An Optimal Transport Perspective on Unpaired Image Super-Resolution

Milena Gazdieva
Skoltech*, Moscow, Russia
m.gazdieva@skoltech.ru

Petr Mokrov
Skoltech*
Moscow, Russia

Litu Rout
University of Texas Austin
Austin, Texas, US

Alexander Korotin
Skoltech*, Moscow, Russia
AIRI†, Moscow, Russia

Alexander Filippov
Huawei Noah's Ark Lab
Moscow, Russia

Evgeny Burnaev
Skoltech*, Moscow, Russia
AIRI†, Moscow, Russia

Abstract

Real-world image super-resolution (SR) tasks often do not have paired datasets, which limits the application of supervised techniques. As a result, the tasks are usually approached by *unpaired* techniques based on Generative Adversarial Networks (GANs), which yield complex training losses with several regularization terms, e.g., content or identity losses. While GANs usually provide good practical performance, they are used heuristically, i.e., theoretical understanding of their behaviour is yet rather limited. We theoretically investigate optimization problems which arise in such models and find two surprising observations. First, the learned SR map is always an *optimal transport* (OT) map. Second, we theoretically prove and empirically show that the learned map is *biased*, i.e., it does not actually transform the distribution of low-resolution images to high-resolution ones. Inspired by these findings, we investigate recent advances in neural OT field to resolve the *bias* issue. We establish an intriguing connection between regularized GANs and neural OT approaches. We show that unlike the existing GAN-based alternatives, these algorithms aim to learn an *unbiased* OT map. We empirically demonstrate our findings via a series of synthetic and real-world unpaired SR experiments.

1 Introduction

The problem of image super-resolution (SR) is to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. In many modern deep learning approaches, SR networks are trained in a supervised manner by using synthetic datasets containing LR-HR *pairs* (Lim et al., 2017, §4.1); (Zhang et al., 2018b, §4.1). For example, it is common to create LR images from HR with a simple downscaling, e.g., bicubic (Ledig et al., 2017, §3.2). However, such an artificial setup barely represents the practical setting, in which the degradation is more sophisticated and unknown (Maeda, 2020). This obstacle suggests the necessity of developing methods capable of learning SR maps from *unpaired* data without considering prescribed degradations.

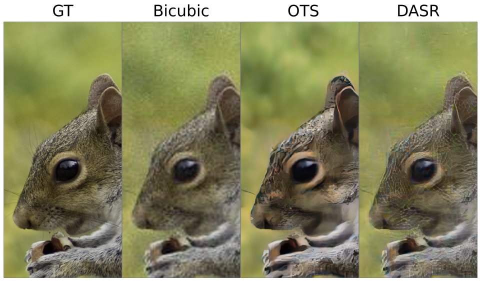


Figure 1: Super-resolution of a squirrel using Bicubic upsample, OTS and DASR (Wei et al., 2021) methods (4×4 upsample, 370×800 crops).

Contributions. We study the unpaired image SR task and its solutions based on Generative Adversarial Networks (Goodfellow et al., 2014, GANs) and analyse them from the Optimal Transport (Villani, 2008, OT) perspective.

1. **Theory I.** We investigate the GAN optimization objectives regularized with content losses, which are common in unpaired image SR methods (§4). We prove that the solution to such objectives is always an optimal transport map which is, in general, biased.

[*] Skolkovo Institute of Science and Technology
[†] Artificial Intelligence Research Institute

2. **Theory II.** We explain the ideas that stand behind recent algorithms from the field of neural OT (Korotin et al., 2023b; Fan et al., 2023) which aim to recover the true (unbiased) OT map. To do this, we show that their algorithms' optimization objective can be viewed as a certain particular case of GAN-based objectives regularized with content losses (§5). We also establish connections between these algorithms and regularized GANs that use integral probability metrics (IPMs) as a loss (§5.1).
3. **Practice.** We empirically show that oppositely to neural OT methods GANs' maps are *biased* (§6.1), i.e., they do not transform the LR image distribution to the true HR image distribution. We demonstrate the findings on the synthetic (§6.1) and real-world (§D) super-resolution task.

Notation. We use $\mathcal{X} = \mathbb{R}^{D_x}, \mathcal{Y} = \mathbb{R}^{D_y}$ to denote data spaces and $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y})$ to denote the respective sets of probability distributions on them. We denote by $\Pi(\mathbb{P}, \mathbb{Q})$ the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. For a measurable map $T : \mathcal{X} \to \mathcal{Y}$, we denote the associated push-forward operator by $T_{\#}$. The expression $\| \cdot \|$ denotes the usual Euclidean norm if not stated otherwise. We denote the space of $\mathbb{Q}$-integrable functions on $\mathcal{Y}$ by $L^1(\mathbb{Q})$.

2 Unpaired Image Super-Resolution Task

In this section, we formalize the *unpaired* image super-resolution task that we consider (Figure 2).

Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions of LR and HR images, respectively, on spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. We assume that $\mathbb{P}$ is obtained from $\mathbb{Q}$ via some *unknown* degradation. The learner has access to unpaired random samples from $\mathbb{P}$ and $\mathbb{Q}$. The task is to fit a map $T : \mathcal{X} \to \mathcal{Y}$ satisfying $T_{\#}\mathbb{P} = \mathbb{Q}$ which *inverts* the degradation.

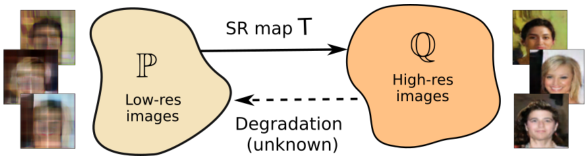


Figure 2: The task of super-resolution we consider.

We highlight that the image SR task is theoretically ill-posed for two reasons.

1. **Non-existence.** The degradation filter may be *non-injective* and, consequently, *non-invertible*. This is a theoretical obstacle to learn one-to-one SR maps T.
2. **Ambiguity.** There might exist *multiple* maps satisfying $T_{\#}\mathbb{P} = \mathbb{Q}$ but only one inverting the degradation. With no prior knowledge about the correspondence between $\mathbb{P}$ and $\mathbb{Q}$, it is unclear how to pick this particular map.

The first issue is usually not taken into account in practice. Most existing paired and unpaired SR methods learn one-to-one SR maps T, see (Ledig et al., 2017; Lai et al., 2017; Wei et al., 2021).

The second issue is typically softened by regularizing the model with the content loss. In the real-world, it is reasonable to assume that HR and the corresponding LR images are close. Thus, the fitted SR map T is expected to only *slightly* change the input image. Formally, one may require the learned map T to have the small value of

$$\mathcal{R}_c(T) \overset{def}{=} \int_{\mathcal{Y}} c\big(x, T(x)\big) d\mathbb{P}(x), \tag{1}$$

where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is a function estimating how different the inputs are. The most popular example is the ℓ^1 *identity* loss, i.e, formulation (1) for $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$ and $c(x, y) = \|x - y\|_1$.

More broadly, losses $\mathcal{R}_c(T)$ are typically called *content* losses and incorporated into training objectives of methods for SR (Lugmayr et al., 2019a, §3.4), (Kim et al., 2020, §3) and other unpaired tasks beside SR (Taigman et al., 2016, §4), (Zhu et al., 2017, §5.2) as regularizers. They stimulate the learned map T to minimally change the image content.

A common approach to solve the unpaired SR via GANs is to define a loss function $\mathcal{D} : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$ and train a generative neural network T via minimizing

$$\inf_{T:\mathcal{X} \mapsto \mathcal{Y}} \big[\mathcal{D}(T_{\#}\mathbb{P}, \mathbb{Q}) + \lambda \mathcal{R}_c(T)\big]. \tag{2}$$

The term $\mathcal{D}(T_{\#}\mathbb{P}, \mathbb{Q})$ ensures that the generated distribution $T_{\#}\mathbb{P}$ of SR images is close to the true HR distribution $\mathbb{Q}$. For convenience, we assume that $\mathcal{D}(\mathbb{Q}, \mathbb{Q}) = 0$ for all $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$. Two most popular examples of $\mathcal{D}$ are the Jensen–Shannon divergence (Goodfellow et al., 2014), i.e., the vanilla GAN loss, and the Wasserstein-1 loss (Arjovsky & Bottou, 2017).

In unpaired SR methods, the optimization objectives are typically more complex than (2). In addition to the content or identity loss (1), several other regularizations are usually introduced. Existing approaches to unpaired image SR mainly solve the problem in two steps. One group of approaches learn the degradation operation at the first step and then train a super-resolution model in a supervised manner using generated pseudo-pairs, see (Bulat et al., 2018; Fritsche et al., 2019). Another group of approaches (Yuan et al., 2018; Maeda, 2020) firstly learn a mapping from real-world LR images to “clean“ LR images, i.e., HR images, downscaled using predetermined (e.g., bicubic) operation, and then a mapping from “clean“ LR to HR images. Most methods are based on CycleGAN (Zhu et al., 2017), initially designed for the domain transfer task, and utilize cycle-consistency loss. Methods are also usually endowed with several other losses, e.g. content (Kim et al., 2020, §3), identity (Wang et al., 2021, §3.2) or perceptual (Lugmayr et al., 2019a, §3.4). However, we emphasize that all methods have unpaired learning step which corresponds to the optimization objective (2). In Appendix E, we show that the learning objectives of popular SR methods can be represented as (2).

3 Background on Optimal Transport

In this section, we give the key concepts of the OT theory (Villani, 2008) that we use in our paper.

