi_1) + \iota_\beta(\pi_2). \tag{17}$$

Entropic ROT admits the following dual formulation:

$$\mathrm{ROT}_\varepsilon(\alpha, \beta) = \max_{(f,g) \in \mathbb{R}^N \times \mathbb{R}^M} \langle a, f \rangle + \langle b, g \rangle - \varepsilon \left\langle a \otimes b, \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) - 1 \right\rangle. \tag{18}$$

Assuming that $f^\star, g^\star$ are the solutions to Problem (18), we have [23, p. 124]:

$$\nabla_{x_i} \mathrm{ROT}_\varepsilon \left(\sum_{i=1}^N a_i \delta_{x_i}, \beta \right) = a_i \nabla \varphi(x_i), \tag{19}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ has the expression

$$\varphi(x) = -\varepsilon \log \left(\sum_{j=1}^M b_j \exp\left(\frac{g_j^\star - c(x, y_j)}{\varepsilon}\right) \right). \tag{20}$$

Problem (18) is a concave maximization problem which can be solved by Sinkhorn's algorithm.

Theorem 9 (Sinkhorn's algorithm [19, 23, 54])**.** *(i) Starting from any $f^0 \in \mathbb{R}^N$, the following algorithm converges to a solution of Problem* (18)*:*

$$\begin{aligned} g_j^{t+1} &= -\varepsilon \log \left(\sum_{i=1}^N a_i \exp\left(\frac{f_i^t - c_{i,j}}{\varepsilon}\right) \right), \ j \in [\![M]\!], \\ f_i^{t+1} &= -\varepsilon \log \left(\sum_{j=1}^M b_j \exp\left(\frac{g_j^{t+1} - c_{i,j}}{\varepsilon}\right) \right), \ i \in [\![N]\!]. \end{aligned} \tag{21}$$

(ii) The sequence of vectors (f^t, g^t) satisfies

$$F(f^t, g^t) = \left\langle a, f^t \right\rangle + \left\langle b, g^t \right\rangle, \tag{22}$$

where $F(f, g)$ is the function to maximize in Problem (18)*.*

(iii) Alternatively, a symmetric fixed-point method [23] can be employed: starting from any $(\tilde{f}^0, \tilde{g}^0) \in \mathbb{R}^N \times \mathbb{R}^M$, the following iterations converge to a solution of (18):

$$\begin{aligned}\tilde{g}_j^{t+1} &= \frac{1}{2}\left(\tilde{g}_j^t - \varepsilon \log\left(\sum_{i=1}^N a_i \exp\left(\frac{\tilde{f}_i^t - c_{i,j}}{\varepsilon}\right)\right)\right), \ j \in [\![M]\!], \\ \tilde{f}_i^{t+1} &= \frac{1}{2}\left(\tilde{f}_i^t - \varepsilon \log\left(\sum_{j=1}^M b_j \exp\left(\frac{\tilde{g}_j^t - c_{i,j}}{\varepsilon}\right)\right)\right), \ i \in [\![N]\!].\end{aligned} \tag{23}$$

In our experiments, we adopt the symmetric scheme (23) with the initial values $(\tilde{f}^0, \tilde{g}^0) = (\mathbf{0}_N, \mathbf{0}_M)$ completed with a final single asymmetric step (21) to output an estimate of ROT_ε using (22) that has a linear complexity (while (18) has quadratic complexity). The gradient in (19) is obtained using automatic differentiation by plugging $g^{t\infty}$ into (20), where t_∞ is the index of the last iteration. To accommodate large distributions α and β, we employ the KeOps library [15], which enables the computation of reductions for large matrices with no risk of memory overflows.

3.2 Regularized Semi-Unbalanced Optimal Transport

In this section, we present an entropy-regularized version of SUOT, which we refer to as the Regularized Semi-Unbalanced Optimal Transport (RSUOT) cost. The corresponding proofs are given in Appendix C.

3.2.1 Regularized Semi-Unbalanced Optimal Transport: Primal, Dual Formulation and Gradient

Definition 10 (Regularized Semi-Unbalanced Optimal Transport)**.** *For $\varepsilon > 0$ and $\rho > 0$ fixed, the RSUOT cost between α and β is defined as*

$$\mathrm{RSUOT}_\varepsilon^\rho(\alpha, \beta) = \min_{\pi \in \mathbb{R}_+^{N \times M}} \langle C, \pi \rangle + \varepsilon \mathrm{KL}(\pi | a \otimes b) + \iota_\alpha(\pi_1) + \rho \mathrm{KL}(\pi_2 | b). \tag{24}$$

Similarly to ROT, the addition of εKL to the RSUOT objective function makes it strongly convex. Consequently, by applying the Fenchel-Rockafellar theorem, we can reformulate Problem (24) in a dual form:

$$\mathrm{RSUOT}_\varepsilon^\rho(\alpha, \beta) = \max_{(f,g) \in \mathbb{R}^N \times \mathbb{R}^M} \langle a, f \rangle - \langle b, \phi^\star(-g) \rangle - \varepsilon \left\langle a \otimes b, \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) - 1 \right\rangle. \tag{25}$$

As for ROT, if $f^\star, g^\star$ denote the solutions of Problem (25), we have

$$\nabla_{x_i} \mathrm{RSUOT}_\varepsilon^\rho \left(\sum_{k=1}^N a_k \delta_{x_k}, \beta\right) = a_i \nabla \varphi(x_i), \tag{26}$$

with

$$\varphi(x) = -\varepsilon \log\left(\sum_{j=1}^M b_j \exp\left(\frac{g_j^\star - c(x, y_j)}{\varepsilon}\right)\right). \tag{27}$$

3.2.2 Resolution by Sinkhorn's algorithm

We next present a Sinkhorn-like algorithm for Problem (25).

Theorem 11 (Sinkhorn's algorithm for RSUOT)**.** *(i) Starting from any $f^0 \in \mathbb{R}^N$, the following sequence $(f^t, g^t)_{t\geq 0}$ converges linearly to the unique solution $(f^\star, g^\star)$ of Problem (25):*

$$\begin{aligned} g_j^{t+1} &= -\frac{\varepsilon}{1+\frac{\varepsilon}{\rho}} \log\left(\sum_{i=1}^N a_i \exp\left(\frac{f_i^t - c_{i,j}}{\varepsilon}\right)\right), \ j \in [\![M]\!], \\ f_i^{t+1} &= -\varepsilon \log\left(\sum_{j=1}^M b_j \exp\left(\frac{g_j^{t+1} - c_{i,j}}{\varepsilon}\right)\right), \ i \in [\![N]\!]. \end{aligned} \tag{28}$$

(ii) Denoting $F(f, g)$ the function to be maximized in Problem (25), the sequence of vectors $(f^t, g^t)_{t\geq 0}$ satisfies

$$F(f^t, g^t) = \left\langle a, f^t \right\rangle - \left\langle b, \phi^\star(-g^t) \right\rangle. \tag{29}$$

(iii) Alternatively, the solution vectors $(f^\star, g^\star)$ can be computed by iterating a symmetric fixed-point method [23]: starting from any $(\tilde{f}^0, \tilde{g}^0) \in \mathbb{R}^N \times \mathbb{R}^M$, the following sequence $(\tilde{f}^t, \tilde{g}^t)_{t\geq 0}$ also converges linearly to the solution of (25):

$$\begin{aligned} \tilde{g}_j^{t+1} &= \frac{1}{2}\left(\tilde{g}_j^t - \frac{\varepsilon}{1+\frac{\varepsilon}{\rho}} \log\left(\sum_{i=1}^N a_i \exp\left(\frac{\tilde{f}_i^t - c_{i,j}}{\varepsilon}\right)\right)\right), \ j \in [\![M]\!], \\ \tilde{f}_i^{t+1} &= \frac{1}{2}\left(\tilde{f}_i^t - \varepsilon \log\left(\sum_{j=1}^M b_j \exp\left(\frac{\tilde{g}_j^t - c_{i,j}}{\varepsilon}\right)\right)\right), \ i \in [\![N]\!]. \end{aligned} \tag{30}$$

In our experiments, unlike ROT, we opt for the non-symmetric scheme (28), which, for the values of ρ we use, converges faster than its symmetric counterpart. We initialize it with $f^0 = \mathbf{0}_N$.

3.2.3 Exact Computation of RSUOT in the Simple Case of Section 2.3.3

Here, we revisit the scenario where $\alpha = \delta_x$ and $\beta = b_1\delta_{y_1} + b_2\delta_{y_2}$, with $b_1 + b_2 = 1$ and x, y_1, $y_2 \in \mathbb{R}^n$. Since, in this case, the product measure $\alpha \otimes \beta$ is the only measure with prescribed marginals α and β, we have that

$$\mathrm{ROT}_\varepsilon(\alpha, \beta) = \mathrm{OT}(\alpha, \beta) = b_1\|y_1 - x\|^2 + b_2\|y_2 - x\|^2,$$

where the minimizing transport plan π^*_{ROT} is again (b_1, b_2).

In RSUOT, the functional to be minimized includes the regularization term $\varepsilon\,\mathrm{KL}(\pi|a\otimes b)$. Since a is scalar, this term simplifies to $\varepsilon\,\mathrm{KL}(\pi_2|b)$. Therefore, RSUOT reduces to SUOT, but with a higher level of regularization, as demonstrated by

$$\mathrm{RSUOT}^\rho_\varepsilon(\alpha, \beta) = \mathrm{SUOT}^{\rho+\varepsilon}(\alpha, \beta) = \|y_1 - x\|^2 - (\rho + \varepsilon)\log(b_1 + b_2 e^{-\frac{d}{\rho+\varepsilon}}),$$

where the minimum is attained for $\pi^*_{\mathrm{RSUOT}} = (1 - \eta^*, \eta^*)$, with

$$\eta^* = \frac{b_2 e^{-\frac{d}{\rho+\varepsilon}}}{b_1 + b_2 e^{-\frac{d}{\rho+\varepsilon}}}.$$

As long as α is a single Dirac, increasing the regularization is equivalent to decreasing the semi-unbalancing effect. However, this equivalence does not hold when α has more than one atom.

4 Applications to Inverse Imaging Problems

4.1 The Semi-Unbalanced Wasserstein Patch Prior

We now explain how we use the SUOT cost in the context of inverse imaging problems. For this, we consider a general linear inverse problem with forward model

$$y = \mathcal{A}x^\star + \eta, \tag{31}$$

where $x^\star \in \mathcal{X} = \mathbb{R}^N$, $y \in \mathcal{Y} = \mathbb{R}^K$, $\mathcal{A} \in \mathbb{R}^{K\times N}$, and $\eta \in \mathcal{Y}$ is an error term. The goal is to construct an estimate $\hat{x}(y)$ of $x^\star$. For this, we adopt a variational approach where we solve the optimization problem

$$\min_{x\in\mathcal{X}} \frac{\lambda}{2}\|\mathcal{A}x - y\|^2 + R(x), \tag{32}$$

for $\lambda > 0$ and a regularizer R. Other fidelity terms can be used in (32) instead of the squared Euclidean distance.

Besides y, we assume that we also have access to a reference image $x_{\text{ref}} \in \mathbb{R}^M$ related to $x^\star$. The first row of Figure 3 shows an example where the first column is the reference x_{ref}, the second one is a simulated noisy version y of the ground-truth $x^\star$ appearing in the third column. As can be seen, x_{ref} is related to $x^\star$ in the sense that its statistics are similar to those of $x^\star$. Pixelwise values of x_{ref} and $x^\star$ are nonetheless not expected to be close.

In this context, Hertrich *et al.* [37] have introduced the use of the Wasserstein patch prior (WPP) $R_0(x) = \text{OT}(\alpha_x, \beta_{x_{\text{ref}}})$, where $\alpha_x = \frac{1}{N}\sum_{i=1}^N \delta_{P_i x}$, $\beta_{x_{\text{ref}}} = \frac{1}{M}\sum_{j=1}^M \delta_{P_j x_{\text{ref}}}$, P_k being the operator that extracts the k^{th} patch from a given image. To simplify notation, we have implicitly assumed that the number of patches extracted from each image is equal to the number of its pixels, however, in practice, there are less patches than pixels due to border considerations. Solving Problem (32) with $R = R_0$ thus favors images whose patch distribution is close to the one of the reference x_{ref} in the sense of OT, while being consistent with the observation y.

The regularizer R_0 involves patches of a fixed given size. In order to capture and incorporate both fine and coarse details, [37] also proposes the regularizer

$$R_L(x) = \frac{1}{L+1}\sum_{\ell=0}^{L} \text{OT}(\alpha_{x^\ell}, \beta_{x^\ell_{\text{ref}}}), \tag{33}$$

where $x^\ell = A^\ell x$, $x^\ell_{\text{ref}} = A^\ell x_{\text{ref}}$, A represents a downsampling operator, A^0 is the identity and $A^{\ell+1} = AA^\ell$. In practice, a convolution with a Gaussian blur kernel of size 4×4 and standard deviation of 1, followed by a subsampling of factor 2 is used for A. The empirical distributions α_{x^ℓ} and $\beta_{x^\ell_{\text{ref}}}$ are given by: $\alpha_{x^\ell} = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \delta_{P_i x^\ell}$ and $\beta_{x^\ell_{\text{ref}}} = \frac{1}{M_\ell}\sum_{j=1}^{M_\ell} \delta_{P_j x^\ell_{\text{ref}}}$. It should be noted that the size of the patches extracted by the operators P_k does not depend on the scale level ℓ. Consistent with our notation, when $L = 0$, the multiscale WPP reverts to the single-scale one.

