Unpaired Image Super-Resolution with Optimal Transport Maps

Milena Gazdieva * 1  Litu Rout * 2  Alexander Korotin * 1  Alexander Filippov 3  Evgeny Burnaev 1

Abstract

Real-world image super-resolution (SR) tasks often do not have paired datasets limiting 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 such as content and identity losses. We theoretically investigate the 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 empirically show that the learned map is biased, i.e., it may not actually transform the distribution of low-resolution images to high-resolution images. Inspired by these findings, we propose an algorithm for unpaired SR which learns an unbiased OT map for the perceptual transport cost. Unlike existing GAN-based alternatives, our algorithm has a simple optimization objective reducing the necessity to perform complex hyperparameter selection and use additional regularizations. At the same time, it provides nearly state-of-the-art performance on the large-scale unpaired AIM-19 dataset.

1. Introduction

The task 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., 2018c, §5.1); (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); (Wang et al., 2019, §4.1). However, such an artificial setup barely represents the practical setting where the degradation is more sophisticated and unknown (Maeda, 2020). This issue points to the necessity of developing efficient methods capable of learning SR maps from unpaired data without considering prescribed degradations.

In this paper, we study the unpaired image SR task and its solutions based on Generative Adversarial Networks (Goodfellow et al., 2014, GANs) and analyse them from an Optimal Transport (OT, see (Villani, 2008)) perspective. The main contributions of this paper are as follows:

1. We investigate the GAN optimization objectives regularized with content losses which are common in unpaired image SR methods (§5, §4). We prove that the solution to such objectives is always an optimal transport map. We empirically show that such maps are biased (§7.1).

2. We provide an algorithm to fit an unbiased OT map for perceptual transport cost (§6.1) and apply it to unpaired image SR problem (§7.2, §7.3). We establish connections between our algorithm and regularized GANs using integral probability metrics (IPMs) as the loss (§6.2).

Our algorithm solves a minimax optimization objective and does not require extensive hyperparameter search which is a high contrast to existing methods for unpaired image SR. At the same time, the algorithm provides nearly state-of-art performance in the unpaired image SR task (§7.3).

Notation. We use $\mathcal{X}, \mathcal{Y}$ to denote Polish spaces and $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y})$ to denote the respective sets of probability distributions on them. We denote by $\Pi(\mathcal{P}, \mathcal{Q})$ the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ with marginals $\mathcal{P}$ and $\mathcal{Q}$.
2. Unpaired Image Super-Resolution Task

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

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 \( \mathcal{Q} \)-integrable functions on \( \mathcal{Y} \) by \( L^1(\mathcal{Q}) \).

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 \( T_\# P = Q \) but only one inverting the degradation. With no prior knowledge about the correspondence between \( P \) and \( Q \), 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; Yuan et al., 2018; Wei et al., 2021; Wang 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

\[
R_c(T) \overset{\text{def}}{=} \int_\mathcal{Y} c(x, T(x)) dP(x),
\]

where \( c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \) is a function which estimates how different the inputs are. The most popular example is \( \mathcal{X} = \mathcal{Y} = \mathbb{R}^D \) and \( c(x, y) = \| x - y \|_1 \). In this case, the value (1) is usually called the \( \ell^1 \) identity loss:

\[
R_{\ell^1} \overset{\text{def}}{=} \int_\mathcal{X} \| x - T(x) \|_1 dP(x).
\]

More broadly, losses (1) and (2) 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 (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.

3. Background on Optimal Transport

In this section, we give the key concepts of OT theory we use in our paper. For a detailed review, see (Villani, 2008).

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

\[
\text{Cost}(P, Q) \overset{\text{def}}{=} \inf_{T_\# P = Q} \int_\mathcal{X} c(x, T(x)) dP(x),
\]

where the minimum is taken over measurable functions \( T_\# \) which inverts the degradation. We highlight that the image SR task is theoretically ill-posed for two reasons.