Primal form. For two distributions $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and a transport cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, Monge's primal formulation of the *optimal transport cost* is as follows:

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) \overset{\text{def}}{=} \inf_{T_{\#}\mathbb{P}=\mathbb{Q}} \int_{\mathcal{X}} c\big(x, T(x)\big) d\mathbb{P}(x), \tag{3}$$

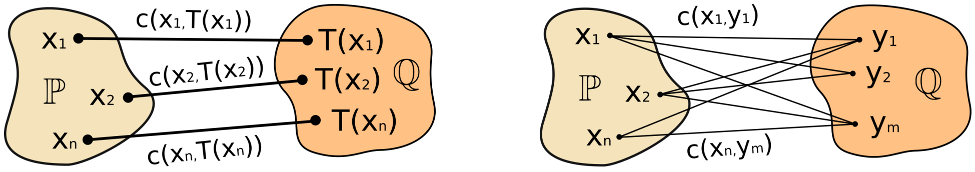
where the minimum is taken over the measurable functions (transport maps) $T : \mathcal{X} \to \mathcal{Y}$ that map $\mathbb{P}$ to $\mathbb{Q}$, see Figure 3a. The optimal T^* is called the *optimal transport map*.

Note that (3) is not symmetric, and this formulation does not allow mass splitting, i.e., for some $\mathbb{P}, \mathbb{Q}$ there may be no map T that satisfies $T_{\#}\mathbb{P} = \mathbb{Q}$. Thus, (Kantorovitch, 1958) proposed the relaxation:

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) \overset{\text{def}}{=} \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y), \tag{4}$$

where the minimum is taken over the transport plans π, i.e., the measures on $\mathcal{X} \times \mathcal{Y}$ whose marginals are $\mathbb{P}$ and $\mathbb{Q}$ (Figure 3b). The optimal $\pi^* \in \Pi(\mathbb{P}, \mathbb{Q})$ is called the *optimal transport plan*.

With mild assumptions on the transport cost $c(x, y)$ and distributions $\mathbb{P}$, $\mathbb{Q}$, the minimizer π^* of (4) always exists (Villani, 2008, Theorem 4.1) but might not be unique. If π^* is of the form $[\text{id}, T^*]_{\#}\mathbb{P} \in \Pi(\mathbb{P}, \mathbb{Q})$ for some T^*, then T^* is an optimal transport map that minimizes (3).



(a) Monge's formulation of OT.

(b) Kantorovich's formulation of OT.

Figure 3: Monge's and Kantorovich's formulations of Optimal Transport.

Dual form. The dual form (Villani, 2003) of OT cost (4) is as follows:

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \sup_{f} \left[\int_{\mathcal{X}} f^c(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} f(y) d\mathbb{Q}(y) \right]; \tag{5}$$

here sup is taken over all $f \in \mathcal{L}^1(\mathbb{Q})$, and $f^c(x) = \inf_{y \in \mathcal{Y}} \big[c(x, y) - f(y)\big]$ is the c-transform of f.

Optimal Transport in Generative Models. The majority of existing OT-based generative models employ OT cost as the loss function to update the generative network, e.g., see (Arjovsky et al., 2017). These methods are out of scope of the present paper, since they do not compute OT maps. Existing methods to compute the OT map approach the primal (3), (4) or dual form (5). Primal-form methods (Lu et al., 2020; Xie et al., 2019; Bousquet et al., 2017; Balaji et al., 2020) optimize complex GAN objectives such as (2) and provide biased solutions (§4, §6.1). For a comprehensive overview of dual-form methods, we refer to (Korotin et al., 2021). The authors conduct an evaluation of OT methods for the quadratic cost $c(x, y) = \|x - y\|^2$. According to them, the best performing method is $\lfloor$MM:R$\rceil$. Extensions of $\lfloor$MM:R$\rceil$ appear in (Rout et al., 2022; Fan et al., 2023).

4 BIASED OT IN GANS

In this section, we establish connections between GAN methods regularized by content losses (1) and OT. Such GANs are popular in a variety of tasks beside SR, e.g., style transfer (Huang et al., 2018). The theoretical analysis in this section holds for these tasks as well. However, since we empirically demonstrate the findings on the SR problem, we keep the corresponding notation in §4.

For a theoretical analysis, we stick to the basic formulation regularized with generic content loss (2). It represents the simplest and straightforward SR setup. We prove the following lemma, which connects the solution T^λ of (2) and OT maps.

Lemma 1 (The solution of the regularized GAN is an OT map)**.** *Assume that $\lambda > 0$ and the minimizer T^λ of* (2) *exists. Then T^λ is an OT map between $\mathbb{P}$ and $\mathbb{Q}^\lambda \overset{\text{def}}{=} T^\lambda_\# \mathbb{P}$ for cost $c(x,y)$, i.e., it minimizes*

$$\inf_{T_\#\mathbb{P}=\mathbb{Q}^\lambda} \mathcal{R}_c(T) = \inf_{T_\#\mathbb{P}=\mathbb{Q}^\lambda} \int_{\mathcal{X}} c\big(x, T(x)\big) d\mathbb{P}(x).$$

Our Lemma 1 states that the minimizer T^λ of a regularized GAN problem is *always* an OT map between $\mathbb{P}$ and the distribution $\mathbb{Q}^\lambda$ generated by the same T^λ from $\mathbb{P}$. However, below we prove that $\mathbb{Q}^\lambda \neq \mathbb{Q}$, i.e., T^λ **does not** actually produce the distribution of HR images (Figure 4). To begin with, we state and prove the following auxiliary result.

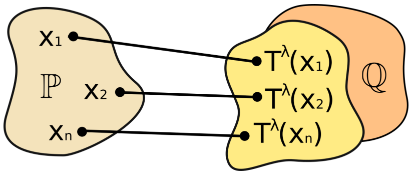


Figure 4: Illustration of Lemma 1. The solution T^λ of (2) is an OT map from $\mathbb{P}$ to $T^\lambda_\#\mathbb{P}$. In general, $T^\lambda_\#\mathbb{P} \neq \mathbb{Q}$ (Thm. 1).

Lemma 2 (Reformulation of the regularized GAN via distributions)**.** *Under the assumptions of Lemma 1, let $\mathcal{X} = \mathcal{Y}$ be a compact subset of $\mathbb{R}^D$ with negligible boundary. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ be absolutely continuous, $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and $c(x,y) = \|x-y\|^p$ with $p > 1$. Then* (2) *is equivalent to*

$$\inf_{\mathbb{Q}' \in \mathcal{P}(\mathcal{Y})} \mathcal{F}(\mathbb{Q}') \overset{\text{def}}{=} \inf_{\mathbb{Q}' \in \mathcal{P}(\mathcal{Y})} \big[\mathcal{D}(\mathbb{Q}', \mathbb{Q}) + \lambda \cdot \text{Cost}(\mathbb{P}, \mathbb{Q}')\big], \tag{6}$$

and the solutions of (2) *and* (6) *are related as $\mathbb{Q}^\lambda = T^\lambda_\#\mathbb{P}$, where $\mathbb{Q}^\lambda$ is the minimizer of* (6)*.*

In the following Theorem, we prove that, in general, $\mathbb{Q}^\lambda \neq \mathbb{Q}$ for the minimizer $\mathbb{Q}^\lambda$ of (6).

Theorem 1 (The distribution solving the regularized GAN problem is always biased)**.** *Under the assumptions of Lemma 2, assume that the first variation (Santambrogio, 2015,* Definition 7.12*) of the functional $\mathbb{Q}' \mapsto \mathcal{D}(\mathbb{Q}', \mathbb{Q})$ at the point $\mathbb{Q}' = \mathbb{Q}$ exists and is equal to zero. This means that $\mathcal{D}(\mathbb{Q} + \epsilon\Delta\mathbb{Q}, \mathbb{Q}) = \mathcal{D}(\mathbb{Q}, \mathbb{Q}) + o(\epsilon)$ for every signed measure $\Delta\mathbb{Q}$ of zero total mass and $\epsilon \geq 0$ such that $\mathbb{Q} + \epsilon\Delta\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$. Then, if $\mathbb{P} \neq \mathbb{Q}$, then $\mathbb{Q}' = \mathbb{Q}$ does not deliver the minimum to $\mathcal{F}$.*

Before proving Theorem 1, we highlight that the assumption about the vanishing first variation of $\mathbb{Q}' \mapsto \mathcal{D}(\mathbb{Q}', \mathbb{Q})$ at $\mathbb{Q}' = \mathbb{Q}$ is *reasonable*. In Appendix B, we prove that this assumption holds for the popular GAN discrepancies $\mathcal{D}(\mathbb{Q}', \mathbb{Q})$, e.g., f-divergences (Nowozin et al., 2016) and certain Wasserstein distances (Arjovsky et al., 2017).

Corollary 1. *Under the assumptions of Theorem 1, the solution T^λ of regularized GAN* (2) *is biased, i.e., it does not satisfy $T^\lambda_\#\mathbb{P} = \mathbb{Q}$ and* does not transform LR images to true HR ones.

Additionally, we provide a toy example that further illustrates the issue with the bias.

Example 1. *Consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^1$. Let $\mathbb{P} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2$, $\mathbb{Q} = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_3$ be distributions concentrated at $\{0, 2\}$ and $\{1, 3\}$, respectively. Put $c(x,y) = |x-y|$ to be the content loss. Also, let $\mathcal{D}$ to be the OT cost for $|x-y|^2$. Then for $\lambda = 0$ there exist two maps between $\mathbb{P}$ and $\mathbb{Q}$ that deliver the same minimal value for* (2)*, namely $T(0) = 1, T(2) = 3$ and $T(0) = 3, T(2) = 1$. For $\lambda > 0$, the optimal solution of the problem* (2) *is unique, **biased** and given by $T(0) = 1 - \frac{\lambda}{2}, T(2) = 3 - \frac{\lambda}{2}$.*

In Example 1, $T^\lambda_\#\mathbb{P} = \mathbb{Q}^\lambda$ *never* matches $\mathbb{Q}$ exactly for $\lambda > 0$. In §6.1, we conduct an evaluation of maps obtained via minimizing objective (2) on the synthetic benchmark by (Korotin et al., 2021). We empirically demonstrate that the bias exists and it is indeed a notable practical issue.

Remarks. Throughout this section, we enforce additional assumptions on (2), e.g., we restrict our analysis to content losses $c(\cdot,\cdot)$, which are powers of Euclidean norms $\|\cdot\|^p$. This is needed to make the derivations concise and to be able to exploit the available results in OT. We think that the provided results hold under more general assumptions and leave this question open for future studies.