4.2 A First Illustrative Example: Denoising Using Single Scale WPP

In this subsection, our goal is to showcase the effectiveness of the single-scale regularizer R_0 in a denoising example, illustrating the potential improvements in image quality

through the utilization of SUOT over OT. Specifically, we consider the scenario where $\mathcal{A}$ is the identity operator and $\eta \sim \mathcal{N}(0, (30/256)^2\mathrm{I})$ is a realization from a Gaussian random variable. We utilize patches of size 6×6. All patches are extracted from the reference and noisy images.

The optimization problem (32) with $R(x) = R_0(x)$ is solved with 500 (outer) iterations of gradient descent, employing the Adam optimizer [43]. In each outer iteration, OT is approximated using 10 (inner) iterations of AGAA on the semi-dual formulation (see (6)). We warm start the semi-dual variable, that is, the last semi-dual variable in inner iteration p is used to initialize the first one in iteration $p + 1$. We later address ROT, SUOT and RSUOT variants in a similar manner, approximating the transport using 10 iterations of AGAA in the case of SUOT, and 10 iterations of Sinkhorn in the case of ROT and RSUOT. We assess the denoising quality by computing the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM) [73], as well as the Learned Perceptual Image Patch Similarity (LPIPS) [74], each time considering a central crop of 6 pixels away from the image edges. We obtained the result labeled as OT in Figure 3.

When minimizing (32) with $R = R_0$, the tendency of OT to preserve the frequencies of the patches of x_{ref} is balanced by the presence of the data fidelity term $\frac{\lambda}{2}\|\mathcal{A} \cdot -y\|^2$. It is however interesting to see if the relaxation provided by SUOT, allowing an explicit control of those frequencies, can improve the denoising results. In particular, we expect the use of SUOT to enable discarding some patches in x_{ref} not suitable for estimating $x^\star$. We thus introduce the semi-unbalanced Wasserstein patch prior $\tilde{R}_0(x) = \mathrm{SUOT}^\rho(\alpha_x, \beta_{x_{\mathrm{ref}}})$, $\rho > 0$, and its multiscale extension $\tilde{R}_L$. With $\rho = 0.01$, the result labeled as SUOT in Figure 3 improves upon the one obtained with OT in terms of all considered metrics.

In Figure 3, we also show the result obtained when minimizing (32) with $R(x)$ being equal to the entropy-regularized fully unbalanced OT cost of [17, 25], hereafter denoted $\mathrm{RUOT}^\rho_\varepsilon(\alpha_x, \beta_{x_{\mathrm{ref}}})$. The strength of entropy regularization is dictated by the value of ε. The result labeled as RUOT in Figure 3 corresponds to $\varepsilon = 10^{-4}$, making RUOT a good approximation of its unregularized counterpart UOT, proposed in [11, 46]. The latter differs from SUOT in that the constraints on *both* marginals of the OT plan are relaxed. As can be expected, when employing $\mathrm{RUOT}^\rho_\varepsilon$ as a regularizer, some patches might not evolve trough the gradient descent iterations. This is confirmed in Figure 3 where, after initializing the ADAM optimizer with the noisy image y, we see that some regions in the image labeled RUOT have not been denoised. In comparison, our proposed SUOT-based result ensures spatially homogeneous denoising and achieves the best performance. To summarize, SUOT is a robust version of OT, that denoises all the patches while allowing for a partial match to the reference distribution, the proportion of matching being tuned by the parameter ρ.

Figure 4 investigates how denoising performance is affected by replacing OT and SUOT by their entropy-regularized counterparts ROT and RSUOT. Both ROT and RSUOT are approximated using Sinkhorn's algorithms given in Theorems 9 and 11. It is observed that as ε increases, the quality of denoising decreases, particularly in terms of LPIPS score. This decline is related to the bias discussed in Section 3.1. Nevertheless, with small values of ε, ROT and RSUOT enable reasonable distribution comparisons.

4.3 Reference-Driven Super-Resolution using Multiscale WPP

We concentrate in this subsection on the problem of super-resolution (SR) with the presence of a reference, as considered in [37]. The forward model is given by (31), with

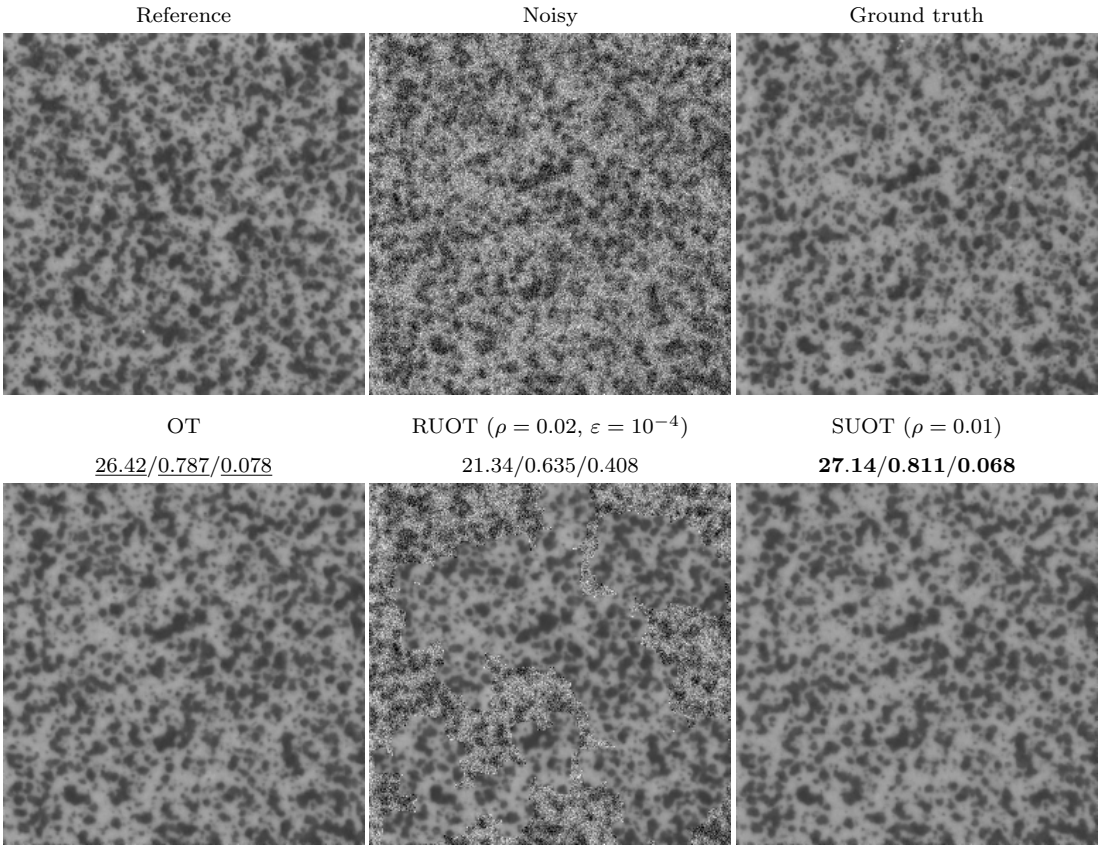

Figure 3: **Application of OT, SUOT and RUOT to denoising**. The top row displays the reference image (Reference), the noisy image (Noisy), and the original image (Ground truth). The bottom row presents denoising results using OT, RUOT, and SUOT with $\lambda = 0.03$. For each result, a triplet of metrics PSNR/SSIM/LPIPS is provided, with the best score highlighted in bold and the second-best underlined. The computation time is 325 seconds for OT, 326 seconds for SUOT, and 725 seconds for RUOT.

$\mathcal{A} = S \circ H$, where H is a low-pass convolution operator, S a downsampling operator and $\eta \sim \mathcal{N}(0, 0.01^2)$. The observation y is thus a low-resolution (LR) version of the high-resolution (HR) ground-truth $x^\star$. The reference x_{ref} is in this context a HR image, whose statistics are similar to those of $x^\star$. While pixelwise differences between $x^\star$ and x_{ref} are not expected to be close to zero, its is nonetheless assumed that the distribution of patches extracted from x_{ref} can serve as a prior for estimating $x^\star$. The reference-driven SR scenario emerges whenever time and/or system resources are allocated to acquire high-quality images, which are subsequently used to improve related lower-quality acquisitions. This scope can be further expanded by using higher-level features instead of patches, in particular if these features are robust to image transformations such as changes in brightness, color, contrast, scale, and rotation.

The authors of [37] propose to use their multiscale WPP prior R_L, defined in (33), as a regularizer in (32). This leads to an estimate $\hat{x}(y)$ whose multiscale patch distribution is close to the one of x_{ref} in the sense of OT, while maintaining pixel values consistent with y. The balance between these two objectives is determined by λ.

In practice, in order to deal with the previously discussed issues encountered when using OT as a cost, the authors of [37] formulate the optimization problem (32) in terms of an unknown image z obtained by artificially expanding the boundaries of x. The

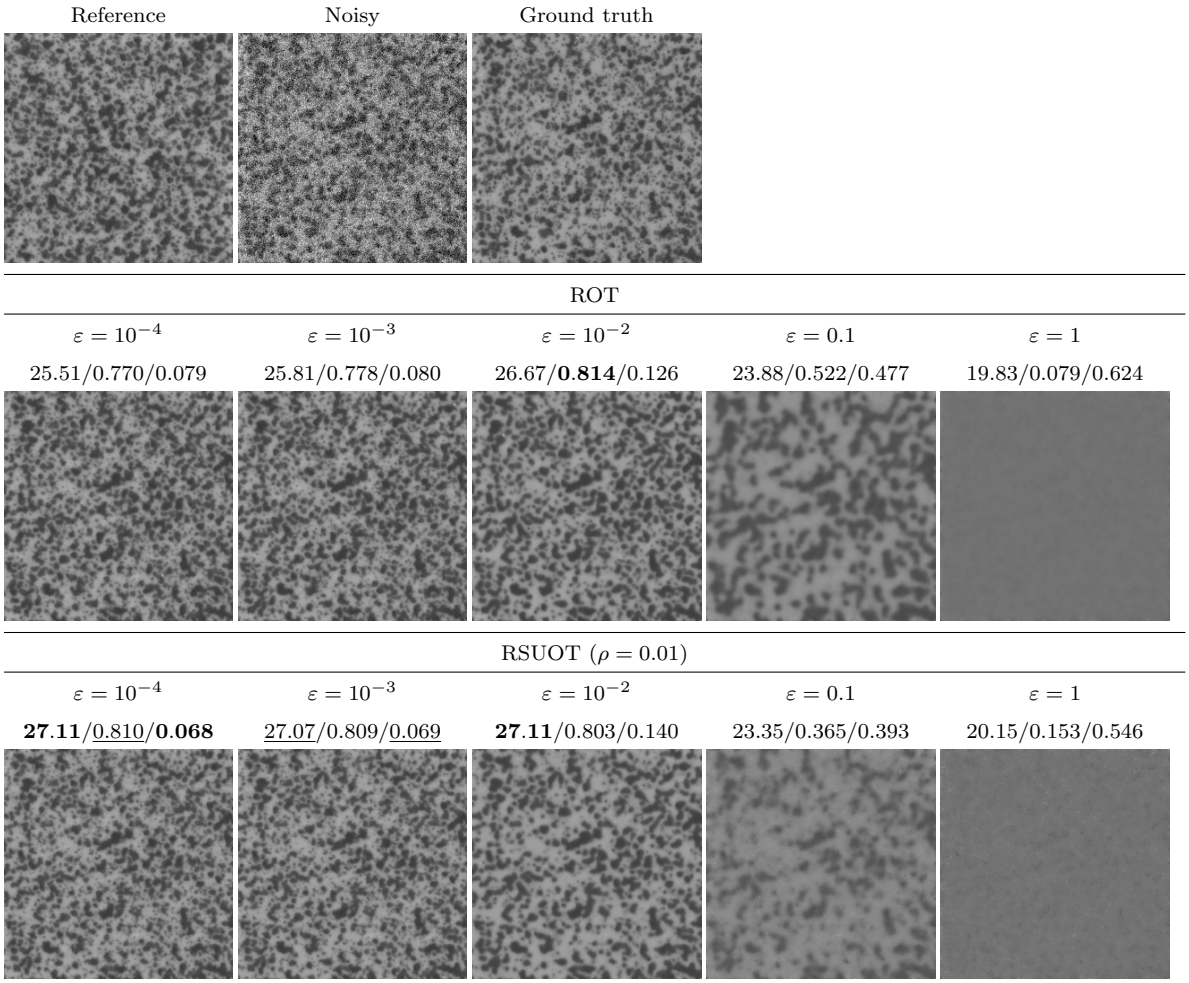

Figure 4: **Impact of the regularization on a denoising example**. The top row displays the original image (Ground truth), the noisy image (Noisy), and the reference image (Reference). The bottom rows present denoising results employing ROT and RSUOT for different values of ε and $\lambda = 0.03$. Each result is accompanied by a triplet of metrics PSNR/SSIM/LPIPS. The best score is denoted in bold, and the second-best is underlined. The computation time for each result was 725 seconds.

fidelity term is applied to the central part of z, leading to the problem

$$\min_z \frac{\lambda}{2}\|SHCz - y\|^2 + \frac{1}{L+1}\sum_{\ell=0}^{L} \mathrm{OT}(\alpha_{z^\ell}, \beta_{x^\ell_{\mathrm{ref}}}), \tag{34}$$

where C is the operator that discards the artificial boundaries. In this way, patches from x_{ref} that are unsuitable for super-resolving y are expected to accumulate within the artificial boundaries of z instead of being concentrated at its center.