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

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

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

Note that (3) is not symmetric, and this formulation does not allow for mass splitting, i.e., for some \( P, Q \) there might be no map \( T \) that satisfies \( T_\# P = Q \). Thus, (Kantorovitch, 1958) proposed the following relaxation (Figure 4):

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

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

With mild assumptions on the transport cost \( c(x, y) \) and distributions \( P, Q \), the minimizer \( \pi^* \) of (4) always exists (Villani, 2008, Theorem 4.1) but might not be unique.

If \( \pi^* \) is of the form \( [\text{id}, T^*]_\# P \in \Pi(P, Q) \) for some \( T^* \), then \( T^* \) is an optimal transport map that minimizes (3).

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

\[
\text{Cost}(P, Q) = \sup_{\theta} \left[ \int_{\mathcal{X}} f^c(x) dP(x) + \int_{\mathcal{Y}} f(y) dQ(y) \right];
\]

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

**Unpaired Image Super-Resolution.** Existing approaches to unpaired image super-resolution mainly solve the problem in two steps. One group of approaches learn 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 to define a loss function for training a generative neural network such as (6) and provide biased solutions (e.g., see (Arjovsky et al., 2017). These methods are out of scope of our paper, since they do not compute OT maps.

**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 our 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) optimize complex GAN objectives such as (6) and provide biased solutions (§5, §7.1). 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 [MM:R]. It is based on the variational reformulation of (5) which is a particular case of our formulation (10). Extensions of [MM:R] appear in (Rout et al., 2021; Fan et al., 2021).

5. Biased Optimal Transport in GANs

In this section, we establish connections between GAN methods regularized by content losses (1) and OT.

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

$$\inf_{T} \left[ D(T_{#}^\mathbb{P}, \mathbb{Q}) + \lambda R_{c}(T) \right].$$

(6)

The term $D(T_{#}^\mathbb{P}, \mathbb{Q})$ ensures that the generated distribution $T_{#}^\mathbb{P}$ of SR images is close to the true HR distribution $\mathbb{Q}$; the second term $R_{c}(T)$ is the content loss (1). Two most popular examples of $D$ 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 complicated than (6). In addition to the content (1) or identity loss (2), several other regularizations are usually introduced, see §4.

For the theoretical analysis, we stick to the basic formulation regularized with generic content loss (6). It represents the simplest and straightforward SR setup. We prove the following lemma which connects the solution $T^{\lambda}$ of (6) and optimal maps for transport cost $c(x, y)$.

**Lemma 1.** Assume that $\lambda > 0$ and the minimizer $T^{\lambda}$ of (6) exists. Then $T^{\lambda}$ is an optimal transport map between $\mathbb{P}$ and $T_{#}^\mathbb{P}$ for transport cost $c(x, y)$, i.e., it minimizes

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

**Proof.** Assume that $T^{\lambda}$ is not an optimal map between $\mathbb{P}$ and $T_{#}^\mathbb{P}$. Then there exists a more optimal $T^{\dagger}$ satisfying $T_{#}^\mathbb{P} = T_{#}^\mathbb{P}$ and $R_{c}(T^{\dagger}) < R_{c}(T^{\lambda})$. We substitute this $T^{\dagger}$ to (6) and derive

$$D(T_{#}^\mathbb{P}, \mathbb{Q}) + \lambda R_{c}(T_{#}^\mathbb{P}) = D(T_{#}^\mathbb{P}, \mathbb{Q}) + \lambda R_{c}(T^{\dagger}) < D(T_{#}^\mathbb{P}, \mathbb{Q}) + \lambda R_{c}(T^{\lambda}),$$

which is a contradiction, since $T^{\lambda}$ is a minimizer of (6), but $T^{\dagger}$ provides the value smaller than the optimal. □

Our Lemma 1 states that the minimizer $T^{\lambda}$ of regularized GAN problem is always an OT map between $\mathbb{P}$ and the distribution generated by the same $T^{\lambda}$ from $\mathbb{P}$. However, there are no guarantees that $T_{#}^\mathbb{P} = \mathbb{Q}$, i.e., that $T^{\lambda}$ actually produces the distribution of HR images, see Figure 5. We argue that, in general, for (6) it holds $T_{#}^\mathbb{P} \neq \mathbb{Q}$.