5 Relation between GANs and Neural Optimal Transport Solvers

In this section, we analyze recent neural algorithms to compute OT maps (Fan et al., 2023; Korotin et al., 2023b; Rout et al., 2022) and show their connection with regularized GANs. Below we show that their loss can be viewed as a particular (in a certain sense) GAN objective regularized with the content loss. To begin with, we recall that typical OT optimization objective is minimax and given by

$$[\text{Cost}(\mathbb{P},\mathbb{Q}) =] \qquad \sup_{f} \inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \Big[\int_{\mathcal{Y}} f(y)d\mathbb{Q}(y) + \int_{\mathcal{X}} \big\{ c\big(x,T(x)\big) - f(T(x))\big\}\, d\mathbb{P}(x)\Big], \tag{7}$$

where $\sup_f$ is taken w.r.t. all potentials $f \in \mathcal{L}^1(\mathbb{Q})$. Under mild assumptions[1], by solving (7) one may recover the true (*unbiased*) OT map T^*, see (Korotin et al., 2023b, Lemma 4), (Fan et al., 2023, §3,4). In practice, T, f are replaced with neural networks; as in GANs, they are optimized with the stochastic gradient descent-ascent techniques using the empirical samples from $\mathbb{P}, \mathbb{Q}$.

Now let us get back to GANs. In §4, we show that solutions of (2) are, in general, *biased* OT maps. Note, that this bias is related to the trade-off between components of GANs optimization objective (2), i.e., the quality of generated image and its similarity to the input. In order to resolve the bias issue, one can consider the loss $\mathcal{D}(T_{\#}\mathbb{P},\mathbb{Q}) \equiv \mathcal{I}(T_{\#}\mathbb{P},\mathbb{Q})$ where $\mathcal{I}$ is the indicator function which takes two values: zero if its inputs coincide and $+\infty$ when they differ. Then we can rewrite (2) as

$$\lambda \cdot \inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \big[\frac{1}{\lambda}\mathcal{I}(T_{\#}\mathbb{P},\mathbb{Q}) + \mathcal{R}_c(T)\big] = \lambda \cdot \inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \big[\mathcal{I}(T_{\#}\mathbb{P},\mathbb{Q}) + \mathcal{R}_c(T)\big]. \tag{8}$$

Here we used the fact that $\lambda \cdot \mathcal{I}(\cdot,\cdot) = \mathcal{I}(\cdot,\cdot)$. Note that the solution $\widehat{T}$ (if it exists) of (8) satisfies $\widehat{T}_{\#}\mathbb{P} = \mathbb{Q}$. Otherwise, the objective yields the value $+\infty$. Therefore, problem (8) is equivalent to the optimization of the functional $\mathcal{R}_c(T)$ with the constraint $T_{\#}\mathbb{P} = \mathbb{Q}$. As a result, (8) turns to be just the Monge OT problem (3) multiplied by $\lambda > 0$ and with the constraint incorporated directly to the loss via the indicator function $\mathcal{I}$. We conclude that its solution is an OT map, i.e., $\widehat{T} = T^*$, and the optimal value of (8) is exactly $\lambda \cdot \text{Cost}(T_{\#}\mathbb{P},\mathbb{Q})$.

Unfortunately, optimizing objective (8) in practice is non-trivial: even testing the condition $T_{\#}\mathbb{P} = \mathbb{Q}$ (i.e., computing $\mathcal{I}$) is hard, which makes it challenging to compute the loss. Note that

$$\mathcal{I}(T_{\#}\mathbb{P},\mathbb{Q}) = \sup_{f} \big[-\int_{\mathcal{X}} f(T(x))d\mathbb{P}(x) + \int_{\mathcal{Y}} f(y)d\mathbb{Q}(y)\big], \tag{9}$$

where f skims through all integrable w.r.t. $\mathbb{Q}$ and $T_{\#}\mathbb{P}$ functions. Indeed, if $T_{\#}\mathbb{P} = \mathbb{Q}$, the two integrals always coincide. Otherwise, there always exists a measurable function f whose integrals over distributions differ. One may then multiply it by an arbitrary number to get any value of the expression, i.e., in this case, sup equals $+\infty$. We substitute (9) to (8) multiplied by $\frac{1}{\lambda}$ and get

$$\text{Cost}(\mathbb{P},\mathbb{Q}) = \inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \mathcal{I}(T_{\#}\mathbb{P},\mathbb{Q}) + \mathcal{R}_c(T) =$$
$$\inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \sup_{f} \big[\int_{y\in\mathcal{Y}} f(y)d\mathbb{Q}(y) + \int_{x\in\mathcal{X}} \big\{c\big(x,T(x)\big) - f(T(x))\big\}\, d\mathbb{P}(x)\big] \tag{10}$$

which almost coincides with (7); the only difference is the order of inf and sup. At this point, a natural question arises: what is the conceptual difference between (7) and (10), and why neural OT works typically consider (7) rather than (10)? We believe that this is simply because the loss for the Neural OT methods is usually derived from the conventional dual formulation of OT (5) by expressing the c-transform, which yields the additional inner problem. In fact, when it comes to the practical optimization of (7) or (10), the actual order of optimization does not matter too much. The overall performance depends more on a proper choice of hyperparameters of the optimization.

5.1 Regularized GANs vs. Optimal Transport Solver

In this subsection, we discuss similarities and differences between neural OT optimization objective (7) and the objective of regularized GANs (2). We establish an intriguing connection between GANs that use *integral probability metrics* (IPMs) as $\mathcal{D}$. A discrepancy $\mathcal{D}:\mathcal{P}(\mathcal{Y})\times\mathcal{P}(\mathcal{Y})\rightarrow\mathbb{R}_+$ is an IPM if

[1] In certain cases among the solutions in such a problem may be so-called *fake* solutions which are not the OT maps. We refer to Korotin et al. (2023a) for a fruitful discussion of this phenomena.

$$\mathcal{D}(\mathbb{Q}_1,\mathbb{Q}_2)=\sup_{f\in\mathcal{F}}\Big[\int_{\mathcal{Y}} f(y)d\mathbb{Q}_2(y)-\int_{\mathcal{Y}} f(y)d\mathbb{Q}_1(y)\Big], \tag{11}$$

where the maximization is performed over some certain class $\mathcal{F}$ of functions (discriminators) $f:\mathcal{Y}\rightarrow\mathbb{R}$. The most popular example of $\mathcal{D}$ is the Wasserstein-1 loss (Arjovsky & Bottou, 2017), where $\mathcal{F}$ is a class of 1-Lipschitz functions. For other IPMs, see (Mroueh et al., 2017, Table 1).

Substituting (11) to (2) yields the saddle-point optimization problem for the **regularized IPM GAN**:

$$\inf_{T:\mathcal{X}\rightarrow\mathcal{Y}}\left[\sup_{f\in\mathcal{F}}\left\{\int_{\mathcal{Y}} f(y)d\mathbb{Q}(y)-\int_{\mathcal{X}} f\big(T(x)\big)d\mathbb{P}(x)\right\}+\lambda\int_{\mathcal{X}} c\big(x,T(x)\big)d\mathbb{P}(x)\}\right]$$
$$=\inf_{T:\mathcal{X}\rightarrow\mathcal{Y}}\sup_{f\in\mathcal{F}}\left[\int_{\mathcal{Y}} f(y)d\mathbb{Q}(y)+\int_{\mathcal{X}}\left\{\lambda\cdot c\big(x,T(x)\big)-f\big(T(x)\big)\right\}d\mathbb{P}(x)\right]. \tag{12}$$

We emphasize that the expression inside (12) for $\lambda=1$ is similar to the expression in OTS optimization (7). Below we highlight the **key differences** between (7) and (12).

First, in OTS the optimization over potential f is unconstrained, while in IPM GAN it must belong to $\mathcal{F}$, some certain restricted class of functions. For example, when $\mathcal{D}$ is the Wasserstein-1 ($\mathbb{W}_1$) IPM, one has to use an additional penalization, e.g., the gradient penalty (Gulrajani et al., 2017). This further complicates the optimization and adds hyperparameters which have to be carefully selected.

Second, the optimization of IPM GAN requires selecting a parameter λ that balances the content loss $\mathcal{R}_c$ and the discrepancy $\mathcal{D}$. In OTS for all costs $\lambda\cdot c(x,y)$ with $\lambda>0$, the OT map T^* is the same.

To conclude, even for $\lambda=1$, the IPM GAN problem generally does not match that of OTS. Table 1 summarizes the differences and the similarities between OTS and regularized IPM GANs.

	Optimal Transport Solver (OTS)	**Regularized IPM GAN**
Minimax optimization objective	$\sup_{f}\inf_{T:\mathcal{X}\rightarrow\mathcal{Y}}\Big[\int_{\mathcal{Y}} f(y)d\mathbb{Q}(y)+$ $\int_{\mathcal{X}}\left\{c\big(x,T(x)\big)-f\big(T(x)\big)\right\}d\mathbb{P}(x)\Big]$	$\inf_{T:\mathcal{X}\rightarrow\mathcal{Y}}\sup_{f\in\mathcal{F}}\Big[\int_{\mathcal{Y}} f(y)d\mathbb{Q}(y)+$ $\int_{\mathcal{X}}\left\{\lambda\cdot c\big(x,T(x)\big)-f\big(T(x)\big)\right\}d\mathbb{P}(x)\Big]$
Potential f (discriminator)	Unconstrained $f\in L^1(\mathbb{Q})$	Constrained $f\in\mathcal{F}\subset L^1(\mathbb{Q})$ A method to impose the constraint is needed.
Regularization weight λ	N/A	Hyperparameter choice required

Table 1: Comparison of the optimization objectives of OTS and regularized IPM GAN.