Instead of dealing with the rigidity of OT by artificially expanding the boundaries of the unknown, we propose to use our SUOT and RSUOT regularizers, as they intrinsically allow for partial mass displacement. We thus consider Problem (32) with $\mathcal{A} = S \circ H$ and $R(x) = \frac{1}{L+1}\sum_{\ell=0}^{L} \mathrm{SUOT}^\rho(\alpha_{x^\ell}, \beta_{x^\ell_{\mathrm{ref}}})$ or $R(x) = \frac{1}{L+1}\sum_{\ell=0}^{L} \mathrm{RSUOT}^\rho_\epsilon(\alpha_{x^\ell}, \beta_{x^\ell_{\mathrm{ref}}})$.

In our experiments, we use images sourced from the MVTec database [8, 9], resized to 256×256. This database contains images of different object and texture categories, such as "Tile" and "Wood" (see Columns 2 and 3 in Figure 5 for grayscale versions, and

in Figure 6 for the corresponding color versions). Each category contains images with and without anomalies. Considering Column 2 in Figure 5, the topmost row shows a "Tile" without anomaly, while the next row shows a "Tile" presenting an anomaly. The subsequent rows show two other categories, without and with anomalies.

From the MVTec database, we constructed two sub-datasets, the "anomaly-free" and "with anomalies" datasets. The "anomaly-free" dataset consists of eighteen pairs of images $(y_i, x_{\mathrm{ref},i})$, where each y_i is a LR *anomaly-free* image and $x_{\mathrm{ref},i}$ a HR *anomaly-free* image from the same category as y_i. The "with anomalies" dataset follows a similar structure except that now, the references are chosen among the HR images of the same category but *with anomalies*. In both datasets, the LR images y_i have been simulated by convolving the corresponding original HR images $x_i^\star$ with a Gaussian kernel of size 16×16 and standard deviation equal to 2. A subsampling of factor 4 along each spatial dimension has been applied and Gaussian white noise of standard deviation 0.01 added. The two sub-datasets allow to compare the relative performance of OT and SUOT in two different regimes. With the "anomaly-free" dataset, the patch distribution of each reference is expected to differ only marginally from the one of the corresponding ground-truth. In the "with anomalies" dataset, a limited portion of patches from each reference correspond to outliers, that is, they are very different from the typical patches found in the corresponding ground-truth.

To super-resolve each LR image y_i, we used the associated reference $x_{\mathrm{ref},i}$ as a prior and set the parameters as follows: $\lambda = 0.006$, $\rho = 0.01$, $\varepsilon = 10^{-4}$, $L = 1$ (2 scales), and a patch size of 6. All patches were extracted from x (or z) and 10000 patches were extracted, once and for all, from each reference image. Similar to [37], we solve Problem (34) with 500 (outer) iterations of gradient descent using the Adam optimizer. In each outer iteration, OT is approximated with 10 iterations of AGAA. The semi-dual variables are warm-started as explained in the previous subsection. We address our SUOT and RSUOT variants in a similar manner, approximating SUOT using AGAA and RSUOT using 10 iterations of Sinkhorn's algorithm presented in Theorem 11.

We evaluate PSNR, SSIM and LPIPS scores for each dataset, each time considering a central crop of 6 pixels away from the image edges. The results are presented in Table 1 and Figure 5. In Figure 5, each LR image is associated with two rows: the first row corresponds to the result obtained using an anomaly-free reference HR image, while the second row corresponds to the one obtained using a reference HR image with anomalies. The results shown in Table 1 and Figure 5 demonstrate that the improved robustness achieved by replacing OT with SUOT or RSUOT leads to higher PSNR values when the references lack anomalies. Furthermore, it leads to significantly higher PSNR and LPIPS scores when the references present outliers.

Table 2 and Figure 6 lead to the same conclusions in the case of color images. Considering the computational time, we opted to solely compare with SUOT.

4.4 Neural Network Amortization

We consider in this subsection a neural network amortization [2] of Problem (32) with $\mathcal{A} = S \circ H$ and $R = \mathrm{SUOT}$. In this context, amortization amounts to training a neural network G_θ, whose input is a LR image y, and whose output should be close to the estimate $\hat{x}(y)$ obtained as a solution of Problem (32). The amortization of Problem (32) with the WPP regularizer $R = R_L$ has been developed in [1]. Therein, the authors

Table 1: **Average PSNR, SSIM and LPIPS obtained on two datasets of 18 grayscale image pairs from the MVTec image database [8, 9]**. Reference images from the first dataset are anomaly-free while those in the second dataset contain anomalies (see Figure 5). The average running time is provided.

	Anomaly-free			With anomalies			
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Runtime
WPP [37]	30.11	0.651	0.094	30.15	0.670	0.099	<u>109s</u>
ROT	29.81	**0.659**	**0.088**	29.74	**0.682**	0.101	230s
RSUOT (Ours)	**30.48**	<u>0.655</u>	0.094	**30.90**	<u>0.676</u>	**0.092**	230s
SUOT (Ours)	<u>30.43</u>	0.653	<u>0.092</u>	<u>30.80</u>	0.672	<u>0.093</u>	**73s**

Table 2: **Average PSNR, SSIM and LPIPS obtained on two datasets of 18 color image pairs from the MVTec image database [8, 9]**. Reference images from the first dataset are anomaly-free while those in the second dataset contain anomalies like in Figure 5.

	Anomaly-free			With anomalies			
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Runtime
WPP [37]	29.75	0.660	0.135	30.01	0.680	0.131	161s
SUOT (Ours)	**30.37**	**0.676**	**0.131**	**30.83**	**0.701**	**0.128**	**136s**

propose to minimize the following loss:

$$\min_{\theta} \frac{1}{N_B} \sum_{j=1}^{N_B} \frac{1}{b} \sum_{i \in B_j} \|f(G_\theta(y_i)) - y_i\|^2 + \lambda \, \mathrm{OT}\Big(\frac{1}{b} \sum_{i \in B_j} \alpha_{G_\theta(y_i)}, \beta_{x_{\mathrm{ref}}}\Big), \quad \lambda > 0, \tag{35}$$

referred to as the WPPNets loss. In order to augment the number of training examples, all training LR images y_i in (35) are smaller in size compared to the reference image. They are thus expected to have a patch distribution close to *a subset* of the one of x_{ref}. Once trained, G_θ is used to super-resolve LR images whose patch distribution is supposed to be close to the *specific* reference used during training. To put it another way, retraining G_θ is necessary as soon as the reference changes. In (35), the division of training images into batches $B = (B_j)_{j \in N_B}$ aims at alleviating the rigidity induced by using OT as a cost. This rigidity is especially pronounced here, since LR images may present only a small part of the considered object or texture. With this strategy, the aggregate patch distribution inside a batch closely matches the one of the reference in the sense of OT, while still allowing for variation *within* the batch. The architecture for G_θ proposed in [1] is adapted from the fully convolutional CNN proposed in [69].

To further augment the flexibility of the approach, we suggest replacing OT in Prob-

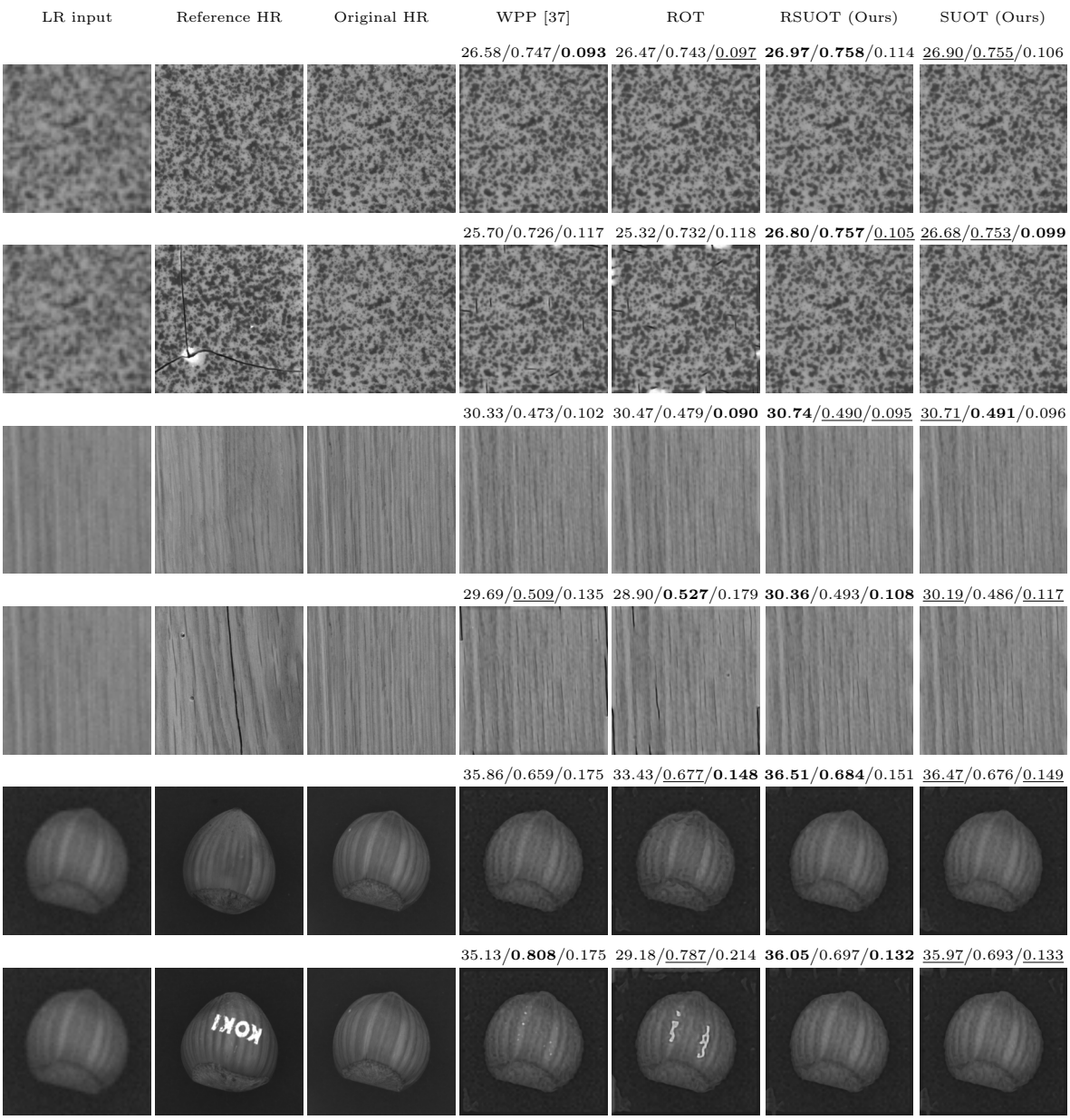

Figure 5: **Application to super-resolution with a reference image**. From left to right: LR input, reference HR input, original HR image, sSR by WPP [37], ROT, RSUOT and SUOT priors. For each result we provide the triplet PSNR/SSIM/LPIPS.

lem (35) with our SUOT variant. We denote this modified problem as WPPNetsSU:

$$\min_{\theta} \frac{1}{N_B} \sum_{j=1}^{N_B} \frac{1}{b} \sum_{i \in B_j} \|f(G_\theta(y_i)) - y_i\|^2 + \lambda \, \mathrm{SUOT}^\rho \left(\frac{1}{b} \sum_{i \in B_j} \alpha_{G_\theta(y_i)}, \beta_{x_{\mathrm{ref}}} \right), \quad \lambda > 0. \tag{36}$$

The authors of [1] solve Problem (35) with the Adam optimizer. To approximate OT, they utilize AGAA with 20 iterations per step. We address Problem (36) with the implementation provided by the authors, substituting the computation of OT with that of SUOT.