![Figure 5: Illustration of Lemma 1. The solution $T^{\lambda}$ of (6) is an OT map from $\mathbb{P}$ to $T_{#}^\mathbb{P}$. In general, $T_{#}^\mathbb{P} \neq \mathbb{Q}$.](image)

The following toy example illustrates the issue with the bias.

**Example 1.** Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$. Let $\mathbb{P} = \delta_0$, $\mathbb{Q} = \delta_1$ be distributions concentrated at 0 and 1, respectively. Let $\lambda > 0$. Put $c(x, y) = ||x - y||^2 = ||x - y||^2$ to be the quadratic cost. Also, let $D = W_1$ be the Wasserstein-1 distance, i.e., the optimal transport cost for $||x - y|| = ||x - y||$. In this case, the optimal solution is $T^{\lambda}(0) = \min(1, \frac{1}{2\lambda})$.

**Proof.** The problem (6) simplifies to

$$\inf_{T} \left[ |T(0) - 1| + \lambda(T(0) - 0)^2 \right].$$

(7)

The minimum is attained when $T(0) = \min(1, \frac{1}{2\lambda})$. □
Example 1 shows that it is possible that $T^*_\#_P$ never matches $Q$ exactly if $\lambda > \frac{1}{2}$. To reduce the bias one should use small $\lambda \leq \frac{1}{2}$. However, practically, using very small $\lambda$ will provide no actual effect on the optimization. On the other hand, our OT solver recovers the actual OT map (§6).

In §7.1, we conduct an evaluation of maps obtained via minimizing objective (6) on the artificial benchmark by (Korotin et al., 2021). We empirically demonstrate that the bias exists anyway and it is indeed a notable practical issue.

6. Unbiased Optimal Transport Solver

In §6.1, we derive our algorithm to compute OT maps. Importantly, in §6.2, we detail its differences and similarities with regularized GANs, which we discussed in §5.

6.1. Minimax Optimization Algorithm

We derive a minimax optimization problem to recover the optimal transport map from $P$ to $Q$. We expand the dual form (5). To do this, we first note that

$$\int_X f^*(x) d\mathbb{P}(x) = \int_X \inf_{y \in Y} \{c(x, y) - f(y)\} d\mathbb{P}(x) = \int_X \inf_{T: X \rightarrow Y} \{c(x, T(x)) - f(T(x))\} d\mathbb{P}(x)$$

(8)

In transition from (8) to (9), we replace the optimization over points $y \in Y$ with the equivalent optimization over functions $T: X \rightarrow Y$. This is possible due to the Rockafellar interchange theorem (Rockafellar, 1976, Theorem 3A). We substitute (9) to (5) and derive

$$\text{Cost}(P, Q) = \sup_f \int_X \arg\inf_T \left[ \int_Y f(y) dQ(y) \right] + \int_X \left\{ c(x, T(x)) - f(T(x)) \right\} d\mathbb{P}(x)$$

(9)

(10)

We define the expression under the $\sup$ by $L(f, T)$. Now we show that by solving the saddle point problem (10) one can obtain the optimal transport map $T^*$.

Lemma 2. Assume that the optimal transport map $T^*$ between $P$, $Q$ for cost $c(x, y)$ exists. Then, for every optimal potential $f^* \in \arg\sup_f \int_T L(f, T)$ of (10),

$$T^* \in \arg\inf_T \int_X \left\{ c(x, T(x)) - f(T(x)) \right\} d\mathbb{P}(x).$$

(11)

Proof. Since $f^*$ is optimal, we have

$$\int_T L(f^*, T) = \text{Cost}(P, Q).$$

(12)

We use $T^*_\#_P = Q$ and change of variables $y = T^*(x)$ to derive $\int_X f^*(T^*(x)) d\mathbb{P} = \int_Y f^*(y) dQ$. As a result,

$$L(f^*, T^*) = \int_X c(x, T^*(x)) d\mathbb{P}(x) = \text{Cost}(P, Q).$$

(13)

By comparing (12) and (13), we obtain the desired (11).