6 Experimental Illustration

In §6.1, we assess the bias of regularized IPM GANs by using the Wasserstein-2 benchmark (Korotin et al., 2021). In §6.2, we evaluate OTS on the large-scale unpaired AIM-19 dataset from (Lugmayr et al., 2019b) and compare it with popular GAN-based solutions for unpaired image SR. The code is written in `PyTorch`. We list the hyperparameters for Algorithm 1 in Table 4 of Appendix C.

Neural network architectures. We use WGAN-QC's (Liu et al., 2019) ResNet (He et al., 2016) architecture for the potential f_ω. In §6.1, where input and output images have the same size, we use UNet[2] (Ronneberger et al., 2015) as a transport map T_θ. In §6.2, the LR input images are 4×4 times smaller than HR, so we use EDSR network (Lim et al., 2017).

Transport costs. In §6.1, we use the *mean squared error* (MSE), i.e., $c(x,y)=\frac{\|x-y\|^2}{\dim(\mathcal{Y})}$. It is equivalent to the quadratic cost but is more convenient due to the normalization. In §6.2, we consider $c(x,y)=b(\text{Up}(x),y)$, where b is a cost between the bicubically upsampled LR image $x^{\text{up}}=\text{Up}(x)$ and HR image y. We test b defined as MSE and the *perceptual cost* using features of a pre-trained VGG-16 network (Simonyan & Zisserman, 2014), see Appendix C for details.

6.1 Assessing the Bias in Regularized GANs

In this section, we empirically confirm the insight of §4 that the solution T^λ of (2) may not satisfy $T^\lambda_\#\mathbb{P}=\mathbb{Q}$. Notably, if $T^\lambda_\#\mathbb{P}=\mathbb{Q}$, then from our Lemma 1 it follows that $T^\lambda\equiv T^*$, where T^* is an

[2] github.com/milesial/Pytorch-UNet

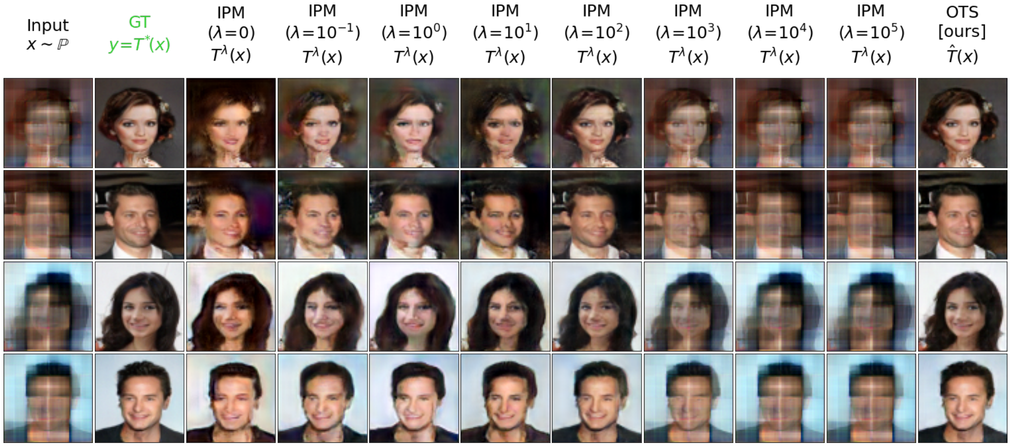


Figure 5: Comparison of OTS, regularized IPM GAN on the Wasserstein-2 benchmark. The 1st line shows blurry faces $x \sim \mathbb{P}$, the 2nd line, clean faces $y = T^*(x)$, where T^* is the OT map from $\mathbb{P}$ to $\mathbb{Q}$. Next lines show maps from $\mathbb{P}$ to $\mathbb{Q}$ fitted by the methods.

Metrics/ Method	**Regularized IPM GAN** (WGAN-GP, $\lambda_{\text{GP}} = 10$)								**OTS**
	$\lambda = 0$	$\lambda = 10^{-1}$	$\lambda = 10^{0}$	$\lambda = 10^{1}$	$\lambda = 10^{2}$	$\lambda = 10^{3}$	$\lambda = 10^{4}$	$\lambda = 10^{5}$	
$\mathcal{L}^2$-UVP ↓	25.2%	16.7%	17.7%	12.0%	**4.0%**	14.0%	28.5%	30.5%	**1.4%**
FID↓	57.24	46.23	40.04	42.89	**24.25**	187.95	332.7	334.7	**15.65**
PSNR↑	17.90	19.76	19.34	20.81	**25.58**	19.91	16.90	16.52	**30.02**
SSIM↑	0.565	0.655	0.656	0.689	**0.859**	0.702	0.520	0.498	**0.933**
LPIPS↓	0.135	0.093	0.099	0.081	**0.031**	0.172	0.429	0.446	**0.013**

Table 2: Quantitative evaluation of restoration maps fitted by the regularized IPM GAN, OTS using the Wasserstein-2 images benchmark (Korotin et al., 2021).

OT map from $\mathbb{P}$ to $\mathbb{Q}$ for $c(x, y)$. Thus, to access the bias, it is reasonable to compare the learned map T^λ with the ground truth OT map T^* for $\mathbb{P}$, $\mathbb{Q}$.

For evaluation, we use the Wasserstein-2 benchmark (Korotin et al., 2021). It provides high-dimensional continuous pairs $\mathbb{P}$, $\mathbb{Q}$ with an analytically known OT map T^* for the quadratic cost $c(x, y) = \|x - y\|^2$. We use their "Early" images benchmark pair. It simulates the image deblurring setup, i.e., $\mathcal{X} = \mathcal{Y}$ is the space of 64×64 RGB images, $\mathbb{P}$ is blurry faces, $\mathbb{Q}$ is clean faces satisfying $\mathbb{Q} = T^*_\# \mathbb{P}$, where T^* is an analytically known OT map, see the 1st and 2nd lines in Figure 5. To quantify the learned maps from $\mathbb{P}$ to $\mathbb{Q}$, we use PSNR, SSIM, LPIPS (Zhang et al., 2018a), FID (Heusel et al., 2017) metrics. Similar to (Wei et al., 2021), we use the AlexNet-based (Krizhevsky et al., 2012) LPIPS. FID and LPIPS are practically the *most important* since they better correlate with the human perception of the image quality. We include PSNR, SSIM as popular evaluation metrics, but they are known to *badly measure perceptual quality* (Zhang et al., 2018a; Nilsson & Akenine-Möller, 2020). Due to this, higher PSNR, SSIM values do not necessarily mean better performance. We calculate metrics using `scikit-image` for SSIM and open source implementations for PSNR[3], LPIPS[4] and FID[5]. In this section, we additionally use the $\mathcal{L}^2$-UVP (Korotin et al., 2021, §4.2) metric.

On the benchmark, we compare OTS (7) and IPM GAN (2). We use MSE as the content loss $c(x, y)$. In IPM GAN, we use the Wasserstein-1 ($\mathbb{W}_1$) loss with the gradient penalty $\lambda_{\text{GP}} = 10$ (Gulrajani et al., 2017) as $\mathcal{D}$. We do 10 discriminator updates per 1 generator update and train the model for 15K generator updates. For fair comparison, the rest hyperparameters match those of OTS algorithm. We train the regularized WGAN-GP with various coefficients of content loss $\lambda \in \{0, 10^{-1}, \ldots, 10^5\}$ and show the learned maps T^λ and the map $\hat{T}$ obtained by OTS in Figure 5.

Results. The performance of the regularized IPM GAN *significantly* depends on the choice of the content loss value λ. For high values $\lambda \geq 10^3$, the learned map is close to the identity as expected. For small values $\lambda \leq 10^1$, the regularization has little effect, and WGAN-GP solely struggles to fit a good restoration map. Even for the best performing $\lambda = 10^2$ all metrics are notably worse than for OTS. Importantly, *OTS decreases the burden of parameter searching* as there is no parameter λ.

[3] `github.com/photosynthesis-team/piq`

[4] `github.com/richzhang/PerceptualSimilarity`

[5] `github.com/mseitzer/pytorch-fid`

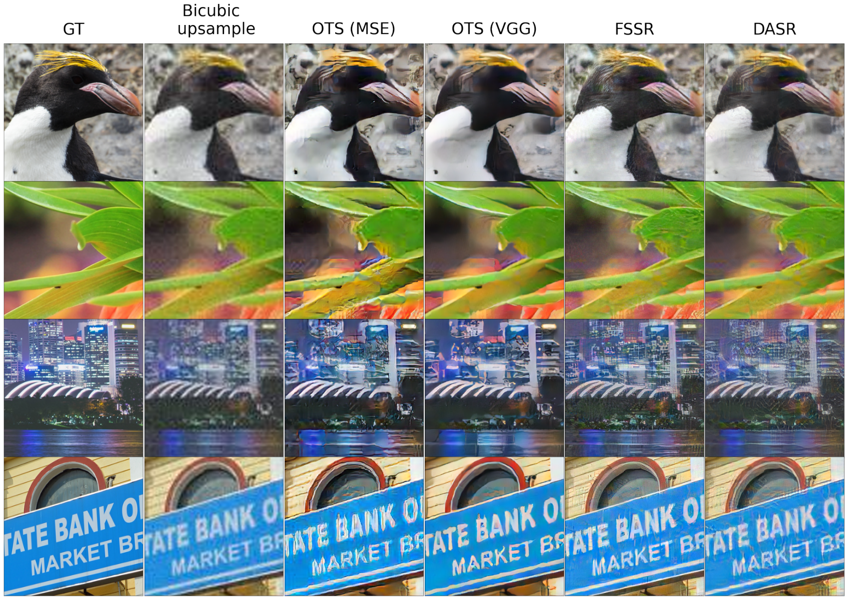


Figure 6: Qualitative results of OTS, bicubic upsample, FSSR and DASR on AIM 2019 dataset (350×350 crops).

6.2 LARGE-SCALE EVALUATION

For evaluating OTS method at a large-scale, we employ the dataset by (Lugmayr et al., 2019b) of AIM 2019 Real-World Super-Resolution Challenge (Track 2). The train part contains 800 HR images with up to 2040 pixels width or height and 2650 unpaired LR images of the same shape. They are constructed using artificial, but realistic, image degradations. We quantitatively evaluate OTS method on the validation part of AIM dataset that contains 100 pairs of LR-HR images.