In our experiments, we focused on images belonging to the "Wood" and "Tile" classes of the MVTec dataset [8, 9]. These images underwent grayscale conversion and were

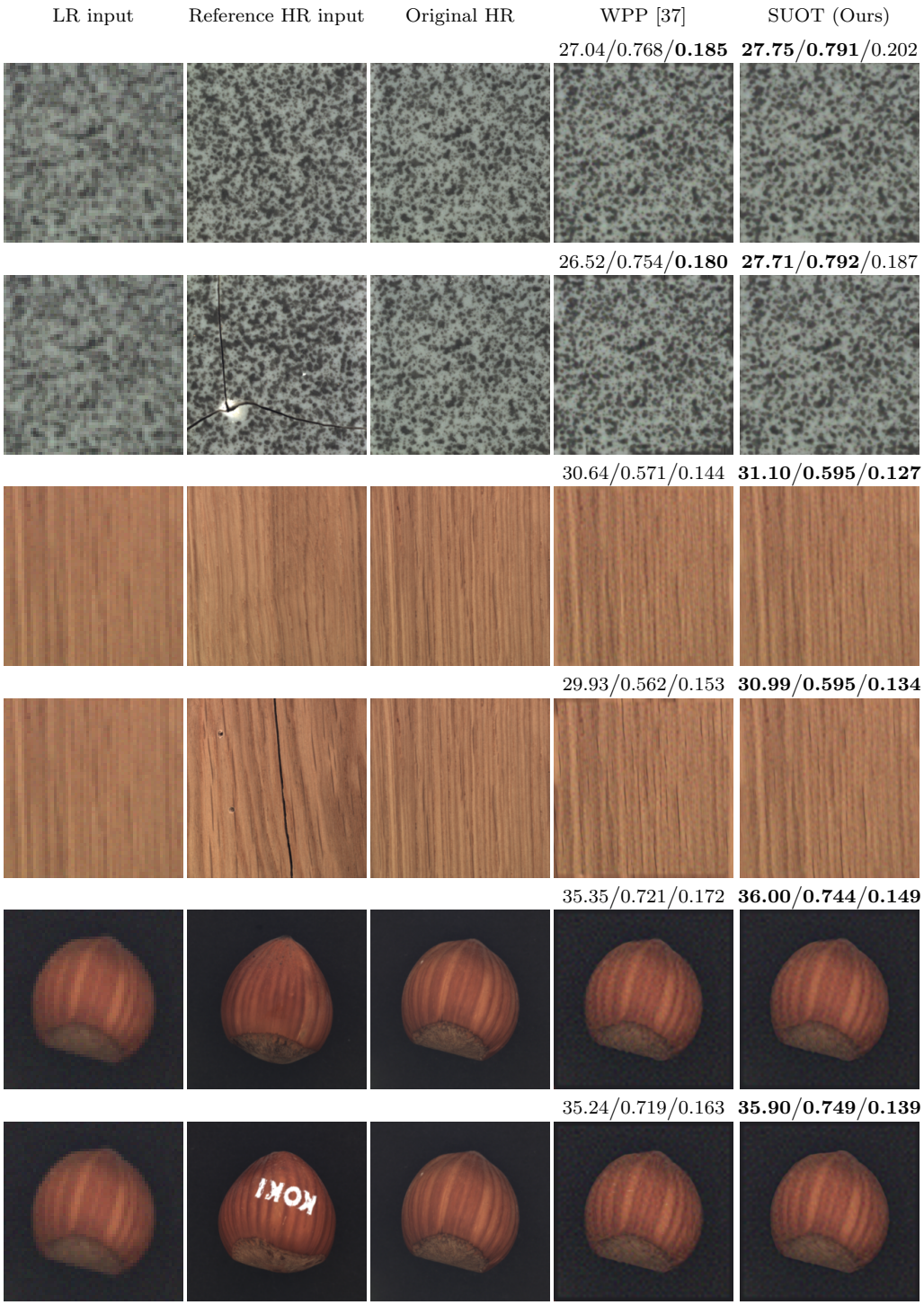

Figure 6: **Application to color super-resolution with a reference image**. From left to right: LR input, reference HR input, original HR image, SR by WPP [37] and SUOT priors. For each result we provide the triplet PSNR/SSIM/LPIPS.

resized to a resolution of 600×600 pixels. We then extracted six 100×100 HR crops from the training sets associated with each category. We simulated the LR version of each 100×100 HR image by applying a Gaussian blur kernel of size 16×16 and standard deviation of 2. We then applied a subsampling factor of 4 in each spatial dimension, and added Gaussian noise $\eta \sim \mathcal{N}(0, 0.01^2)$. We selected an anomaly-free reference image from each category.

Following [1], we choose $b = 25$, $\lambda = 12.5$, a patch size of 6, and opted for $\rho = 0.01$. Table 3 presents the average PSNR, SSIM, and LPIPS values for WPPNets and WPPNetsSU, calculated over the entire test dataset in the "Wood" and "Tile" classes. The dataset consists of 18 samples for the "Wood" class and 32 samples for the"Tile" class. It is worth noting that while the overall results appear comparable, a noticeable improvement in PSNR is observed specifically for the "Wood" class, which exhibits a more diverse texture compared to the "Tile" class. It demonstrates the capability of SUOT to focus on regions of the reference distribution that align with the ground-truth. We present some results from Table 3 in Figure 7. Furthermore, we generated Table 4 and Figure 8 for color images, observing consistent conclusions with those drawn for grayscale images. The neural network training time was 2.5 hours for grayscale images and 5 hours for color images using an Nvidia A100 GPU. After training, each image is processed with an average time of 0.013 seconds.

Table 3: **Average PSNR, SSIM and LPIPS scores for the SR of two grayscale image sets: "Wood" and "Tile" from [8, 9]**. The network was trained using the reference images shown in Figure 7. The "Wood" dataset consists of 18 images, while the "Tile" dataset contains 32 images. From top to bottom: SR by WPPNets [1], WPPNetsSU.

	Wood			Tile		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
WPPNets [37]	31.99	0.5307	**0.1804**	**32.88**	0.8504	**0.2257**
WPPNetsSU (Ours)	**32.33**	**0.5494**	0.1896	32.83	**0.8514**	0.2358

Table 4: **Average PSNR, SSIM, and LPIPS scores for the SR of two sets of color images:"Wood" and "Tile" sourced from [8, 9]**. The model was trained using the reference images shown in Figure 8. The "Wood" dataset consists of 18 images, while the "Tile" dataset contains 32 images. From top to bottom: SR by WPPNets [1], WPPNetsSU.

	Wood			Tile		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
WPPNets [37]	32.07	0.6651	0.1743	33.83	0.8747	**0.2282**
WPPNetsSD (Ours)	**32.88**	**0.6758**	**0.1694**	**33.93**	**0.8769**	0.2646

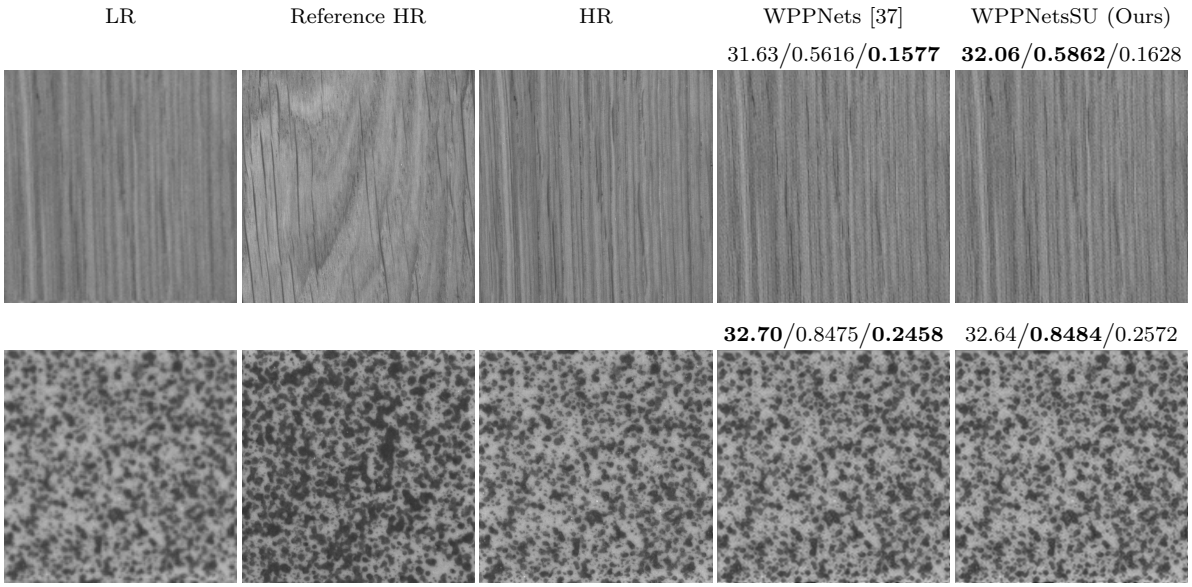

Figure 7: **Application of WPPNets [1] and WPPNetsSU to a grayscale "Wood" and "Tile" image from [8,9].** Images are arranged from left to right: Low-resolution image (LR), reference image (Reference HR), high resolution image (HR), SR using WPPNets [1], and SR using WPPNetsSU. For each result, a triplet of metrics PSNR/SSIM/LPIPS is provided.

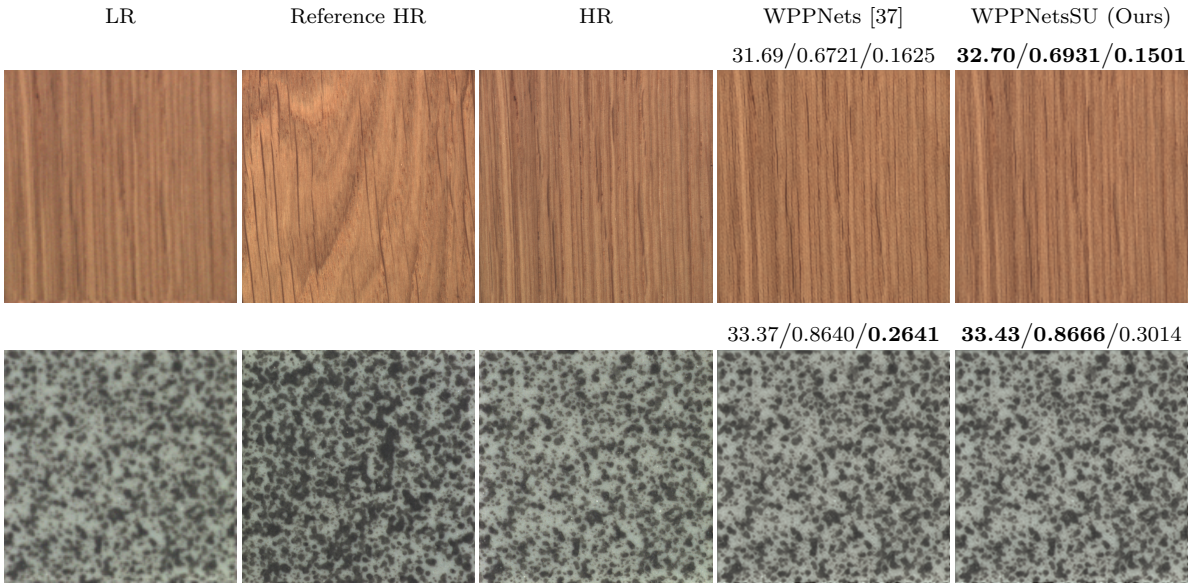

Figure 8: **Application of WPPNets [1] and WPPNetsSU to a color "Wood" and "Tile" image from [8, 9].** Images are presented from left to right: Low-resolution image (LR), reference high-resolution image (Reference HR), high-resolution image (HR), SR using WPPNets [1], and SR using WPPNetsSU. For each result, a triplet of metrics PSNR/SSIM/LPIPS is provided.

5 Single Image Generation using SUOT

5.1 PSinOT and its Semi-Unbalanced Version

Let us turn to another task for which OT between image patches is particularly adapted, namely single-image generation (SIG). The goal of SIG is to generate realistic images resembling a given reference image x_{ref}. The Patch-based algorithm for Single image generation with OT (PSinOT), introduced in [16], addresses this challenge by implementing OT projection and Nearest Neighbor Matching (NNM) at various scales. Namely, the reference image undergoes L successive downscalings using an operator A, which combines a convolution with a 3×3 Gaussian kernel of standard deviation 1.5 and a downsampling operator of factor 2. This results in the image sequence $\left(x_{\text{ref}}^{\ell}\right)_{\ell \in \{0,\dots,L\}}$, where $x_{\text{ref}}^{0} = x_{\text{ref}}$ and $x_{\text{ref}}^{\ell+1} = A x_{\text{ref}}^{\ell}$. PSinOT starts by synthesizing the image at the coarsest scale L. At each scale ℓ, the generated image $\hat{x}^{\ell}$, after being scaled up, serves as initialization for the previous finer scale $\ell - 1$, until the initial resolution is reached. To synthesize the image at some scale ℓ, PSinOT works either with OT or NNM. OT is chosen at coarse scales, in order to establish the overall structure of the image, while at finer scales, NNM is preferred, in order to enhance the details of the generated image while preserving the structure. At a given scale ℓ, synthesizing an image using OT is done by solving (with gradient descent, starting from a given initialization)

$$\min_{x} \mathrm{OT}(\alpha_x, \beta_{x_{\text{ref}}^{\ell}}), \tag{37}$$

where α_x is the empirical distribution of all patches extracted from x, and $\beta_{x_{\text{ref}}^{\ell}}$ the empirical distribution of all patches extracted from x_{ref}^{ℓ}. NNM-based synthesis at scale ℓ works by solving

$$\mathrm{NNM}(x_{\text{ref}}^{\ell}) = \arg\min_{x} \sum_{i=1}^{N} \min_{j \in [\![M]\!]} \|P_i x - P_j x_{\text{ref}}^{\ell}\|_2^2, \tag{38}$$

where P_k is the operator that extracts the k^{th} patch. The NNM problem is efficiently solved by means of the PatchMatch algorithm [5].