Lemma 2 states that one can solve a saddle point problem (10), obtain an optimal pair $(f^*, T^*)$, and use $T^*$ as an OT map from $P, Q$. For general $P, Q$, the $\arg\inf_T$ set for an optimal $f^*$ might contain not only OT map $T^*$ but other functions as well. However, our experiments (§7) show that this is not a serious issue in practice.

To solve the optimization problem (10), we approximate the potential $f$ and map $T$ with neural networks $f_\omega$ and $T_\theta$, respectively. We train them with stochastic gradient ascent-descent by using random batches from $P, Q$. The practical optimization procedure is detailed in Algorithm 1. We call this procedure an Optimal Transport Solver (OTS).

6.2. Regularized GANs vs. Optimal Transport Solver

In this subsection, we discuss similarities and differences between our optimization objective (10) and the objective of regularized GANs (6). We establish an intriguing connection between GANs that use integral probability metrics (IPMs) as discrepancies $D$. A discrepancy $D: \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \rightarrow \mathbb{R}_+$ is an IPM if

$$D(Q_1, Q_2) = \sup_{f \in F} \left[ \int_Y f(y) dQ_2(y) - \int_Y f(y) dQ_1(y) \right]$$

(14)

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

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

**Input**: distributions $P, Q$ accessible by samples; mapping network $T_\theta : \mathcal{X} \rightarrow \mathcal{Y}$; potential $f_\omega : \mathcal{X} \rightarrow \mathbb{R}$; transport cost $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$; number of $K_T$ of inner iterations;

**Output**: approximate OT map $(T_\theta)_\#_P \approx Q$

repeat

Sample batches $X \sim P, Y \sim Q$;

$L_f \leftarrow \frac{1}{|X|} \sum_{y \in Y} f_\omega(y) - \frac{1}{|X|} \sum_{x \in X} f_\omega(T_\theta(x))$;

Update $\omega$ by using $\frac{\partial L_f}{\partial \omega}$;

for $k_T = 1, 2, \ldots, K_T$ do

Sample batch $X \sim P$;

$L_T = \frac{1}{|X|} \sum_{x \in X} [c(x, T_\theta(x)) - f_\omega(T_\theta(x))]$;

Update $\theta$ by using $\frac{\partial L_T}{\partial \theta}$;

until not converged;
Substituting (14) to (6) yields the following saddle-point optimization problem for the regularized IPM GAN:

\[
\inf_T \left[ \sup_{f \in \mathcal{F}} \left\{ \int_Y f(y) dQ(y) - \int_X f(T(x)) dP(x) \right\} + \lambda \int_X c(x, T(x)) dP(x) \right] = \inf_T \left[ \sup_{f \in \mathcal{F}} \left\{ \int_Y f(y) dQ(y) + \int_X \left\{ \lambda \cdot c(x, T(x)) - f(T(x)) \right\} dP(x) \right\} \right], \tag{15}
\]

We emphasize that the expression inside (15) for \( \lambda = 1 \) exactly matches the expression in OTS optimization (10). Below we highlight the differences between (10) and (15).

First, in OTS the map \( T \) is a solution to the inner optimization problem, while in regularized IPM GAN the generator \( T \) is a solution to the outer optimization problem.

Second, 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 \( D \) is the Wasserstein-1 \((\mathcal{W}_1)\) IPM, one has to use an additional penalization, e.g., the gradient penalty (Gulrajani et al., 2017). This further complicates the optimization and adds extra hyperparameters which have to be carefully selected.

Third, the optimization of IPM GAN requires selecting the hyperparameter \( \lambda \) which balances the content loss \( \int_X c(x, T(x)) dP(x) \) and the discrepancy \( D \). In OTS for all costs \( \lambda \cdot c(x, y) \) with \( \lambda > 0 \), the OT map \( T^* \) is the same.

We summarize the differences and the similarities between OTS and regularized IPM GANs in Table 1.