Baselines. We compare OTS on AIM dataset with the bicubic upsample, FSSR (Fritsche et al., 2019) and DASR (Wei et al., 2021) methods. FSSR method is the winner of AIM 2019 Challenge; DASR is a current state-of-the-art method for unpaired image SR. Both methods utilize the idea of frequency separation and solve the problem in two steps. First, they train a network to generate LR images. Next, they train a super-resolution network using generated pseudo-pairs. Differently to FSSR, DASR also employs real-world LR images for training SR network taking into consideration the domain gap between generated and real-world LR images. Both methods utilize several losses, e.g., adversarial and perceptual, either on the entire image or on its high/low frequency components. For testing FSSR and DASR, we use their official code and pretrained models.

Implementation details. We train the networks using 128×128 HR, 32×32 LR random *patches* of images augmented via random flips, rotations. We conduct separate experiments using EDSR as the transport map and either MSE or perceptual cost, and denote them as OTS (MSE), OTS (VGG) respectively.

Metrics. We calculate PSNR, SSIM, LPIPS, FID. FID is computed on 32×32 patches of LR test images upsampled by the method in view w.r.t. random patches of test HR. We use 50k patches to compute FID. The other metrics are computed on the *entire* upsampled LR test and HR test images.

Experimental results are given in Table 3, Figure 6. The results show that the usage perceptual cost function in OTS boosts performance. According to FID, OTS with perceptual cost function beats DASR. On the other hand, it outperforms FSSR in PSNR, SSIM and, importantly, LPIPS. Note that bicubic upsample outperforms all the methods, according only to PSNR and SSIM, which have issues stated in §6.1. According to visual analysis, OTS with the perceptual cost better deals with noise artifacts. Additional results are given in Appendix F. We also demonstrate the bias issue of FSSR and DASR in Appendix D.

Method	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic upsample	178.59	22.39	0.613	0.688
OTS (MSE)	139.17	19.73	0.533	0.456
OTS (VGG)	89.04	20.96	0.605	0.380
FSSR	53.92	20.83	0.514	0.390
DASR	124.09	21.79	0.577	0.346

Table 3: Comparison of OTS with FSSR, DASR on AIM19 dataset. The 1st, 2nd, 3rd best results are highlighted in green, blue and underlined, respectively.

7 Conclusion

Our analysis connects content losses in GANs with OT and reveals the bias issue. Content losses are used in a wide range of tasks besides SR, e.g., in the style transfer and domain adaptation tasks. Our results demonstrate that GAN-based methods in all these tasks may *a priori lead to biased solutions*. In certain cases it is undesirable, e.g., in medical applications (Bissoto et al., 2021). Failing to learn true data statistics (and learning biased ones instead), e.g., in the super-resolution of MRI images, might lead to a wrong diagnosis made by a doctor due to SR algorithm drawing **inexistent details** on the scan. Thus, we think it is essential to emphasize and alleviate the bias issue.

8 Acknowledgements

The work of Skoltech was supported by the Analytical center under the RF Government (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021).

References

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

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.

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1847–1856, 2021.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. *European Conference on Computer Vision*, 09 2018.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=2mZSlQscj3`. Featured Certification.

Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Leonid Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

Gwantae Kim, Jaihyun Park, Kanghyu Lee, Junyeop Lee, Jeongki Min, Bokyeung Lee, David K. Han, and Hanseok Ko. Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1862–1871, June 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34, 2021.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Kernel neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023a.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 624–632, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein GAN with quadratic transport cost. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4832–4841, 2019.

Guansong Lu, Zhiming Zhou, Jian Shen, Cheng Chen, Weinan Zhang, and Yong Yu. Large-scale optimal transport via adversarial training with cycle-consistency. *arXiv preprint arXiv:2003.06635*, 2020.

Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3408–3416, 2019a.

Andreas Lugmayr, Martin Danelljan, Radu Timofte, Manuel Fritsche, Shuhang Gu, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, AN Rajagoapalan, Nam Hyung Joon, et al. Aim 2019 challenge on real-world image super-resolution: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3575–3583. IEEE, 2019b.

Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 291–300, 2020.

Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q, p)-Wasserstein GANs: Comparing ground metrics for Wasserstein GANs. *arXiv preprint arXiv:1902.03642*, 2019.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=5JdLZg346Lw`.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wei Wang, Haochen Zhang, Zehuan Yuan, and Changhu Wang. Unsupervised real-world super-resolution: A domain adaptation perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4318–4327, 2021.

Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13385–13394, June 2021.

Yujia Xie, Minshuo Chen, Haoming Jiang, Tuo Zhao, and Hongyuan Zha. On scalable and efficient computation of large scale optimal transport. volume 97 of *Proceedings of Machine Learning Research*, pp. 6882–6892, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/xie19a.html`.

Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yong bing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 814–81409, 2018.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018a.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018b.

Yuanbo Zhou, Wei Deng, Tong Tong, and Qinquan Gao. Guided frequency separation network for real-world super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 428–429, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A Proofs

Proof of Lemma 1. Assume that T^{λ} is not an optimal map between $\mathbb{P}$ and $T^{\lambda}_{\#}\mathbb{P}$. Then there exists a more optimal $T^{\dagger}$ satisfying $T^{\dagger}_{\#}\mathbb{P} = T^{\lambda}_{\#}\mathbb{P}$ and $\mathcal{R}_c(T^{\dagger}) < \mathcal{R}_c(T^{\lambda})$. We substitute this $T^{\dagger}$ to (2) and derive

$$\mathcal{D}(T^{\dagger}_{\#}\mathbb{P}, \mathbb{Q}) + \lambda\mathcal{R}_c(T^{\dagger}) = \mathcal{D}(T^{\lambda}_{\#}\mathbb{P}, \mathbb{Q}) + \lambda\mathcal{R}_c(T^{\dagger}) < \mathcal{D}(T^{\lambda}_{\#}\mathbb{P}, \mathbb{Q}) + \lambda\mathcal{R}_c(T^{\lambda}),$$

which is a contradiction, since T^{λ} is a minimizer of (2), but $T^{\dagger}$ provides the smaller value. □

Proof of Lemma 2. We derive

$$\inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \big[\mathcal{D}(T_{\#}\mathbb{P}, \mathbb{Q}) + \lambda\mathcal{R}_c(T)\big] = \inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \big[\mathcal{D}(T_{\#}\mathbb{P}, \mathbb{Q}) + \lambda \int_{\mathcal{X}} c\big(x, T(x)\big) d\mathbb{P}(x)\big] = \tag{13}$$

$$\inf_{T:\mathcal{X}\mapsto\mathcal{Y}} \big[\mathcal{D}(T_{\#}\mathbb{P}, \mathbb{Q}) + \lambda \cdot \text{Cost}(\mathbb{P}, T_{\#}\mathbb{P})\big] = \inf_{\mathbb{Q}'\in\mathcal{P}(\mathcal{Y})} \big[\mathcal{D}(\mathbb{Q}', \mathbb{Q}) + \lambda \cdot \text{Cost}(\mathbb{P}, \mathbb{Q}')\big]. \tag{14}$$

In transition from (13) to (14), we use the definition of OT cost (3) and our Lemma 1, which states that the minimizer T^{λ} of (2) is an OT map, i.e., $\int_{\mathcal{X}} c\big(x, T^{\lambda}(x)\big) d\mathbb{P}(x) = \text{Cost}(\mathbb{P}, T^{\lambda}_{\#}\mathbb{P})$. The equality in (14) follows from the fact that $\mathbb{P}$ is abs. cont. and $c(x, y) = \|x - y\|^p$: for all $\mathbb{Q}' \in \mathcal{P}(\mathcal{Y})$ there exists a (unique) solution T to the Monge OT problem (3) for $\mathbb{P}, \mathbb{Q}'$ (Santambrogio, 2015, Thm. 1.17). □

Proof of Theorem 1. Let $\Delta\mathbb{Q} = \mathbb{P} - \mathbb{Q}$ denote the difference measure of $\mathbb{P}$ and $\mathbb{Q}$. It has zero total mass and $\forall\epsilon \in [0, 1]$ it holds that $\mathbb{Q} + \epsilon\Delta\mathbb{Q} = \epsilon\mathbb{P} + (1 - \epsilon)\mathbb{Q}$ is a mixture distribution of probability distributions $\mathbb{P}$ and $\mathbb{Q}$. As a result, for all $\epsilon \in [0, 1]$, we have

$$\mathcal{F}(\mathbb{Q} + \epsilon\Delta\mathbb{Q}) = \mathcal{D}(\mathbb{Q} + \epsilon\Delta\mathbb{Q}, \mathbb{Q}) + \lambda \cdot \text{Cost}(\mathbb{P}, \mathbb{Q} + \epsilon\Delta\mathbb{Q}) =$$

$$\mathcal{D}(\mathbb{Q}, \mathbb{Q}) + o(\epsilon) + \lambda \cdot \text{Cost}(\mathbb{P}, \epsilon\mathbb{P} + (1 - \epsilon)\mathbb{Q}) \leq \tag{15}$$

$$o(\epsilon) + \lambda \cdot \epsilon \cdot \text{Cost}(\mathbb{P}, \mathbb{P}) + \lambda \cdot (1 - \epsilon) \cdot \text{Cost}(\mathbb{P}, \mathbb{Q}) = o(\epsilon) + \lambda \cdot (1 - \epsilon) \cdot \text{Cost}(\mathbb{P}, \mathbb{Q}) = \tag{16}$$

$$\underbrace{\lambda \cdot \text{Cost}(\mathbb{P}, \mathbb{Q})}_{=\mathcal{F}(\mathbb{Q})} - \lambda \cdot \epsilon \cdot \underbrace{\text{Cost}(\mathbb{P}, \mathbb{Q})}_{>0} + o(\epsilon),$$

where in transition from (15) to (16), we use $\mathcal{D}(\mathbb{Q}, \mathbb{Q}) = 0$ and exploit the convexity of the OT cost (Villani, 2003, Theorem 4.8). In (16), we use $\text{Cost}(\mathbb{P}, \mathbb{P}) = 0$. We see that $\mathcal{F}(\mathbb{Q}+\epsilon\Delta\mathbb{Q})$ is smaller then $\mathcal{F}(\mathbb{Q})$ for sufficiently small $\epsilon > 0$, i.e., $\mathbb{Q}'=\mathbb{Q}$ does not minimize $\mathcal{F}$. □