In order to give a clear algorithmic description of the process, we introduce the functions $\mathrm{solveOT}(x_{\text{init}}, x_{\text{ref}})$ and $\mathrm{solveNNM}(x_{\text{init}}, x_{\text{ref}})$ that respectively solve Problems (37) and (38), starting from an initial image x_{init}. Regardless of the function used, the output $\hat{x}^{\ell}$ is upscaled with a factor of 2 by determining the Nearest Neighbor Assignment (NNA) between the patches of $\hat{x}^{\ell}$ and those of x_{ref}^{ℓ}, and then by interpolating the NNA map [16]. The resulting image is used as the initialization at scale $\ell - 1$. The initialization at the coarsest scale is a Gaussian noise image. PSinOT uses 4 scales; at coarse scales ($\ell \in \{2, 3\}$), OT is used, whereas NNM is used at fine scales $\ell \in \{0, 1\}$. The procedure is detailed in Algorithm 1.

When the number of patches in the synthesized image equals the one in the reference image, solving (37) leads to a unique matching of each patch in the synthesized image with a single, distinct patch from the reference. This results in a one-to-one correspondence, which often makes the synthesized image look very similar to the reference.

To increase the variety in the synthesized images, we propose to replace OT with SUOT in (37). This transforms the optimization problem (37) into

$$\min_{x} \mathrm{SUOT}^{\rho}(\alpha_x, \beta_{x_{\text{ref}}^{\ell}}), \tag{39}$$

Algorithm 1 PSinOT/PSinSUOT

Input: z (initialization image), $\left(x_{\text{ref}}^{\ell}\right)_{\ell\in\{0,\dots,3\}}$ (downscaled versions of x_{ref}), ρ (the unbalance parameter, for the SUOT variant)
Output: $\hat{x}^0$ (the synthesized image at the initial resolution)
for $\ell = 3$ **to** 0 **do**
  if $\ell == 3$ **then**
    $x_{\text{init}}^{\ell} \leftarrow z$
  else
    $x_{\text{init}}^{\ell} \leftarrow \text{upscale}(\hat{x}^{\ell+1})$
  end if
  if $\ell \in \{2, 3\}$ **then**
    $\hat{x}^{\ell} \leftarrow \text{solveOT}(x_{\text{init}}^{\ell}, x_{\text{ref}}^{\ell})$
  else
    $\hat{x}^{\ell} \leftarrow \text{solveNNM}(x_{\text{init}}^{\ell}, x_{\text{ref}}^{\ell})$
  end if
end for
return $\hat{x}^0$

and allows multiple patches in the synthesized image to evolve to the same closest reference patch. We show below that this flexibility leads to variations in the synthesized images, enhancing their diversity. Our variant of PSinOT using SUOT, denoted as PSinSUOT, is also described by Algorithm 1, considering that the function solveOT now performs a gradient descent on (39) instead of (37).

5.2 Numerical Results

In our experiments, we utilized reference images from the
Places50 and SIGD16 datasets [32]. To solve Problem (37) and Problem (39), we used, as in Houdard et al. [40], 1000 iterations of the Adam gradient descent optimizer. To approximate OT and SUOT, we utilized AGAA with 10 iterations per step. An exception was made for the initial step, for which we allowed 10000 iterations for improved convergence. Following the procedure detailed in [16], we initialized the image with white Gaussian noise sampled from $\mathcal{N}(0.5, 1)$ at the coarsest scale and chose to work with patches of size 11×11.

In Table 5, we present the SIFID and Diversity scores averaged on 50 generated images on the three reference images displayed in Figure 9 using PSinOT and PSinSUOT with $\rho = 0.01$. As explained in [63], SIFID is a single-image version of the Fréchet Inception Distance (FID) [38]. Diversity is defined as the per-pixel standard deviation calculated from 50 generated images, averaged over all pixels, and normalized by the standard deviation of the reference image. Consequently, a low SIFID suggests that the synthesized image shares a similar distribution of features with the reference image, indicating similar content. Conversely, a high Diversity score indicates a wide range of possible generated images. As expected, PSinSUOT demonstrates higher SIFID and Diversity scores compared to PSinOT, indicating a more varied set of elements and a greater diversity in the generated images. Additionally, in Figure 9, we showcase six image synthesis results for each reference image, visually emphasizing the diversity of images obtained with SUOT compared to OT. Finally, in Figure 10, we explore the

impact of the unbalance parameter ρ on the images generated with PSinSUOT using various initialization images. We observe that the choice of the initialization image has a significant influence on the generated image with PSinSUOT, whereas PSinOT is less affected by this change. Moreover, the parameter ρ enables interpolating between NNA and OT, significantly influencing the final image. Higher values of ρ shift the result closer to OT, while lower values tilt it towards NNA. Indeed, when $\rho \to 0$, the penalty on π_2 is relaxed, leading to a solution that corresponds to NNA.

Table 5: **SIFID and Diversity scores obtained from 50 generated images with PSinOT and PSinSUOT on three reference images depicted in Figure 9**. SUOT achieves a higher diversity score compared to OT in the generated images.

	Reference 1		Reference 2		Reference 3	
	SIFID↑	Diversity↑	SIFID↑	Diversity↑	SIFID↑	Diversity↑
PSinOT [16]	1.0e-5	0.75	3.5e-6	0.40	1.0e-5	0.62
PSinSUOT (Ours)	**7.9e-5**	**0.77**	**9.3e-6**	**0.51**	**3.4e-5**	**0.78**

6 Conclusion

In this paper, we introduced SUOT, an asymmetric form of OT targeted towards inverse imaging and synthesis problems. SUOT is explicitly designed to alleviate the issues encountered when using OT as a cost in imaging problems. We studied both the unregularized and entropy-regularized version of SUOT, deriving dual formulations, corresponding minimization algorithms and formulas for the gradient of the cost. Those derivations allowed us to use a gradient descent scheme either to minimize an energy, or to learn a neural network, both involving a SUOT cost. We evaluated our proposal for a reference-driven SR problem and showed its benefits. We also incorporated SUOT into a state-of-the-art single-image generation algorithm and showed that it leads to increased diversity, as measured by SIFID and per-pixel standard deviation metrics. Future work will focus on achieving fine-grained control over the proportion of the target distribution considered by SUOT and exploring other relevant image processing problems.

References

[1] Fabian Altekrüger and Johannes Hertrich. WPPNets and WPPFlows: The power of Wasserstein patch priors for superresolution. *SIAM Journal on Imaging Sciences*, 16(3):1033–1067, 2023.

[2] Brandon Amos. Tutorial on amortized optimization for learning to optimize over continuous domains. *arXiv preprint arXiv:2202.00665*, 2022.

[3] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

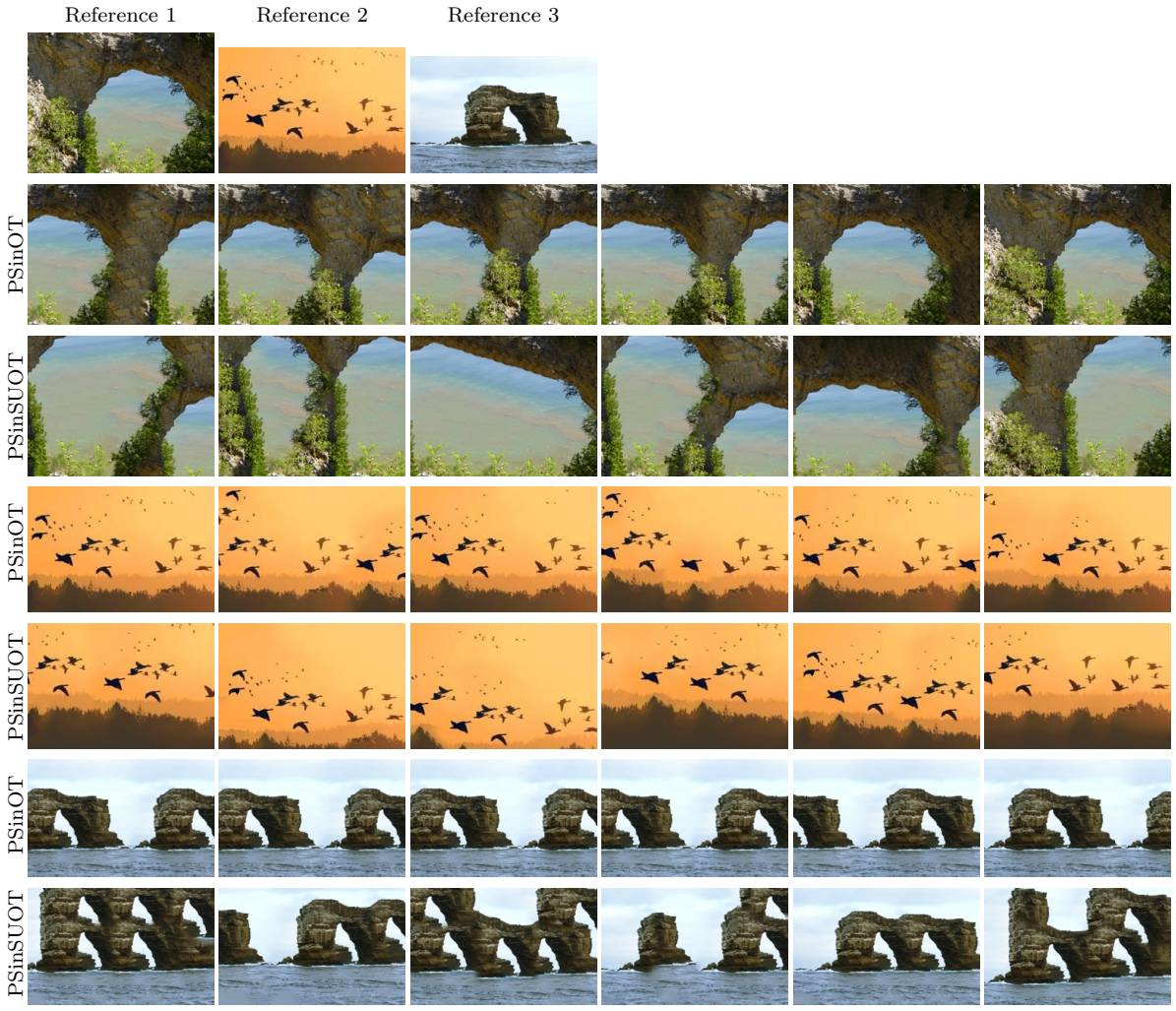

Figure 9: **Selected results showcasing the diversity of images obtained from three reference images using PSinOT and PSinSUOT**. From top to bottom: Reference images, image synthesis using PSinOT or PSinSUOT. For each column, images obtained with PSinOT and PSinSUOT were generated using the same initialization $z \sim \mathcal{N}(0.5, 1)$.

[4] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

[5] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

[6] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76(12):1298–1302, 2006.

[7] Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

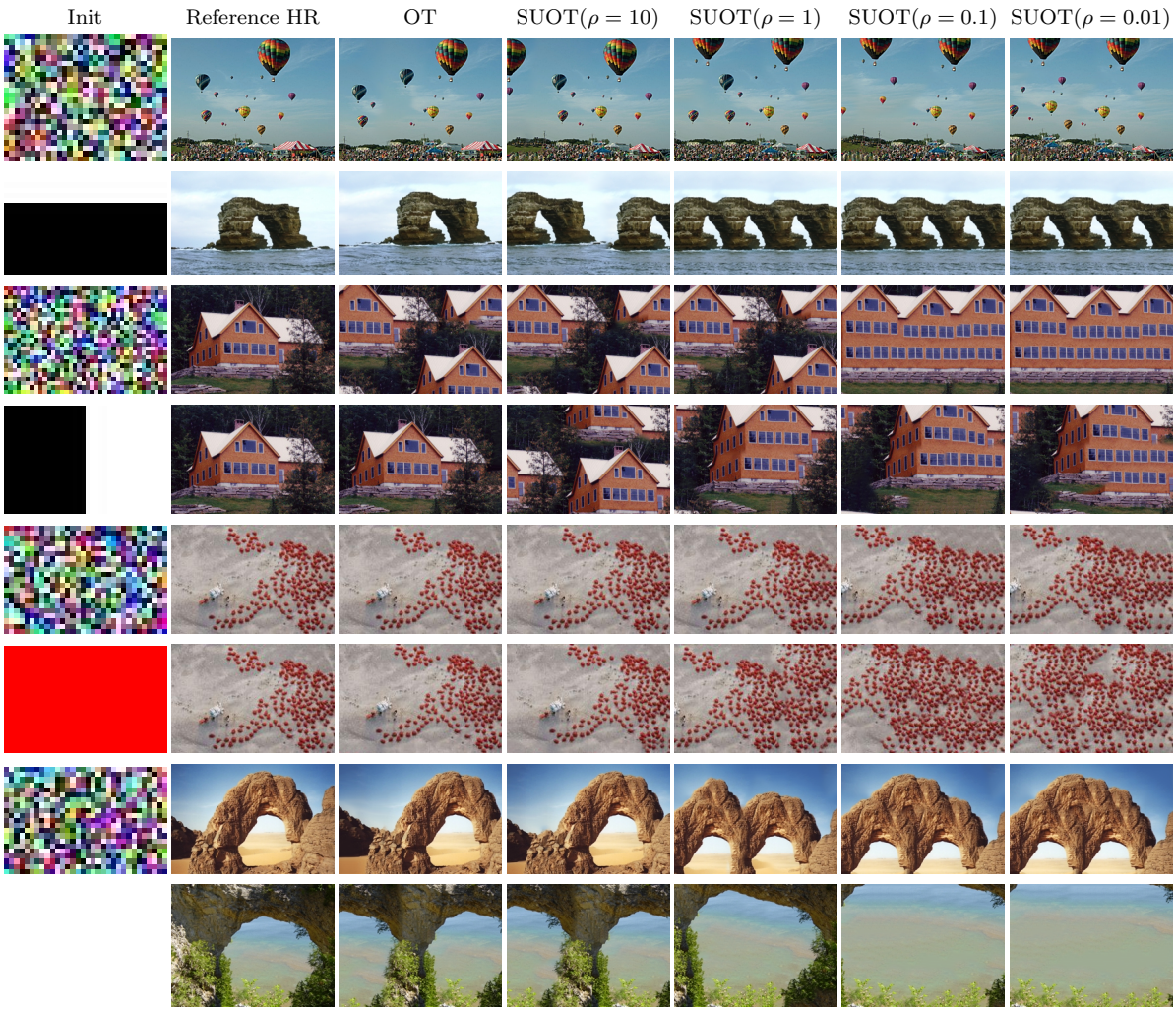

Figure 10: **Impact of the initialization image and the unbalance parameter ρ on PSinOT and PSinSUOT across different images**. From left to right: initialization image, HR reference image, image synthesis using OT, SUOT ($\rho = 10$), SUOT ($\rho = 1$), SUOT ($\rho = 0.1$), and SUOT ($\rho = 0.01$). Random initializations are obtained with a realization of $\mathcal{N}(0.5, 1)$ white noise.