### Table 1: Comparison of the optimization objectives of OTS (ours) and regularized IPM GAN.

|                  | Optimal Transport Solver (Ours)                                                                 | Regularized IPM GAN                                                                 |
|------------------|------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Minimax          | \( \sup_{f \in \mathcal{F}} \int Y f(y) dQ(y) - \int X f(T(x)) dP(x) \)                        | \( \inf_T \sup_{f \in \mathcal{F}} \left[ \int_Y f(y) dQ(y) + \int_X \{ \lambda \cdot c(x, T(x)) - f(T(x)) \} dP(x) \right] \) |
| Optimization     | \( T^* \) solves the inner problem (for optimal \( f^* \)); it is an OT map from \( P \) to \( Q \) (Lemma 2) | \( T^* \) solves the outer problem; it is a biased OT map (§5, §7.1)           |
| Potential \( f  \) | Unconstrained \( f \in L^1(Q) \)                                                                | Constrained \( f \in F \subset L^1(Q) \) A method to impose the constraint is needed. |
| Regularization   | \( \lambda \)                                                                                   | Hyperparameter choice required                                                  |

#### 7. Evaluation

In §7.1, we assess the bias of regularized IPM GANs by using the Wasserstein-2 benchmark (Korotin et al., 2021). In §7.2, we test our method on the CelebA dataset (Liu et al., 2015). In §7.3, we evaluate our method on the large-scale unpaired AIM-19 dataset (Lugmayr et al., 2019b).

The code was written using PyTorch, and each experiment was conducted on a single Tesla V100 GPU. We list the hyperparameters for our Algorithm 1 in Table 5.

#### Neural network architectures.

We use WGAN-QC’s (Liu et al., 2019) ResNet (He et al., 2016) architecture for the potential \( f_{\omega} \). In §7.1, where input and output images have the same size, we use UNet \(^1\) (Ronneberger et al., 2015) as the transport map \( T_\theta \). In §7.2, §7.3, the LR input images are 4 × 4 times smaller than HR, so we test bilinear upsampling plus UNet and, alternatively, EDSR net (Lim et al., 2017).

#### Transport costs.

In §7.1, where input and output images have the same size \((\mathcal{X} = \mathcal{Y})\), 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 §7.2, §7.3, the LR input images are 4 × 4 times smaller than HR, so 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 \( b = \text{MSE} \) and the following perceptual cost as \( b \):

\[
b(x^{\text{up}}, y) = \text{MSE}(x^{\text{up}}, y) + \frac{1}{3} : \text{MAE}(x^{\text{up}}, y) + \sum_{k\in\{3,8,15,22\}} \text{MSE}(f_k(x^{\text{up}}), f_k(y)) + \text{lpips}(x^{\text{up}}, y), \tag{16}
\]

where \( f_k \) denotes the features of the \( k \)-th layer of a pre-trained VGG-16 network (Simonyan & Zisserman, 2014), \( \text{MAE} \) is the mean absolute error \( \frac{\|x - y\|}{\dim(\mathcal{Y})} \), and \( \text{lpips}^2 \) is the learned perceptual image similarity (Zhang et al., 2018a).

#### Dynamic transport cost.

In the preliminary experiments, we used bicubic upsampling as the “Up” operation in the cost. 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 \( f_{\omega} \) (see Table 5), we

\[^1\text{github.com/milesial/Pytorch-UNet}\]

\[^2\text{github.com/richzhang/PerceptualSimilarity}\]
change the cost to \( c(x, y) = b(T'_\theta(x), y) \), where \( T'_\theta \) is a fixed frozen copy of the currently learned SR map \( T_\theta \).

**Optimizers.** We employ Adam (Kingma & Ba, 2014).

### 7.1. Assessing the Bias in Regularized GANs

In this section, we empirically confirm the insight of §5 that the solution \( T^\lambda \) of (6) may not satisfy \( T^\lambda \|P = Q \). Note if \( T^\lambda \|P = Q \), then by our Lemma 1, we conclude that \( T^\lambda = T^* \), where \( T^* \) is an OT map from \( P \) to \( Q \) for \( c(x, y) \). Thus, to access the bias, it is reasonable to compare the learned map \( T^\lambda \) with the ground truth OT map \( T^* \) for \( P \), \( Q \).