Proof of Example 1. Let $T(0) = t_0$ and $T(2) = t_2$. Then $T_{\#}\mathbb{P} = \frac{1}{2}\delta_{t_0} + \frac{1}{2}\delta_{t_2}$, and now (2) becomes

$$\min_{t_0,t_2} \Big[\min\big\{\frac{1}{2}(t_0 - 1)^2 + \frac{1}{2}(t_2 - 3)^2; \frac{1}{2}(t_0 - 3)^2 + \frac{1}{2}(t_2 - 1)^2\big\} + \lambda\big\{\frac{1}{2}|0 - t_0| + \frac{1}{2}|2 - t_2|\big\}\Big],$$

where the second term is $\mathcal{R}_c(T)$ and the first term is the OT cost $\mathcal{D}(T_{\#}\mathbb{P}, \mathbb{Q})$ expressed as the minimum over the transport costs of two possible transport maps $t_0 \mapsto 1; t_2 \mapsto 3$ and $t_0 \mapsto 3; t_2 \mapsto 1$. The minimizer can be derived analytically and equals $t_0 = 1 - \frac{\lambda}{2}, t_2 = 3 - \frac{\lambda}{2}$. □

B First Variations of GAN Discrepancies Vanish at the Optimum

We demonstrate that the first variation of $\mathbb{Q}' \mapsto \mathcal{D}(\mathbb{Q}', \mathbb{Q})$ is equal to zero at $\mathbb{Q}' = \mathbb{Q}$ for common GAN discrepancies $\mathcal{D}$. This suggests that the corresponding assumption of our Theorem 1 is relevant.

To begin with, for a functional $\mathcal{G} : \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{\infty\}$, we recall the definition of its **first variation**. A measurable function $\delta\mathcal{G}[\mathbb{Q}] : \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ is called **the first variation** of $\mathcal{G}$ at a point $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$, if, for every measure $\Delta\mathbb{Q}$ on $\mathcal{Y}$ with zero total mass ($\int_{\mathcal{Y}} 1\, d\Delta\mathbb{Q}(y) = 0$),

$$\mathcal{G}(\mathbb{Q} + \epsilon\Delta\mathbb{Q}) = \mathcal{G}(\mathbb{Q}) + \epsilon \int_{\mathcal{Y}} \delta\mathcal{G}[\mathbb{Q}](y)\, d\Delta\mathbb{Q}(y) + o(\epsilon) \tag{17}$$

for all $\epsilon \geq 0$ such that $\mathbb{Q} + \epsilon\Delta\mathbb{Q}$ is a probability distribution. Here for the sake of simplicity we suppressed several minor technical aspects, see (Santambrogio, 2015, Definition 7.12) for details. Note that the first variation is defined **up to an additive constant**.

Now we recall the definitions of three most popular GAN discrepancies and demonstrate that their first variation is zero at an optimal point. We consider f-divergences (Nowozin et al., 2016), Wasserstein distances (Arjovsky et al., 2017).

Case 1 (f-divergence). Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex and differentiable function satisfying $f(1) = 0$. The f-divergence between $\mathbb{Q}', \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ is defined by

$$\mathcal{D}_f(\mathbb{Q}', \mathbb{Q}) \overset{def}{=} \int_{\mathcal{Y}} f\left(\frac{d\mathbb{Q}'(y)}{d\mathbb{Q}(y)}\right) d\mathbb{Q}(y). \tag{18}$$

The divergence takes finite value only if $\mathbb{Q}' \ll \mathbb{Q}$, i.e., $\mathbb{Q}'$ is absolutely continuous w.r.t. $\mathbb{Q}$. Vanilla GAN loss (Goodfellow et al., 2014) is a case of f-divergence (Nowozin et al., 2016, Table 1).

We define $\mathcal{G}(\mathbb{Q}') \overset{def}{=} \mathcal{D}_f(\mathbb{Q}', \mathbb{Q})$. For $\mathbb{Q}' = \mathbb{Q}$ and some $\Delta\mathbb{Q}$ such that $\mathbb{Q} + \epsilon\Delta\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ we derive

$$\mathcal{G}(\mathbb{Q} + \epsilon\Delta\mathbb{Q}) = \int_{\mathcal{Y}} f\left(\frac{d\mathbb{Q}(y)}{d\mathbb{Q}(y)} + \epsilon\frac{d\Delta\mathbb{Q}(y)}{d\mathbb{Q}(y)}\right) d\mathbb{Q}(y) = \int_{\mathcal{Y}} f\left(1 + \epsilon\frac{d\Delta\mathbb{Q}(y)}{d\mathbb{Q}(y)}\right) d\mathbb{Q}(y) \tag{19}$$

$$= \int_{\mathcal{Y}} f(1) d\mathbb{Q}(y) + \int_{\mathcal{Y}} f'(1)\frac{d\Delta\mathbb{Q}(y)}{d\mathbb{Q}(y)} d\mathbb{Q}(y) + o(\epsilon) = \mathcal{G}(\mathbb{Q}) + \int_{\mathcal{Y}} f'(1) d\Delta\mathbb{Q}(y) + o(\epsilon), \tag{20}$$

where in transition from (19) to (20), we consider the Taylor series w.r.t. ϵ at $\epsilon = 0$. We see that $\delta\mathcal{G}[\mathbb{Q}](y) \equiv f'(1)$ is constant, i.e., the first variation of $\mathbb{Q}' \mapsto \mathcal{D}_f(\mathbb{Q}', \mathbb{Q})$ vanishes at $\mathbb{Q}' = \mathbb{Q}$.

Case 2 (Wasserstein distance). If in OT formulation (4) the cost function $c(x, y)$ equals $\|x - y\|^p$ with $p \geq 1$, then $\left[\text{Cost}(\mathbb{P}, \mathbb{Q})\right]^{1/p}$ is called the *Wasserstein distance* ($\mathbb{W}_p$). Generative models which use $\mathbb{W}_p^p$ as the discrepancy are typically called the Wasserstein GANs (WGANs). The most popular case is $p = 1$ (Arjovsky et al., 2017; Gulrajani et al., 2017), but more general cases appear in related work as well, see (Liu et al., 2019; Mallasto et al., 2019).

The first variation of $\mathcal{G}(\mathbb{Q}') \overset{\text{def}}{=} \mathbb{W}_p^p(\mathbb{Q}', \mathbb{Q})$ at a point $\mathbb{Q}'$ is given by $\mathcal{G}[\mathbb{Q}'](y) = (f^*)^c(y)$, where f^* is the optimal dual potential (provided it is unique up to a constant) in (5) for a pair $(\mathbb{Q}', \mathbb{Q})$, see (Santambrogio, 2015, §7.2). Our particular interest is to compute the optimal potential $(f^*)^c$ at $\mathbb{Q}' = \mathbb{Q}$. We recall (5) and use $\mathbb{W}_p^p(\mathbb{Q}, \mathbb{Q}) = 0$ to derive

$$\mathbb{W}_p^p(\mathbb{Q}, \mathbb{Q}) = 0 = \sup_f \left[\int_{\mathcal{X}} f^c(y') d\mathbb{Q}'(y') + \int_{\mathcal{Y}} f(y) d\mathbb{Q}(y)\right].$$

One may see that $f^* \equiv 0$ attains the supremum (its c-transform $(f^*)^c$ is also zero). Thus, **if** $(f^*)^c \equiv 0$ **is a unique potential** (up to a constant), the first variation of $\mathbb{Q}' \mapsto \mathbb{W}_p^p(\mathbb{Q}', \mathbb{Q})$ at $\mathbb{Q}' = \mathbb{Q}$ vanishes.

C Training Details

The practical optimization procedure of **Optimal Transport Solver** (OTS) is detailed in Algorithm 1.

Perceptual cost. In 6.2 we test following *perceptual cost* as b:

$$b(x^{\text{up}}, y) = \text{MSE}(x^{\text{up}}, y) + 1/3 \cdot \text{MAE}(x^{\text{up}}, y) + 1/50 \cdot \sum_{k \in \{3,8,15,22\}} \text{MSE}\big(f_k(x^{\text{up}}), f_k(y)\big),$$

where f_k denotes the features of the kth layer of a pre-trained VGG-16 network (Simonyan & Zisserman, 2014), MAE is the mean absolute error $\text{MAE}(x, y) = \frac{\|x-y\|_1}{\dim(\mathcal{Y})}$.

Dynamic transport cost. In the preliminary experiments, we used bicubic upsampling as the "Up" operation. Later, we found that the method works better if we gradually change the upsampling. We start from the bicubic upsampling. Every k_c iterations of f_ω (see Table 4), we change the cost to $c(x, y) = b\big(T'_\theta(x), y\big)$, where T'_θ is a fixed frozen copy of the currently learned SR map T_θ.

Hyperparameters. For EDSR, we set the number of residual blocks to 64, the number of features to 128, and the residual scaling to 1. For UNet, we set the base factor to 64. The training details are given in Table 4. We provide a comparison of the hyperparameters of FSSR, DASR and OTS in Table 5. In contrast to FSSR and DASR, OTS method does not contain a degradation part. This helps to notably reduce the amount of tunable hyperparameters.

Optimizer. We employ Adam (Kingma & Ba, 2014).

Computational complexity. Training OTS with EDSR as the transport map and the perceptual transport cost on AIM 2019 dataset takes ≈ 4 days on a single Tesla V100 GPU.