[8] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. In *International Journal of Computer Vision*, volume 129(4), pages 1038–1059, 2021.

[9] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. MVTec AD — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on CVPR*, pages 9584–9592, 2019.

[10] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial GAN. *arXiv preprint arXiv:1705.06566*, 2017.

[11] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pages 880–889. PMLR, 2018.

[12] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer New York, 2005.

[13] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005.

[14] Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282, 2021.

[15] Benjamin Charlier, Jean Feydy, Joan Alexis Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

[16] Nicolas Cherel, Andrés Almansa, Yann Gousseau, and Alasdair Newson. A patch-based algorithm for diverse and high fidelity single image generation. In *29th IEEE International Conference on Image Processing (ICIP) 2022*, Bordeaux, France, October 2022.

[17] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[18] Taeg Sang Cho, Charles L Zitnick, Neel Joshi, Sing Bing Kang, Richard Szeliski, and William T Freeman. Image restoration by matching gradient distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):683–694, 2011.

[19] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

[20] Mauricio Delbracio, Hossein Talebei, and Peyman Milanfar. Projected distribution loss for image enhancement. In *IEEE International Conference on Computational Photography (ICCP 2021)*, pages 1–12. IEEE, 2021.

[21] M. El Gheche, J.-F. Aujol, Y. Berthoumieu, and C.-A. Deledalle. Texture reconstruction guided by a high-resolution patch. *IEEE Transactions on Image Processing*, 26(2):549–560, 2016.

[22] Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 544–560. Springer, 2022.

[23] J. Feydy. *Geometric data analysis, beyond convolutions*. PhD thesis, University of Paris-Saclay, 2020.

[24] J. Feydy, T. Séjourné, F. Vialard, S. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. *AiSTATS*, pages 2681–2690, 2019.

[25] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

[26] B. Galerne, A. Leclaire, and J. Rabin. A texture synthesis model based on semi-discrete optimal transport in patch space. *SIAM Journal on Imaging Sciences*, 11(4):2456–2493, 2018.

[27] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[28] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[29] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[31] N.M. Gottschling, V. Antun, B. Adcock, and A.C. Hansen. The troublesome kernel: why deep learning for inverse problems is typically unstable. *arXiv preprint:2001.01258*, 2020.

[32] Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani. Drop the GAN: In defense of patches nearest neighbors as single image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13460–13469, June 2022.

[33] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[34] J. Gutierrez, B. Galerne, J. Rabin, and T. Hurtut. Optimal patch assignment for statistically constrained texture synthesis. In *Scale Space and Variational Methods in Computer Vision 2017*, pages 172–183, 2017.

[35] Andreas Habring and Martin Holler. Neural-network-based regularization methods for inverse problems in imaging. *arXiv preprint arXiv:2312.14849*, 2023.

[36] Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced Wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9412–9420, June 2021.

[37] J. Hertrich, A. Houdard, and C. Redenbach. Wasserstein patch prior for image superresolution. *IEEE Transactions on Computational Imaging*, 8:693–704, 2022.

[38] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[39] A. Houdard, A. Leclaire, N. Papadakis, and J. Rabin. Wasserstein generative models for patch-based texture synthesis. In *Scale Space and Variational Methods in Computer Vision*, pages 269–280, 2021.

[40] Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin. A generative model for texture synthesis based on optimal transport between feature distributions. *Journal of Mathematical Imaging and Vision*, 65(1):4–28, 2023.

[41] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[42] Johan Karlsson and Axel Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

[43] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[44] Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21947–21959. Curran Associates, Inc., 2021.

[45] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2230–2236, 2017.

[46] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[47] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 427–443, 2019.

[48] Simon Mignon, Bruno Galerne, Moncef Hidane, Cécile Louchet, and Julien Mille. Semi-unbalanced regularized optimal transport for image restoration. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 466–470. IEEE, 2023.

[49] V. Monga, Y. Li, and Y.C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[50] Debarghya Mukherjee, Aritra Guha, Justin M Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR, 2021.

[51] Sloan Nietert, Ziv Goldfeld, and Rachel Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11691–11719. PMLR, 28–30 Mar 2022.

[52] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

[53] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[54] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[55] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40:49–70, 2000.

[56] Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856. IEEE, 2014.

[57] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.

[58] P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus de l'Académie des Sciences, Mathématique*, 356(11-12):1228–1235, 2018.

[59] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[60] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63), 2015.

[61] T. Séjourné, J. Feydy, F. Vialard, A. Trouvé, and G. Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint:1910.12958*, 2019.

[62] Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *arXiv preprint arXiv:2211.08775*, 2022.

[63] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. SinGAN: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4570–4580, 2019.

[64] Karen Simonyan and Andrew Zisserman. Very deep convnets for large-scale image recognition. *Computing Research Repository*, 2014.

[65] Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. When OT meets MOM: Robust estimation of Wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2021.

[66] Jean-Luc Starck, Fionn Murtagh, and Jalal M Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge university press, 2010.

[67] Paul Swoboda and Christoph Schnörr. Convex variational image restoration with histogram priors. *SIAM Journal on Imaging Sciences*, 6(3):1719–1735, 2013.

[68] G. Tartavel, G. Peyré, and Y. Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016.

[69] Chunwei Tian, Yong Xu, Wangmeng Zuo, Chia-Wen Lin, and David Zhang. Asymmetric CNN for image superresolution. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(6):3718–3730, 2021.

[70] Dmitry Ulyanov, Vadim Lebedev, Vedaldi Andrea, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1349–1357, New York, USA, 20–22 Jun 2016. PMLR.

[71] Jonathan Vacher, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Texture interpolation for probing visual perception. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22146–22157. Curran Associates, Inc., 2020.

[72] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[73] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[74] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on CVPR*, pages 586–595, 2018.

A Proofs of Section 2.3

A.1 Proof of Theorem 3

The proof consists in applying the Fenchel-Rockafellar theorem that we now recall.

Theorem 12 (Fenchel-Rockafellar)**.** *[12, Theorem 3.3.5] Let E and H be two normed vector space, $\mathrm{A} : E \to F$ a linear mapping, $r : E \to (-\infty, +\infty]$ and $h : H \to (\infty, +\infty]$ two convex functions satisfying the qualification condition:*

$$0 \in \mathrm{core}(\mathrm{dom}(h) - \mathrm{A}\,\mathrm{dom}(r)), \tag{40}$$

where the core of a set S denoted as core(S) is defined as the set of points x in S such that for every direction d, $x + td$ lies in S for all sufficiently small t.

Moreover, assume that $\inf_E(r + h \circ \mathrm{A}) > -\infty$. *Then, we have:*

$$\inf_{x \in E} r(x) + h(\mathrm{A}\, x) = \max_{y \in H} -r^\star(\mathrm{A}^T\, y) - h^\star(-y), \tag{41}$$

where $r^\star$ and $h^\star$ are the Legendre conjugates of r and h, respectively.

The proof is conducted for discrete distributions α and β with strictly positive weights a and b, respectively. We can write primal SUOT as follows:

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \inf_{\pi \in \mathbb{R}^{M \times N}} \langle C, \pi \rangle + \iota_{\mathbb{R}_+^{M \times N}}(\pi) + \iota_a(\pi_1) + \rho \mathrm{KL}(\pi_2 | b).$$

Defining:

$$\begin{array}{rccc} \mathrm{A} & : & \mathbb{R}^{N \times M} & \to & \mathbb{R}^N \times \mathbb{R}^M \\ & & \pi & \mapsto & (\pi_1, \pi_2) \end{array},$$

the linear operator that extracts the first and second marginals from π, the function r as

$$\begin{aligned} r : \mathbb{R}^{N \times M} &\to \mathbb{R} \\ \pi &\mapsto \langle C, \pi \rangle + \iota_{\mathbb{R}_+^{M \times N}}(\pi), \end{aligned}$$

and the function h as

$$\begin{aligned} h : \mathbb{R}^N \times \mathbb{R}^M &\to \mathbb{R} \\ (\pi_1, \pi_2) &\mapsto \iota_a(\pi_1) + \rho \mathrm{KL}(\pi_2 | b), \end{aligned}$$

we can rewrite the primal SUOT as follows

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \inf_{\pi \in \mathbb{R}^{M \times N}} r(\pi) + h \circ \mathrm{A}(\pi).$$

The conditions of the Fenchel-Rockafellar theorem 12 are satisfied in this context:

- r and h are two convex functions,
- since the function to be minimized is positive, its minimum is positive,
- the condition (40) is satisfied: $0 \in \mathrm{core}(\mathrm{dom}(h) - \mathrm{A}\,\mathrm{dom}(r))$.

Let us justify the latter condition. Firstly, consider that $(a, b) \in \mathrm{dom}(h)$ and $(a \otimes b) \in \mathrm{dom}(r)$. As $(a, b) - \mathrm{A}(a \otimes b) = 0$, it implies that $0 \in \mathrm{dom}(h) - \mathrm{A}\,\mathrm{dom}(r)$. Now, let $d = (d_1, d_2) \in \mathbb{R}^N \times \mathbb{R}^M$ be an arbitrary direction. Since $a > 0$ and $b > 0$, we can select a sufficiently small value for t such that all components of $(a - td_1, b - td_2)$ are strictly positive. As a consequence of $(a - td_1) \otimes (b - td_2) \in \mathrm{dom}(r)$ and since $(a, b) - \mathrm{A}(a - td_1) \otimes (b - td_2) = (td_1, td_2)$, we have $0 + td \in (\mathrm{dom}(h) - \mathrm{A}\,\mathrm{dom}(r))$. By observing this, we can conclude that the condition (40) is satisfied.