For evaluation, we use the Wasserstein-2 benchmark (Korotin et al., 2021). It provides high-dimensional continuous pairs \( P \), \( Q \) with an analytically known OT map \( T^* \) for the quadratic cost \( c(x, y) = \| x - y \|^2 \). We use their “Early” images benchmark pair (Korotin et al., 2021, §4.1). It simulates the image deblurring setup, that is, \( X = Y \) is the space of \( 64 \times 64 \) RGB images, \( P \) is blurry faces, \( Q \) is clean faces satisfying \( Q = T^* \|P \), where \( T^* \) is an analytically known OT map, see the 1st and 2nd lines in Figure 6.

On the benchmark, we compare OTS (10) and IPM GAN (6). We use MSE as the content loss \( c(x, y) \). In IPM GAN, we use the Wasserstein-1 (\( \mathcal{W}_1 \)) loss with the gradient penalty \( \lambda_{GP} = 10 \) (Gulrajani et al., 2017) as \( D \). 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 our 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 \( T^\lambda \) and the map \( \tilde{T} \) obtained by OTS in Figure 6.

To quantify the learned maps from \( P \) to \( Q \), we use the \( \mathcal{L}^2 \) unexplained variance percentage (Korotin et al., 2021, §4.2): \( \mathcal{L}^2\text{-UVP}(T) = \frac{100 \cdot \|T - T^\lambda \|_2^2 / \text{Var}(Q)}{\|T - T^* \|_2^2}, \) where \( \| \cdot \|_P \) denotes the \( \mathcal{L}^2(\mathbb{R}^D \rightarrow \mathbb{R}^D, P) \) norm. We estimate \( \mathcal{L}^2\text{-UVP} \) on 2\,147 samples from \( P \), see Table 2 with the results.

The results show that the performance of the regularized IPM GAN significantly depends on the choice of the content loss value \( \lambda \). For high values \( \lambda \geq 10^3 \), the learned map is close to the identity as expected. For small values \( \lambda \leq 10^3 \), the regularization has little effect, and WGAN-GP solely struggles to fit a good restoration map. Even for the best performing \( \lambda = 10^2 \) the \( \mathcal{L}^2\text{-UVP} \) error is notably bigger than in OTS. Note that in OTS there is no parameter \( \lambda \).

### 7.2. Image Super-Resolution of Faces

We conduct an experiment using CelebA (Liu et al., 2015) faces dataset to test the applicability of OT for unpaired SR.

**Pre-processing and train-test split.** We resize images to \( 64 \times 64 \) px. We adopt the exact unpaired train-test split

![Figure 6: Comparison of OTS (ours) and regularized IPM GAN on the Wasserstein-2 benchmark. The 1st line shows blurry faces \( x \sim P \), the 2nd line, clean faces \( y = T^*(x) \), where \( T^* \) is the OT map from \( P \) to \( Q \). Next lines show maps from \( P \) to \( Q \) fitted by methods in view.](image)

![Table 2: Quantitative evaluation of restoration maps fitted by the regularized IPM GAN and OTS (ours) using the Wasserstein-2 images benchmark by (Korotin et al., 2021).](image)
Table 3: Comparison of OTS (ours) with the bicubic upsample on CelebA dataset. The 1st and 2nd best results are highlighted in green and blue, respectively.

| Method               | FID  | PSNR | SSIM | LPIPS |
|----------------------|------|------|------|-------|
| Bicubic upsample     | 130.72 | 22.73 | 0.665 | 0.308 |
| OTS (ours) UNet      | 14.76  | 22.47 | 0.726 | 0.071 |
| OTS (ours) EDSR      | 13.72  | 22.41 | 0.702 | 0.053 |

Table 4: Comparison of OTS (ours) with FSSR and DASR on AIM 2019 dataset. The 1st, 2nd, 3rd best results are highlighted in green, blue and underlined, respectively.