Algorithm 1: OT solver to compute the OT map between $\mathbb{P}$ and $\mathbb{Q}$ for transport cost $c(x, y)$.

```
Input  : distributions $\mathbb{P}, \mathbb{Q}$ accessible by samples; mapping network $T_\theta : \mathcal{X} \to \mathcal{Y}$;
         potential $f_\omega : \mathcal{X} \to \mathbb{R}$; transport cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; number $K_T$ of inner iters;
Output : approximate OT map $(T_\theta)_\# \mathbb{P} \approx \mathbb{Q}$;
repeat
    Sample batches $X \sim \mathbb{P}$, $Y \sim \mathbb{Q}$;
    $\mathcal{L}_f \leftarrow \frac{1}{|Y|} \sum_{y \in Y} f_\omega(y) - \frac{1}{|X|} \sum_{x \in X} f_\omega\big(T_\theta(x)\big)$;
    Update $\omega$ by using $\frac{\partial \mathcal{L}_f}{\partial \omega}$ to maximize $\mathcal{L}_f$;
    for $k_T = 1, 2, \dots, K_T$ do
        Sample batch $X \sim \mathbb{P}$;
        $\mathcal{L}_T \leftarrow \frac{1}{|X|} \sum_{x \in X} \Big[c\big(x, T_\theta(x)\big) - f_\omega\big(T_\theta(x)\big)\Big]$;
        Update $\theta$ by using $\frac{\partial \mathcal{L}_T}{\partial \theta}$ to minimize $\mathcal{L}_T$;
until not converged;
```

Experiment	dim($\mathcal{X}$)	dim($\mathcal{Y}$)	f	T	k_T	lr_f	lr_T	**Initial cost**	**Total iters** (f)	**Cost update every**	**Batch size**
Benchmark (§6.1)	$3 \times 64 \times 64$	$3 \times 64 \times 64$	ResNet	UNet	10	10^{-4}	10^{-4}	MSE	10K	–	64
AIM-19 (§6.2)	$3 \times 32 \times 32$ (patches)	$3 \times 128 \times 128$ (patches)		EDSR	15			Bicubic + MSE	50K	25K	8
				EDSR	10			Bicubic + VGG	50K	20K	8

Table 4: Hyperparameters that we use in the experiments with OTS Algorithm 1.

Method	**Degradation part**	**Super-resolution part**	**Total**

FSSR	2 neural networks;
	2 optimizers;
	2 schedulers;
	1 adversarial loss;
	1 content loss (ℓ_1+perceptual)
		2 neural networks;
		2 optimizers;
		2 schedulers;
		1 adversarial loss;
		1 content loss (ℓ_1+perceptual)
			4 neural networks;
			4 optimizers;
			4 schedulers;
			2 adversarial losses;
			2 content losses (ℓ_1+perceptual)

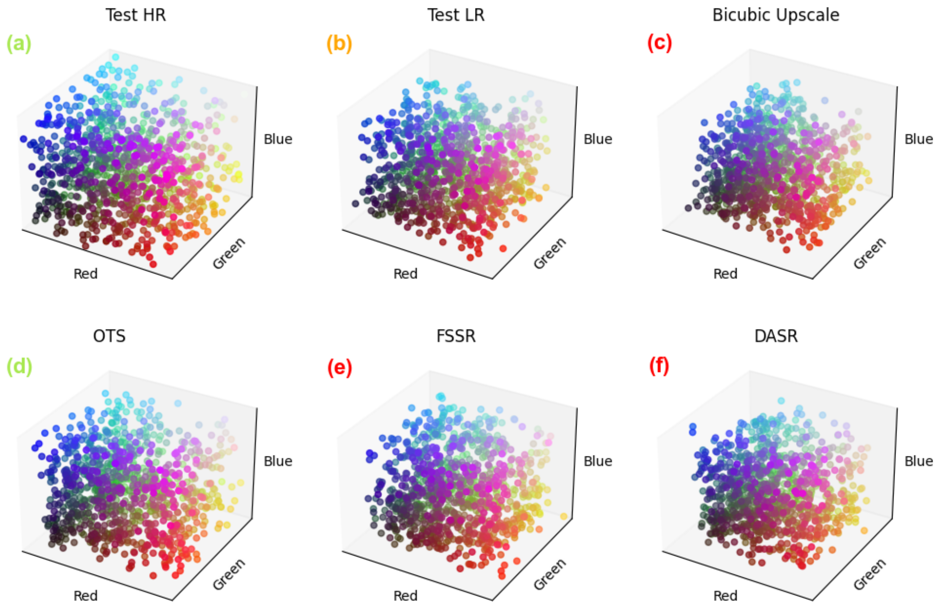
DASR	2 neural networks;
	2 optimizers;
	2 schedulers;
	1 adversarial loss;
	1 content loss (ℓ_1+perceptual)
		2 neural networks;
		2 optimizers;
		2 schedulers;
		1 adversarial loss;
		1 content loss (ℓ_1+perceptual)
			4 neural networks;
			4 optimizers;
			4 schedulers;
			2 adversarial losses;
			2 content losses (ℓ_1+perceptual)

OTS	–	2 neural networks;
		2 optimizers;
		1 cost (ℓ_2+ℓ_1+perceptual)
			2 neural networks;
			2 optimizers;
			1 cost (ℓ_2+ℓ_1+perceptual)

Table 5: Comparison of hyperparameters used in FSSR, DASR and OTS methods.

D ASSESSING THE BIAS OF METHODS ON AIM19 DATASET

We additionally demonstrate the bias issue by comparing color palettes of HR images and super-resolution results of different methods, see Figure 7. We construct palettes by choosing random image pixels from dataset images and representing them as an RGB point cloud in $[0, 1]^3 \subset \mathbb{R}^3$. Figure 7 shows that OTS (d) captures *large contrast* of HR (a) images (variance of its palette), while FSSR (e), DASR (f), Bicubic Upscale (c) palettes are *less contrastive* and closer to LR (b). We construct palettes 100 times to evaluate their average contrast (variance). The metric *quantitatively* confirms that OTS method better captures the contrast of HR dataset, while GAN-based methods (FSSR and DASR) are notably *biased* towards LR dataset statistics (low contrast).



Dataset	*Test HR*	*Test LR*	*Bicubic*	*OTS (VGG)*	*FSSR*	*DASR*
Variance	**0.24** ±0.01	0.17 ±0.01	0.15 ±0.02	**0.20** ±0.03	0.17 ±0.02	0.15 ±0.02

Figure 7: Color palettes and their variance for Test HR, LR datasets and solutions of Bicubic Upscale, OTS, FSSR, DASR methods on AIM19.

E Connection between GAN objectives and Equation (2)

Typical objectives of GAN-based approaches consist of multiple losses – usually one adversarial and several content losses. To make the exposition simple, in our paper, we represented all the content losses as a single loss $c(\cdot, \cdot)$. Below we provide several examples showing how the objectives of popular GAN-based approaches to unpaired image SR could be viewed as (2). For all of these methods, our Lemma 1 applies without any changes. We include in brackets the number of papers citations according to Google Scholar to show that chosen methods are widely used.

FaceSR (2018, 447 citations) The paper of (Bulat et al., 2018) presents one of the first GAN-based approaches to unpaired image SR problem. The method is composed of two steps. First, it learns a degradation between unpaired HR and LR images. Then it employs a second GAN to learn a supervised mapping between paired generated LR and corresponding HR images. The objective of the unpaired step (see their Eq. (1)) is as follows:

$$l = \underbrace{\alpha l_{\text{pixel}}}_{\text{content loss}} + \underbrace{\beta l_{\text{GAN}}}_{\text{adversarial loss}}.$$

Here l_{pixel} is the MSE loss between the generated LR image and downsampled HR. Thus, the objective of this method exactly follows Equation (2).

CinCGAN (2018, 780 citations) The method of (Yuan et al., 2018) is an other pioneering GAN-based approach to unpaired image SR problem, which establishes a different to FaceSR group of two-step methods. First, it uses one CycleGAN to learn a mapping between given noisy LR images and downsampled HR ("clean LR") images. Then, a second CycleGAN fine-tunes a mapping between real LR and HR images. The objective for the first GAN (see their Eq. (5)) is as follows:

$$\mathcal{L}^{LR}_{\text{total}} = \underbrace{\mathcal{L}^{\text{LR}}_{\text{GAN}}}_{\text{adversarial loss}} + \underbrace{w_1\mathcal{L}^{\text{LR}}_{\text{cyc}} + w_2\mathcal{L}^{\text{LR}}_{\text{idt}} + w_3\mathcal{L}^{\text{LR}}_{\text{TV}}}_{\text{content loss}}.$$

Here $\mathcal{L}_{\text{cyc}}^{LR}$ is the cycle-consistency loss[6], $\mathcal{L}_{\text{idt}}^{\text{LR}} - l_1$ identity loss, and $\mathcal{L}_{\text{TV}}^{\text{LR}}$ − total variation loss.

FSSR (Winner of the AIM Challenge on Real-World SR (Lugmayr et al., 2019b), 2019, 228 citations) FSSR (Fritsche et al., 2019) method employs a similar to FaceSR strategy. It firstly learns a mapping between downsampled HR images and given unpaired LR images, and then uses the generated pairs to learn a supervised SR model. The objective of the unpaired step (see their Eq. (6)) is defined by:

$$\mathcal{L}_d = \underbrace{0.005\mathcal{L}_{\text{tex, d}}}_{\text{adversarial loss}} + \underbrace{\mathcal{L}_{\text{col, d}} + 0.01\mathcal{L}_{\text{per, d}}}_{\text{content loss}},$$

where the texture (adversarial) loss $\mathcal{L}_{\text{tex, d}}$ and the color (l_1 identity) loss $\mathcal{L}_{\text{col, d}}$ are applied to low frequencies of the images, while the perceptual loss $\mathcal{L}_{\text{per, d}}$ − to the features of the full images.