According to the Fenchel-Rockafellar theorem, we obtain

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \max_{(f,g) \in \mathbb{R}^N \times \mathbb{R}^M} -r^\star \circ \mathrm{A}^T(f, g) - h^\star(-(f, g)),$$

with

$$\begin{aligned} r^\star \circ A^T(f,g) &= r^\star(f \oplus g) \\ &= \sup_{\tilde{\pi} \in \mathbb{R}_+^{N \times M}} \langle f \oplus g, \tilde{\pi} \rangle - \langle C, \tilde{\pi} \rangle \\ &= \sup_{\tilde{\pi} \in \mathbb{R}_+^{N \times M}} \langle f \oplus g - C, \tilde{\pi} \rangle = \begin{cases} 0 & \text{if } f \oplus g \leq C, \\ +\infty & \text{otherwise,} \end{cases} \end{aligned}$$

and

$$\begin{aligned} h^\star(f,g) &= \sup_{(\tilde{f},\tilde{g}) \in \mathbb{R}^N \times \mathbb{R}^M} \left\langle (f,g), (\tilde{f},\tilde{g}) \right\rangle - h(\tilde{f},\tilde{g}) \\ &= \sup_{(\tilde{f},\tilde{g}) \in \mathbb{R}^N \times \mathbb{R}^M} \left\langle f, \tilde{f} \right\rangle + \left\langle g, \tilde{g} \right\rangle - \iota_a(\tilde{f}) - \rho \mathrm{KL}(\tilde{g}|b) \\ &= \sup_{\tilde{f} \in \mathbb{R}^N} \left\langle f, \tilde{f} \right\rangle - \iota_a(\tilde{f}) + \sup_{\tilde{g} \in \mathbb{R}^M} \left\langle g, \tilde{g} \right\rangle - \rho \mathrm{KL}(\tilde{g}|b). \end{aligned}$$

We observe that this problem is separable, and by the definition of ι_a, we have:

$$\sup_{\tilde{f} \in \mathbb{R}^N} \left\langle f, \tilde{f} \right\rangle - \iota_a(\tilde{f}) = \left\langle f, a \right\rangle.$$

We are left to compute:

$$\sup_{\tilde{g} \in \mathbb{R}^M} \left\langle g, \tilde{g} \right\rangle - \rho \mathrm{KL}(\tilde{g}|b).$$

Due to the strict convexity of the ρKL divergence, it is a strictly concave function that attains its maximum when its gradient is zero:

$$\nabla_{\tilde{g}} \langle g, \tilde{g} \rangle - \rho \mathrm{KL}(\tilde{g}|b) = 0_M \iff g - \rho \log(\frac{\tilde{g}}{b}) = 0_M \iff \tilde{g} = b \odot \exp\left(\frac{g}{\rho}\right).$$

Defining $g^* = b \odot \exp\left(\frac{g}{\rho}\right)$, and using the expression of $\mathrm{KL}(g^*|b)$, we have:

$$\begin{aligned} \sup_{\tilde{g} \in \mathbb{R}^M} \langle g, \tilde{g} \rangle - \rho \mathrm{KL}(\tilde{g}|b) &= \langle g, g^* \rangle - \rho \mathrm{KL}(g^*|b) \\ &= \langle g, g^* \rangle - \rho \left(\left\langle g^*, \log(\frac{g^*}{b}) \right\rangle - \langle g^*, 1 \rangle + \langle b, 1 \rangle \right) \\ &= \langle g, g^* \rangle - \rho \left(\left\langle g^*, \frac{g}{\rho} \right\rangle - \langle g^*, 1 \rangle + \langle b, 1 \rangle \right) \\ &= \rho \langle g^* - b, 1 \rangle \\ &= \rho \left\langle b \odot \exp\left(\frac{g}{\rho}\right) - b, 1 \right\rangle \\ &= \left\langle b, \rho(\exp\left(\frac{g}{\rho}\right) - 1) \right\rangle \\ &= \langle b, \phi^\star(g) \rangle. \end{aligned}$$

So, we have

$$h^\star(f,g) = \langle a, f \rangle - \langle b, \phi^\star(-g) \rangle.$$

Finally,

$$\begin{aligned}\mathrm{SUOT}^{\rho}(\alpha,\beta) &= \max_{(f,g)\in\mathbb{R}^N\times\mathbb{R}^M} -r^\star\circ \mathrm{A}^T(f,g) - h^\star(-(f,g)) \\ &= \max_{(f,g)\in\Gamma(C)} \langle a,f\rangle - \langle b,\phi^\star(-g)\rangle\,,\end{aligned}$$

where $\Gamma(C)=\{(f,g)\in\mathbb{R}^N\times\mathbb{R}^M; f\oplus g\leq C\}$. This concludes the proof.

A.2 Proof of Proposition 5

For each $(i,j)\in[\![N]\!]\times[\![M]\!]$, and $(f,g)\in\Gamma(C)$, we have:

$$f_i \leq c(x_i,y_j) - g_j.$$

Thus, $f_i \leq \min_{j\in[\![M]\!]} c(x_i,y_j) - g_j$. By definition of g^c, and noting that α is a positive measure, we have:

$$\langle f,a\rangle \leq \langle g^c,a\rangle\,,$$

and therefore

$$\langle a,f\rangle - \langle b,\phi^\star(-g)\rangle \leq \langle a,g^c\rangle - \langle b,\phi^\star(-g)\rangle\,. \tag{42}$$

Since $(g^c,g)\in\Gamma(C)$, it follows that if $(f^\star,g^\star)$ is a solution of the dual problem of SUOT, as indicated by inequality (42), then $(g^{\star c},g^\star)$ is also a solution of the dual SUOT.

A.3 Proof of Proposition 6

This is the same proof as the one for OT presented in [39]. Let $g^\star$ be a solution of SUOT. Then $\mathrm{SUOT}^{\rho}(\alpha,\beta)=F(g^\star,x=(x_1,x_2,\ldots,x_N))$, with $F(g,x)=\langle a,g^c\rangle-\langle b,\phi^\star(-g)\rangle$. Since by hypothesis, $x_i\in\mathcal{L}(g^\star)$, it belongs to a unique Laguerre cell indexed by $j(i,g^\star)$, hence

$$F(g^\star,x=(x_1,x_2,\ldots,x_N)) = a_i\left(c(x_i,y_{j(i,g^\star)}) - g^\star_{j(i,g^\star)}\right) + C$$

where C is a constant wrt to x_i. Consequently, one has $\nabla_{x_i}\mathrm{SUOT}^{\rho}(\alpha,\beta) = a_i\nabla_{x_i}c(x_i,y_{j(i,g^\star)})$.

A.4 Proof of Proposition 7

(i) For every $(i,j)\in[\![N]\!]\times[\![M]\!]$, we introduce the functions $f_i: g\mapsto g_i^c$ and $h_j: g\mapsto -\phi^\star(-g_j/\rho)$. Let us observe that

$$F(g) = \langle a,g^c\rangle - \langle b,\phi^\star(-g)\rangle = \sum_{i=1}^N a_i f_i(g) + \sum_{j=1}^M b_j h_j(g). \tag{43}$$

First, we establish the concavity of each f_i. Consider $g^1,g^2\in\mathbb{R}^M$ and $t\in[0,1]$ and define $g=tg^1+(1-t)g^2$. We have:

$$\begin{aligned}f_i(g)=f_i(tg^1+(1-t)g^2) &= \min_{j\in[\![M]\!]}\left[c(x_i,y_j) - tg^1_j - (1-t)g^2_j\right] \\ &= c(x_i,y_{j(i,g)}) - tg^1_{j(i,g)} - (1-t)g^2_{j(i,g)} \\ &= t\left[c(x_i,y_{j(i,g)}) - g^1_{j(i,g)}\right] + (1-t)\left[c(x_i,y_{j(i,g)}) - g^2_{j(i,g)}\right] \\ &\geq t\min_{j\in[\![M]\!]}\left[c(x_i,y_j) - g^1_j\right] + (1-t)\min_{j\in[\![M]\!]}\left[c(x_i,y_j) - g^2_j\right] \\ &= tf_i(g^1) + (1-t)f_i(g^2).\end{aligned}$$

The concavity of h_j directly stems from the convexity of the exponential function. Thus, for all $(i,j) \in [\![N]\!] \times [\![M]\!]$, both f_i and h_j are concave functions. Thanks to (43) we can conclude that F, as a linear combination of concave functions with positive weights, is a concave function.

(ii) For a given $g \in \mathbb{R}^M$, $SG_F(g)$ is a super-gradient of F in g if

$$\forall g' \in \mathbb{R}^M, \quad F(g) - F(g') + \langle SG_F(g), g' - g \rangle \geq 0. \tag{44}$$

We have, for every $g' \in \mathbb{R}^M$,

$$F(g) - F(g') + \langle SG_F(g), g' - g \rangle = \sum_{i=1}^{N} a_i \min_{j \in [\![M]\!]} \Big[c(x_i, y_j) - g_j \Big] + \sum_{j=1}^{M} b_j \rho (1 - \exp\big(-\frac{g_j}{\rho}\big)) \tag{45}$$

$$- \sum_{i=1}^{N} a_i \min_{j \in [\![M]\!]} \Big[c(x_i, y_j) - g'_j \Big] - \sum_{j=1}^{M} b_j \rho (1 - \exp\big(-\frac{g'_j}{\rho}\big)) \tag{46}$$

$$- \sum_{i=1}^{N} a_i (g'_{j(i,g)} - g_{j(i,g)}) + \sum_{j=1}^{M} b_j \exp\big(-\frac{g_j}{\rho}\big)(g'_j - g_j). \tag{47}$$

Let us concentrate on the sum formed by the first terms in lines (45), (46), (47),

$$\sum_{i=1}^{N} a_i \min_{j \in [\![M]\!]} \Big[c(x_i, y_j) - g_j \Big] - \sum_{i=1}^{N} a_i \min_{j \in [\![M]\!]} \Big[c(x_i, y_j) - g'_j \Big] - \sum_{i=1}^{N} a_i (g'_{j(i,g)} - g_{j(i,g)}),$$

and demonstrate that it has a positive value. We note that $j(i,g)$ in the last sum satisfies the minimum in the first sum. We can therefore rewrite the problem:

$$\sum_{i=1}^{N} a_i \Big[c(x_i, y_{j(i,g)}) - g_{j(i,g)} \Big] - \sum_{i=1}^{N} a_i \min_{j \in [\![M]\!]} \Big[c(x_i, y_j) - g'_j \Big] - \sum_{i=1}^{N} a_i (g'_{j(i,g)} - g_{j(i,g)}).$$

which simplifies to :

$$\sum_{i=1}^{N} a_i \Big[c(x_i, y_{j(i,g)}) - g'_{j(i,g)} \Big] - \sum_{i=1}^{N} a_i \min_{j \in [\![M]\!]} \Big[c(x_i, y_j) - g'_j \Big].$$

By definition of the last sum, we can conclude that the whole sum is positive.

Now, let us proceed to demonstrate that the sum formed by the second terms in lines (45), (46), (47),

$$\sum_{j=1}^{M} b_j \rho (1 - \exp\big(-\frac{g_j}{\rho}\big)) - \sum_{j=1}^{M} b_j \rho (1 - \exp\big(-\frac{g'_j}{\rho}\big)) + \sum_{j=1}^{M} b_j \exp\big(-\frac{g_j}{\rho}\big)(g'_j - g_j),$$

is positive. This sum can be rewritten as:

$$\phi(g') = \left\langle b \odot \exp\big(\frac{-g}{\rho}\big), g' - g - \rho \right\rangle + \left\langle b \odot \exp\big(-\frac{g'}{\rho}\big), \rho \right\rangle.$$

ϕ is a convex function and we have $\nabla\phi(g') = 0 \iff g' = g$ with $\nabla\phi = \left\langle b, \exp\left(-\frac{g}{\rho}\right) - \exp\left(-\frac{g'}{\rho}\right)\right\rangle$. So, as a convex function, ϕ has a global minimum in g with $\phi(g) = 0$. We conclude that $\phi \geq 0$.

Since it is the sum of two positive functions, we can conclude that SG_F is a supergradient of F.

B Simple Case Example of Section 2.3.3

Here we give a proof of (15) and (16). For any $\rho > 0$, we have

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \min_{\pi \in \mathbb{R}^{1\times 2}} \pi_{11}\|y_1 - x\|^2 + \pi_{12}\|y_2 - x\|^2 + \iota_{\pi_1}(a) + \rho\,\mathrm{KL}(\pi_2|b)$$

where $\pi \in \mathbb{R}^{1\times 2}$ is denoted as $\pi = (1-\eta, \eta)$. With this, the constraint $\pi_1 = 1$ becomes redundant with $\pi_1 = a$, and we obtain

$$\mathrm{SUOT}^\rho(\alpha, \beta) = \min_{\eta \in [0,1]} (1-\eta)\|y_1 - x\|^2 + \eta\|y_2 - x\|^2 + \rho\big((1-\eta)\log\frac{1-\eta}{b_1} + \eta\log\frac{\eta}{b_2}\big)$$

whose right-hand side is a convex function f of the real variable η. To minimize f, let $d = \|y_2 - x\|^2 - \|y_1 - x\|^2$. Then, we have

$$f'(\eta) = d + \rho\log\frac{\eta(b_1)}{(1-\eta)b_2}.$$

Denoting η^* the minimizer of f, we deduce (15) from $f'(\eta^*) = 0$.

Notice that η^* is still in $[0,1]$, as a fraction of the form $B/(A+B)$ with $A \geq 0$, $B \geq 0$ and $A + B > 0$. The optimal transport plan is hence $\pi^* = (1-\eta^*, \eta^*)$ and $\mathrm{SUOT}^\rho(\alpha, \beta) = f(\eta^*)$ yields (16).