| Method               | FID  | PSNR | SSIM | LPIPS |
|----------------------|------|------|------|-------|
| Bicubic upsample     | 178.59 | 22.39 | 0.609 | 0.689 |
| OTS (ours) EDSR, MSE | 139.17 | 19.73 | 0.504 | 0.457 |
| OTS (ours) UNet, MSE | 133.93 | 18.72 | 0.448 | 0.481 |
| OTS (ours), UNet, VGG| 89.04  | 20.96 | 0.594 | 0.380 |
| OTS (ours), UNet, VGG| 98.52  | 21.35 | 0.580 | 0.367 |
| OTS (ours), UNet, VGG + LPIPS| 69.17 | 20.09 | 0.542 | 0.406 |
| FSSR                 | 53.92  | 20.83 | 0.514 | 0.390 |
| DASR                 | 124.09 | 21.79 | 0.577 | 0.346 |

7.3. Large-Scale Evaluation

For evaluating our 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 our 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 and 32×32 LR random patches of images augmented via random flips and rotations. We conduct separate experiments using either UNet or EDSR as the transport map.
and either MSE or perceptual cost. In the case of perceptual costs for EDSR, we set \((w_{vgg}, w_{lpips}) = (1/50, 0)\). For bilinear upsample + UNet, we test \((w_{vgg}, w_{lpips}) = (3/100, 0)\) and \((1/50, 1/5)\).

**Metrics.** We calculate the same metrics as in §7.2. FID is computed on 128 × 128 patches of LR test images upsampled by the method in view w.r.t. random patches of test HR images. We use 50000 patches to compute FID. The other metrics are computed on the entire upsampled LR test images and HR test images, not patches.

**Experimental results** are given in Table 4, Figure 8. The results show that the usage perceptual cost function in OTS boosts performance. According to FID, OTS with perceptual cost function beats DASR method in all the cases, and for one version it is close to FSSR. It also outperforms FSSR in PSNR, SSIM and, importantly, LPIPS. Bicubic upsample outperforms other methods, according to PSNR and SSIM, but these metrics are known to poorly correlate with perceptual image quality (Blau & Michaeli, 2018). According to visual analysis, OTS deals better with noise artifacts.

To conclude, our method with the perceptual cost is comparable to FSSR and DASR. At the same time, it has a simpler end-to-end learning pipeline and smaller amount of hyperparameters, see Table 6 in Appendix A for comparison.

**Additional experimental results** are given in Appendix B.

### 8. Discussion

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

**Potential Impact.** We expect our optimal transport approach to improve existing applications of image super-resolution. Importantly, it has less hyperparameters than most existing methods — this should simplify its usage in practice. Besides, our method is generic and presumably can be applied to other unpaired learning tasks as well. Studying such applications is a promising avenue for the future work.

**Limitations.** Our method fits a one-to-one optimal mapping (transport map) for super-resolution which, in general, might not exist. Besides, not all optimal solutions of our optimization objective are guaranteed to be OT maps. These limitations point to necessity for further theoretical analysis and improvement of our method for optimal transport.

**Reproductibility.** We provide the code for the experiments and will provide the trained models, see README.md.
References

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

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

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

Fan, J., Liu, S., Ma, S., Chen, Y., and Zhou, H. Scalable computation of monge maps with general costs. *arXiv preprint arXiv:2106.03812*, 2021.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. 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.

Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., and Huang, F. Real-world super-resolution via kernel estimation and noise injection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1914–1923, 2020.

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

Kim, G., Park, J., Lee, K., Lee, J., Min, J., Lee, B., Han, D. K., and Ko, H. 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.

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

Korotin, A., Li, L., Genevay, A., Solomon, J., Filippov, A., and Burnaev, E. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *arXiv preprint arXiv:2106.01954*, 2021.

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

Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. 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.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 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.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. 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.

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

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lu, G., Zhou, Z., Shen, J., Chen, C., Zhang, W., and Yu, Y. Large-scale optimal transport via adversarial training with cycle-consistency. *arXiv preprint arXiv:2003.06635*, 2020.

Lugmayr, A., Danelljan, M., and Timofte, R. Unsupervised learning for real-world super-resolution. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3408–3416, 2019a.
Unpaired Image Super-Resolution with Optimal Transport Maps

Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagoopal, A., Joon, N. H., 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.