DASR (2021, 176 citations) DASR (Wei et al., 2021) structure is also based on the similar to FSSR principles and its two-step structure. In contrast to FSSR, a SR network is trained in a partially supervised manner using not only generated, but also real LR images. The objective of the fully unpaired degradation learning step (see their Eq. (4)) is as follows:

$$\mathcal{L}_{\text{DSN}} = \underbrace{\alpha\mathcal{L}_{\text{con}} + \beta\mathcal{L}_{\text{per}}}_{\text{content loss}} + \underbrace{\gamma\mathcal{L}_{\text{adv}}^{G}}_{\text{adversarial loss}}.$$

Here the adversarial loss $\mathcal{L}_{\text{adv}}^{G}$ is defined on high frequencies of the image, while the content $\mathcal{L}_{\text{con}}$ (l_1 identity) and the perceptual $\mathcal{L}_{\text{per}}$ losses are defined on full images and their features respectively.

ESRGAN-FS (2020, 13 citations) ESRGAN-FS is an other two-step approach based on the principle of learning the degradation, see (Zhou et al., 2020). The objective of its unpaired degradation learning step (see their Eq. (4)) is as follows:

$$\mathcal{L}_{\text{total}} = \underbrace{\lambda_{t1} \cdot \mathcal{L}_{\text{low}} + \lambda_{t2} \cdot \mathcal{L}_{\text{per}}}_{\text{content loss}} + \underbrace{\lambda_{t3} \cdot \mathcal{L}_{\text{high}}}_{\text{adversarial loss}}.$$

Here $\mathcal{L}_{\text{low}}$ (l_1 identity) loss is applied to low frequencies of the images, the perceptual loss $\mathcal{L}_{\text{per}}$ − to the features of the full images, while $\mathcal{L}_{\text{high}}$ (adversarial loss) − high frequencies of the images.

[6] $\mathcal{L}_{\text{cyc}}^{\text{LR}}$ is defined as the MSE loss between given LR image x and $G_2(G_1(x))$, where G_1 learns to map real LR images to "clean" ones and G_2 learns an opposite mapping. For a fixed G_2 this loss can be considered as a part of the content loss.

F ADDITIONAL QUALITATIVE RESULTS ON AIM19

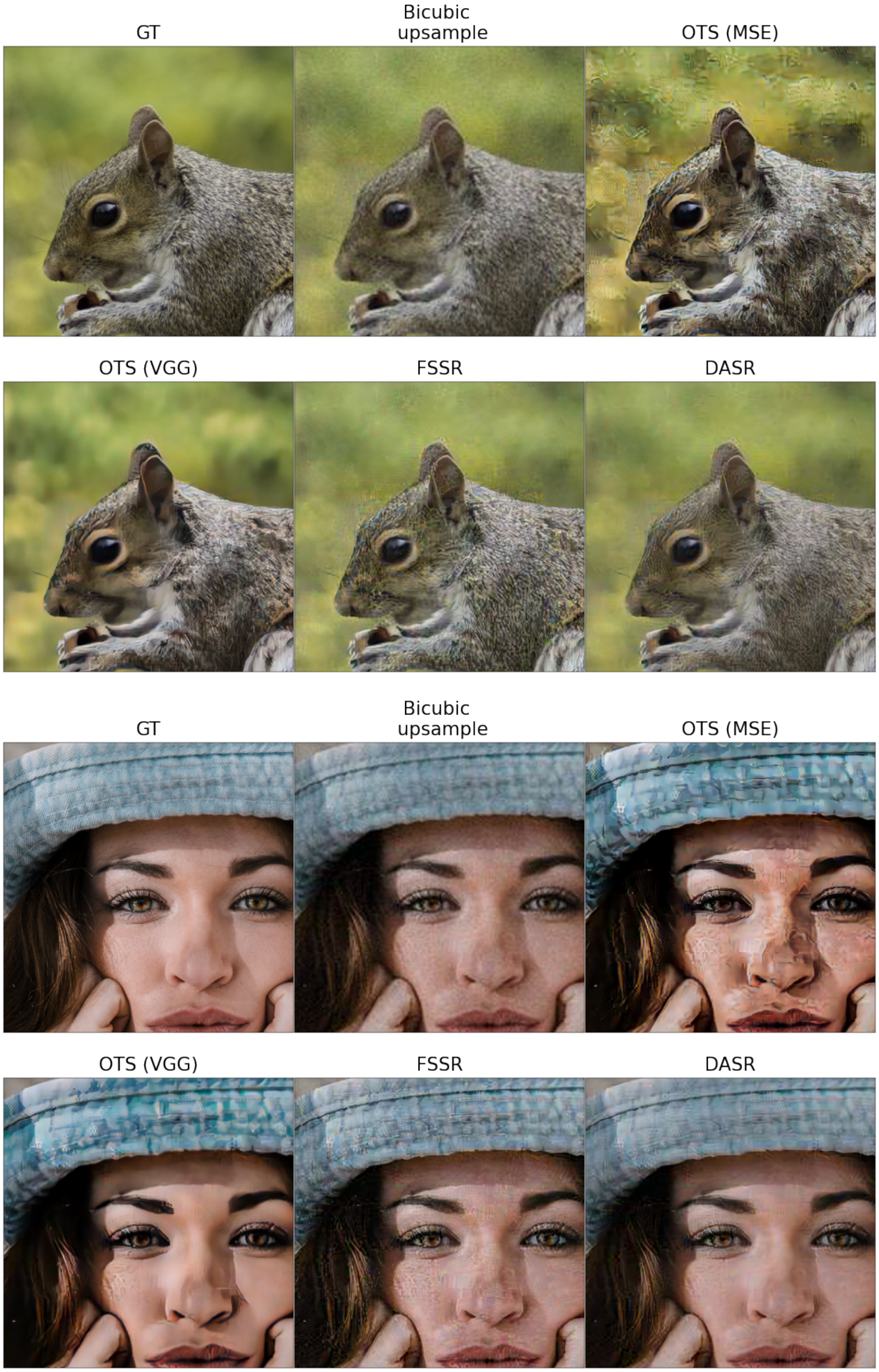


Figure 9: Additional qualitative results of OTS, bicubic upsample, FSSR and DASR on AIM 2019 (800×800 crops).

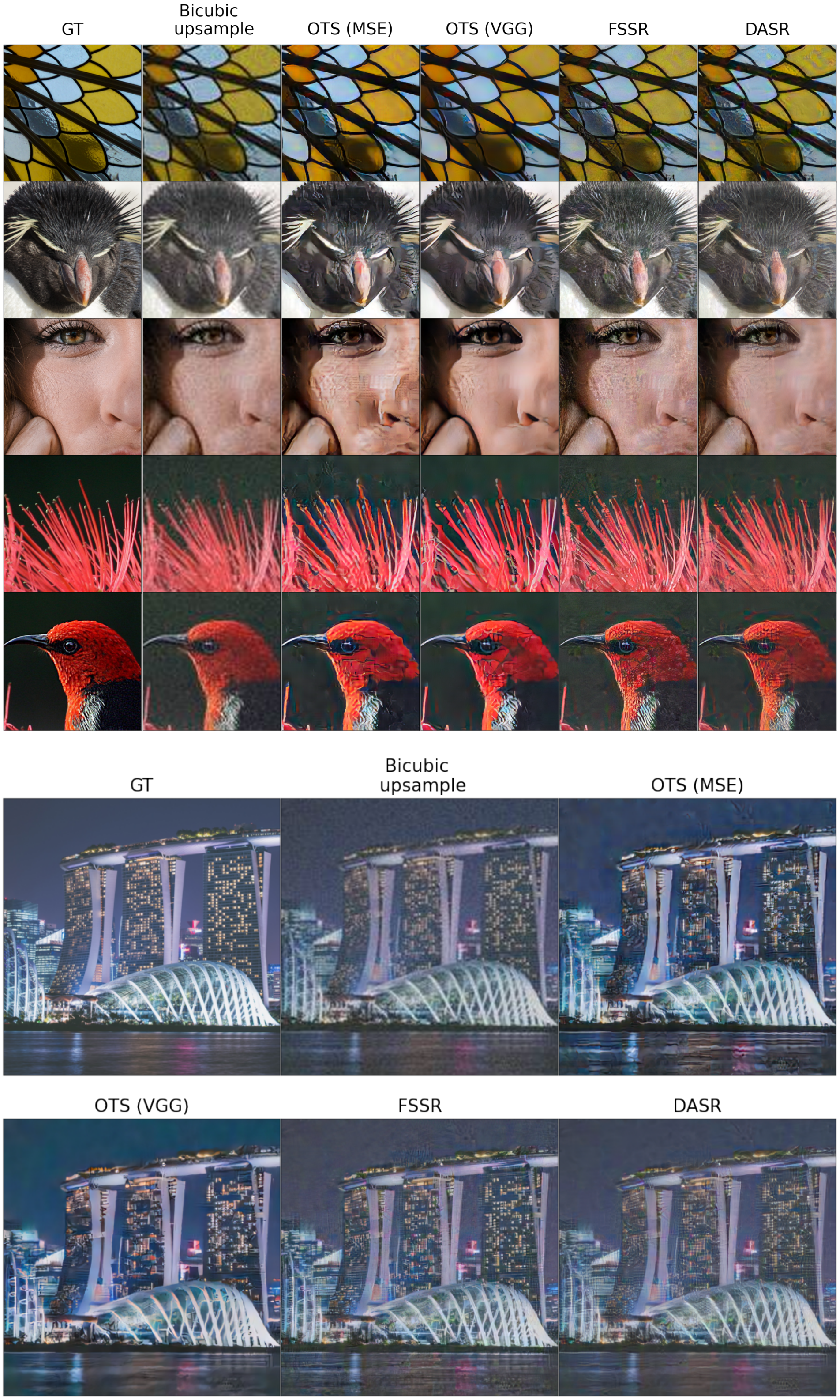


Figure 10: Additional qualitative results of OTS, bicubic upsample, FSSR and DASR on AIM 2019. The sizes of crops on the 1st and 2nd images are 350×350 and 800×800, respectively.