C Proofs of Section 3.2

C.1 Proof of the Dual Formulation in Equation (25)

This proof is an adaptation of Theorem 3. We use the same notations, with the key difference being the incorporation of the regularization term εKL into the function r:

$$\begin{aligned} r : \mathbb{R}^{N\times M} &\to \mathbb{R} \\ \pi &\mapsto \langle C, \pi\rangle + \varepsilon\mathrm{KL}(\pi|a\otimes b), \end{aligned}$$

which leaves its domain unchanged. Since r is a positive convex function defined on $\mathbb{R}_+^{N\times M}$, the conditions of the Fenchel-Rockafellar theorem are met. Therefore, RSUOT satisfies:

$$\mathrm{RSUOT}^\rho(\alpha, \beta) = \max_{(f,g)\in\mathbb{R}^N\times\mathbb{R}^M} -r^\star \circ \mathrm{A}^T(f, g) - h^\star(-(f, g)). \tag{48}$$

As demonstrated in the proof of Theorem 3, recall that:

$$h^\star(f, g) = \langle a, f\rangle - \langle b, \phi^\star(-g)\rangle\,.$$

Next, let us calculate the Legendre transform $r^\star$:

$$\begin{aligned} r^\star \circ \mathrm{A}^T(f,g) &= r^\star(f \oplus g) \\ &= \sup_{\tilde{\pi} \in \mathbb{R}^{N \times M}} \langle f \oplus g, \tilde{\pi} \rangle - r(\tilde{\pi}), \\ &= \sup_{\tilde{\pi} \in \mathbb{R}^{N \times M}} \langle f \oplus g, \tilde{\pi} \rangle - \langle C, \tilde{\pi} \rangle - \varepsilon \mathrm{KL}(\tilde{\pi} | a \otimes b), \\ &= \sup_{\tilde{\pi} \in \mathbb{R}^{N \times M}} \langle f \oplus g - C, \tilde{\pi} \rangle - \varepsilon \mathrm{KL}(\tilde{\pi} | a \otimes b), \end{aligned}$$

which, being a strictly concave function, reaches its maximum when its gradient vanishes, i.e., at:

$$\pi^\star = (a \otimes b) \odot \exp\left(\frac{f \oplus g - C}{\varepsilon}\right),$$

Expanding the KL term leads to the following expression:

$$\begin{aligned} r^\star \circ \mathrm{A}^T(f,g) &= \left\langle (a \otimes b) \odot \exp\left(\frac{f \oplus g - C}{\varepsilon}\right), f \oplus g - C \right\rangle \\ &- \left\langle (a \otimes b) \odot \exp\left(\frac{f \oplus g - C}{\varepsilon}\right), f \oplus g - C \right\rangle \\ &+ \varepsilon \left\langle (a \otimes b) \odot \exp\left(\frac{f \oplus g - C}{\varepsilon}\right), 1 \right\rangle - \varepsilon \langle a \otimes b, 1 \rangle \\ &= \varepsilon \left\langle (a \otimes b) \odot \exp\left(\frac{f \oplus g - C}{\varepsilon}\right), 1 \right\rangle - \varepsilon \langle a \otimes b, 1 \rangle \\ &= \varepsilon \left\langle a \otimes b, \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) - 1 \right\rangle. \end{aligned}$$

Finally, following (48), we have:

$$\mathrm{RSUOT}_\varepsilon^\rho(\alpha, \beta) = \max_{(f,g) \in \mathbb{R}^N \times \mathbb{R}^M} \langle a, f \rangle - \langle b, \phi^\star(-g) \rangle - \varepsilon \left\langle a \otimes b, \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) - 1 \right\rangle,$$

with $\phi^\star(q) = \rho(\exp\left(\frac{q}{\rho}\right) - 1)$. This concludes the proof.

C.2 Proof of the Formulation of RSUOT Gradient in Equation (26)

Let $(f^\star, g^\star)$ be the solution of the dual RSUOT. Then:

$$\mathrm{RSUOT}_\varepsilon^\rho(\sum_{i=1}^N a_i \delta_{x_i}, \beta) = F(f^\star, g^\star, x = (x_1, \ldots, x_N)),$$

with

$$F(f,g,x) = \langle a, f \rangle - \langle b, \phi^\star(-g) \rangle - \varepsilon \left\langle a \otimes b, \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) - 1 \right\rangle.$$

When F is maximum, the gradient $\nabla_f F(f^\star, g^\star, x)$ is zero and a direct calculation leads to the following expression of $f^\star$ wrt $g^\star$:

$$f_i^\star = -\varepsilon \log\left(\sum_{j=1}^M \exp\left(\log(b_j) + \frac{1}{\varepsilon} g_j^\star - \frac{1}{\varepsilon} c_{i,j}\right)\right). \tag{49}$$

In addition, we have :

$$\begin{aligned}\nabla_{x_i} \mathrm{RSUOT}_\varepsilon^\rho(\alpha,\beta) &= \nabla_{x_i} F(f^\star, g^\star, x) \\ &= -\nabla_{x_i}\varepsilon \left\langle a \otimes b, \exp\left(\frac{f^\star \oplus g^\star - C}{\varepsilon}\right) - 1 \right\rangle \\ &= a_i \sum_{j=1}^M b_j \exp\left(\frac{f_i^\star + g_j^\star - c_{ij}}{\varepsilon}\right) \nabla_{x_i} c_{ij} \\ &= a_i \exp\left(\frac{f_i^\star}{\varepsilon}\right) \sum_{j=1}^M b_j \exp\left(\frac{g_j^\star - c_{ij}}{\varepsilon}\right) \nabla_{x_i} c_{ij}.\end{aligned}$$

Now, including the expression of $f^\star$ given in (49), we can write

$$\begin{aligned}\nabla_{x_i} \mathrm{RSUOT}_\varepsilon^\rho(\alpha,\beta) &= a_i \exp\left(\frac{-\varepsilon \log\left(\sum_{j=1}^M \exp\left(\log(b_j) + \frac{1}{\varepsilon} g_j^\star - \frac{1}{\varepsilon} c_{i,j}\right)\right)}{\varepsilon}\right) \\ &\quad \times \sum_{j=1}^M b_j \exp\left(\frac{g_j^\star - c_{ij}}{\varepsilon}\right) \nabla_{x_i} c_{ij} \\ &= a_i \frac{\sum_{j=1}^M \exp\left(\log(b_j) + \frac{1}{\varepsilon} g_j^\star - \frac{1}{\varepsilon} c_{i,j}\right) \nabla_{x_i} c_{i,j}}{\sum_{j=1}^M \exp\left(\log(b_j) + \frac{1}{\varepsilon} g_j^\star - \frac{1}{\varepsilon} c_{i,j}\right)} \\ &= a_i \nabla_{x_i} - \varepsilon \log\left(\sum_{j=1}^M \exp\left(\log(b_j) + \frac{1}{\varepsilon} g_j^\star - \frac{1}{\varepsilon} c_{i,j}\right)\right) \\ &= a_i \nabla \varphi(x_i).\end{aligned}$$

This concludes the proof.

C.3 Proof of Theorem 11

Preliminaries For any $u \in (\mathbb{R}_+^*)^K$, $K \in \mathbb{N}^\star$, $\varepsilon > 0$, we define the operator:

$$\begin{array}{rcccl} \mathrm{Smin}_u^\varepsilon & : & \mathbb{R}^K & \to & \mathbb{R} \\ & & v & \mapsto & -\varepsilon \log\left(\sum_{i=1}^K u_i \exp(-v_i/\varepsilon)\right), \end{array}$$

and refer to as the Smin operator. In addition, let us define

$$\begin{array}{rcccl} F & : & \mathbb{R}^M & \to & \mathbb{R}^N \\ & & g & \mapsto & [\mathrm{Smin}_b^\varepsilon(g - c_{i,\cdot})]_{i \in [\![N]\!]}, \\ G & : & \mathbb{R}^N & \to & \mathbb{R}^M \\ & & f & \mapsto & [\frac{1}{1+\frac{\varepsilon}{\rho}} \mathrm{Smin}_a^\varepsilon(f - c_{\cdot,j})]_{j \in [\![M]\!]}, \end{array}$$

where $c_{i,\cdot}$ and $c_{\cdot,j}$ are respectively the vectors obtained by extracting the i-th row and the j-th column from the cost matrix C. The proof of *(i)* and *(iii)* requires the following two lemmas:

Lemma 13 (Non-expansivity of F)**.** *F is a non-expansive operator:*

$$\forall (g_1, g_2) \in \mathbb{R}^M, \quad \|F(g_1) - F(g_2)\|_\infty \leq \|g_1 - g_2\|_\infty \tag{50}$$

This result follows from the established non-expansivity property of the Smin operator [61, Lemma 1].

Lemma 14 (Contractivity of G)**.** *G is a contractive operator:*

$$\forall (f_1, f_2) \in \mathbb{R}^N, \quad \|G(f_1) - G(f_2)\|_\infty \leq \frac{1}{1+\frac{\varepsilon}{\rho}} \|f_1 - f_2\|_\infty \tag{51}$$

This again is a consequence of the non-expansivity of Smin.
Now, let us come to the proof of Theorem 11.

(i) Using G and F, we can reformulate Sinkhorn's algorithm as the recursion:

$$\begin{aligned} g^{t+1} &= G(f^t) \\ f^{t+1} &= F(g^{t+1}), \end{aligned}$$

Furthermore, let $(f^\star, g^\star) \in \mathbb{R}^N \times \mathbb{R}^M$ be the solution to Problem (25), which satisfies:

$$\begin{aligned} g^\star &= G(f^\star), \\ f^\star &= F(g^\star). \end{aligned}$$

Then we have:

$$\begin{aligned} \|f^{t+1} - f^\star\|_\infty = \|F(g^{t+1}) - F(g^\star)\|_\infty \leq \|g^{t+1} - g^\star\|_\infty = \|G(f^t) - G(f^\star)\|_\infty \\ \leq \frac{1}{1+\frac{\varepsilon}{\rho}} \|f^t - f^\star\|_\infty, \end{aligned}$$

and

$$\begin{aligned} \|g^{t+1} - g^\star\|_\infty = \|G(f^t) - G(f^\star)\|_\infty \leq \frac{1}{1+\frac{\varepsilon}{\rho}} \|f^t - f^\star\|_\infty = \frac{1}{1+\frac{\varepsilon}{\rho}} \|F(g^t) - F(g^\star)\|_\infty \\ \leq \frac{1}{1+\frac{\varepsilon}{\rho}} \|g^t - g^\star\|_\infty. \end{aligned}$$

Hence, we have :

$$\|g^{t+1} - g^\star\|_\infty + \|f^{t+1} - f^\star\|_\infty \leq \frac{1}{1+\frac{\varepsilon}{\rho}} \left(\|g^t - g^\star\|_\infty + \|f^t - f^\star\|_\infty \right).$$

Given that $\frac{1}{1+\frac{\varepsilon}{\rho}} < 1$, the iterates (f^t, g^t) linearly converge to the unique solution $(f^\star, g^\star)$ of the RSUOT problem.

(ii) Upon replacing f^t with $F(g^t)$ in the third term of the dual expression (25) of RSUOT and concomitantly using the definition of F, we obtain:

$$\begin{aligned} \varepsilon \left\langle a \otimes b, \exp\left(\frac{f^t \oplus g^t - C}{\varepsilon}\right) - 1 \right\rangle &= \varepsilon \left\langle a \otimes b, \exp\left(\frac{f^t \oplus g^t - C}{\varepsilon}\right) \right\rangle - \varepsilon \\ &= \varepsilon \left\langle a \odot \exp\left(\frac{F(g^t)}{\varepsilon}\right), \exp\left(-\frac{F(g^t)}{\varepsilon}\right) \right\rangle - \varepsilon \\ &= \varepsilon \langle a, 1 \rangle - \varepsilon = 0. \end{aligned}$$

(iii) Using G and F, we can rewrite the symmetric fixed-point iterations algorithm:

$$\tilde{g}^{t+1} = \frac{1}{2}\left(\tilde{g}^t + G(\tilde{f}^t)\right),$$
$$\tilde{f}^{t+1} = \frac{1}{2}\left(\tilde{f}^t + F(\tilde{g}^t)\right).$$

Furthermore, if $(f^\star, g^\star) \in \mathbb{R}^N \times \mathbb{R}^M$ be a solution to Problem (25), utilizing

$$g^\star = \frac{1}{2}g^\star + \frac{1}{2}G(f^\star),$$
$$f^\star = \frac{1}{2}f^\star + \frac{1}{2}F(g^\star),$$

alongside the non-expansivity of F and the contractivity of G, we can derive the following two inequalities:

$$\begin{aligned}\|\tilde{f}^t - f^\star\|_\infty &\leq \left\|\frac{1}{2}\left(\tilde{f}^{t-1} - f^\star\right) + \frac{1}{2}\left(F(\tilde{g}^{t-1}) - F(g^\star)\right)\right\|_\infty \\ &\leq \frac{1}{2}\|\tilde{f}^{t-1} - f^\star\|_\infty + \frac{1}{2}\|\tilde{g}^{t-1} - g^\star\|_\infty\end{aligned}$$

and

$$\begin{aligned}\|\tilde{g}^t - g^\star\|_\infty &= \left\|\frac{1}{2}\left(\tilde{g}^{t-1} - g^\star\right) + \frac{1}{2}\left(G(\tilde{f}^{t-1}) - G(f^\star)\right)\right\|_\infty \\ &\leq \frac{1}{2}\|\tilde{g}^{t-1} - g^\star\|_\infty + \frac{1}{2}\frac{1}{1+\frac{\varepsilon}{\rho}}\|\tilde{f}^{t-1} - f^\star\|_\infty.\end{aligned}$$

By summing these two inequalities and subsequently applying them at rank $t-1$, we obtain:

$$\|\tilde{f}^t - f^\star\|_\infty + \left\|\tilde{g}^t - g^\star\right\|_\infty \leq \frac{3 + \frac{1}{1+\frac{\varepsilon}{\rho}}}{4}\left(\|\tilde{f}^{t-2} - f^\star\|_\infty + \|\tilde{g}^{t-2} - g^\star\|_\infty\right).$$

Since $\frac{1}{4}(3 + \frac{1}{1+\frac{\varepsilon}{\rho}}) < 1$ we conclude that these symmetrical fixed-point iterations converge linearly to $(f^\star, g^\star)$.