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

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev gan. arXiv preprint arXiv:1711.04894, 2017.

Rockafellar, R. T. Integral functionals, normal integrands and measurable selections. In Nonlinear operators and the calculus of variations, pp. 157–207. Springer, 1976.

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

Rout, L., Korotin, A., and Burnaev, E. Generative modeling with optimal transport maps. arXiv preprint arXiv:2110.02999, 2021.

Simonyn, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200, 2016.

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

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

Wang, W., Zhang, H., Yuan, Z., and Wang, C. Unsupervised real-world super-resolution: A domain adaptation perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4318–4327, 2021.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. Esrgan: Enhanced super-resolution generative adversarial networks. In Leal-Taixé, L. and Roth, S. (eds.), Computer Vision – ECCV 2018 Workshops, pp. 63–79, Cham, 2019. Springer International Publishing.

Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., and Song, H. 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.

Xie, Y., Chen, M., Jiang, H., Zhao, T., and Zha, H. 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, Y., Liu, S., Zhang, J., bing Zhang, Y., Dong, C., and Lin, L. 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.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018a.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. In ECCV, 2018b.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018c.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. 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. Training Details

For EDSR, we set the number of residual blocks to 64, number of features to 128 and residual scaling to 1. For UNet, we set the base factor to 64. Training details are given in Table 5.

| Experiment | dim(\(\mathcal{X}\)) | dim(\(\mathcal{Y}\)) | \(f\) | \(T\) | \(k_T\) | \(l_{r,T}\) | Initial cost | Total iters \((f)\) | Cost update every | Batch size |
|------------|----------------------|----------------------|-------|--------|--------|-----------|--------------|------------------|-----------------|-------------|
| Benchmark (\(M^7.1\)) | 3 \times 64 \times 64 | 3 \times 64 \times 64 | UNet | 10 | 10^{-4} | MSE | 10K | — | 64 |
| Celeba (\(M^7.2\)) | 3 \times 16 \times 16 | 3 \times 64 \times 64 | ResNet | 15 | 10^{-4} | Bicubic + MSE | 100K | 25K | 64 |
| AIM-19 (\(M^7.3\)) | 3 \times 32 \times 32 (patches) | 3 \times 128 \times 128 (patches) | EDSR | 15 | | Bicubic + VGG | 50K | 25K | 8 |

Table 5: Hyperparameters that we use in the experiments with our Algorithm 1.

We provide a comparison of hyperparameters of FSSR, DASR and OTS (ours) in Table 6. In contrast to FSSR and DASR, our method does not contain a degradation part. This helps to notably reduce the amount of tunable hyperparameters.

| Method | Degradation part | Super-resolution part | Total |
|--------|------------------|-----------------------|-------|
| FSSR   | 2 neural networks; 2 optimizers; 2 schedulers; 1 adversarial loss; 1 content loss (\(\ell_1\)+perceptual) | 4 neural networks; 4 optimizers; 4 schedulers; 2 adversarial losses; 2 content losses (\(\ell_1\)+perceptual) |       |
| DASR   | 2 neural networks; 2 optimizers; 2 schedulers; 1 adversarial loss; 1 content loss (\(\ell_1\)+perceptual) | 4 neural networks; 4 optimizers; 4 schedulers; 2 adversarial losses; 2 content losses (\(\ell_1\)+perceptual) |       |
| OTS (ours) | — | 2 neural networks; 2 optimizers; 1 cost (\(\ell_2+\ell_1\)+perceptual) |       |

Table 6: Comparison of hyperparameters used in FSSR, DASR and OTS (ours) methods.
B. Additional Results

Figure 9: Additional qualitative results of OTS (ours), bicubic upsample, FSSR and DASR on AIM 2019. The sizes of crops on the 1st and 2nd images are $350 \times 350$ and $800 \times 800$, respectively.
Figure 10: Additional qualitative results of OTS (ours), bicubic upsample, FSSR and DASR on AIM 2019 (800×800 crops).