Generative Modeling with Optimal Transport Maps

Litu Rout  
Space Applications Centre 
Indian Space Research Organisation  
lr@sac.isro.gov.in

Alexander Korotin  
Skolkovo Institute of Science and Technology  
Artificial Intelligence Research Institute (AIRI)  
a.korotin@skoltech.ru

Evgeny Burnaev  
Skolkovo Institute of Science and Technology  
Artificial Intelligence Research Institute (AIRI)  
e.burnaev@skoltech.ru

Abstract

With the discovery of Wasserstein GANs, Optimal Transport (OT) has become a powerful tool for large-scale generative modeling tasks. In these tasks, OT cost is typically used as the loss for training GANs. In contrast to this approach, we show that the OT map itself can be used as a generative model, providing comparable performance. Previous analogous approaches consider OT maps as generative models only in the latent spaces due to their poor performance in the original high-dimensional ambient space. In contrast, we apply OT maps directly in the ambient space, e.g., a space of high-dimensional images. First, we derive a min-max optimization algorithm to efficiently compute OT maps for the quadratic cost (Wasserstein-2 distance). Next, we extend the approach to the case when the input and output distributions are located in the spaces of different dimensions and derive error bounds for the computed OT map. We evaluate the algorithm on image generation and unpaired image restoration tasks. In particular, we consider denoising, colorization, and inpainting, where the optimality of the restoration map is a desired attribute, since the output (restored) image is expected to be close to the input (degraded) one.

1 Introduction

Since the discovery of Generative Adversarial Networks (GANs, Goodfellow et al. (2014)), there has been a surge in generative modeling (Radford et al., 2016; Arjovsky et al., 2017; Brock et al., 2019; Karras et al., 2019). In the past few years, Optimal Transport (OT, Villani (2008)) theory has been pivotal in addressing important issues of generative models. In particular, the usage of Wasserstein distance has improved diversity (Arjovsky et al., 2017; Gulrajani et al., 2017), convergence (Sanjabi et al., 2018), and stability (Miyato et al., 2018; Kim et al., 2021) of GANs.

Generative models based on OT can be split into two classes depending on what OT is used for. First, the optimal transport cost serves as the loss for generative models, see Figure 1a. This is the most prevalent class of methods which includes WGAN (Arjovsky et al., 2017) and its modifications: WGAN-GP (Gulrajani et al., 2017), WGAN-LP (Petzka et al., 2018), and WGAN-QC (Liu et al., 2019). Second, the optimal transport map is used as a generative model itself, see Figure 1b. Such approaches include LSOT (Seguy et al., 2018), AE-OT (An et al., 2020a), ICNN-OT (Makkuva et al., 2020), W2GN (Korotin et al., 2021a). Models of the first class have been well-studied, but limited attention has been paid to the second class. Existing approaches of the second class primarily consider OT maps in latent spaces of pre-trained autoencoders (AE), see Figure 3. The performance of such generative models depends on the underlying AEs, in which decoding transformations are often not accurate; as a result this deficiency limits practical applications in high-dimensional ambient spaces. For this reason, using OT in the latent space does not necessarily guarantee superior performance in generative modeling.
The focus of our paper is the second class of OT-based models using OT map as the generative map. Finding an optimal mapping is motivated by its ability to preserve specific attributes of the input samples, a desired property in unpaired learning. For example, in unpaired image-to-image translation, the learner has to fit a map between two data distributions which preserves the image content. CycleGAN-based models (Zhu et al., 2017) are widely used for this purpose. However, they typically have complex optimization objectives consisting of several losses (Amodio & Krishnaswamy, 2019; Lu et al., 2019) in order to make the fitted map preserve the required attributes.

The main contributions of this paper are as follows:

1. We propose an end-to-end algorithm (§4.3) to fit OT maps for the quadratic cost (Wasserstein-2 distance) between distributions located on the spaces of equal dimensions (§4.1) and extend the method to unequal dimensions as well (§4.2). We prove error bounds for the method (§4.4).

2. We demonstrate large-scale applications of OT maps in popular computer vision tasks. We consider image generation (§5.1) and unpaired image restoration (§5.2) tasks.

Our strict OT-based framework allows the theoretical analysis of the recovered transport map. The OT map obtained by our method can be directly used in large-scale computer vision problems which is in high contrast to previous related methods relying on autoencoders and OT maps in the latent space. Importantly, the performance and computational complexity of our method is comparable to OT-based generative models using OT cost as the loss.

Notations. In what follows, $\mathcal{X}$ and $\mathcal{Y}$ are two complete metric spaces, $\mu(x)$ and $\nu(y)$ are probability distributions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. For a measurable map $T : \mathcal{X} \to \mathcal{Y}$, $T_\# \mu$ denotes the pushforward distribution of $\mu$, i.e., the distribution for which any measurable set $E \subset \mathcal{Y}$ satisfies $T_\# \mu(E) = \mu(T^{-1}(E))$. For a vector $x$, $\|x\|$ denotes its Euclidean norm. We use $(x, y)$ to denote the inner product of vectors $x$ and $y$. We use $\Pi(\mu, \nu)$ to denote the set of joint probability distributions on $\mathcal{X} \times \mathcal{Y}$ whose marginals are $\mu$ and $\nu$, respectively (couplings). For a function $f : \mathbb{R}^D \to \mathbb{R} \cup \{\pm \infty\}$ its Legendre–Fenchel transform (the convex conjugate) is $\overline{f}(y) = \sup_{x \in \mathbb{R}^D} \{\langle x, y \rangle - f(x)\}$. It is convex, even if $f$ is not.

## 2 Background on Optimal Transport

Consider a cost of transportation, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined over the product space of $\mathcal{X}$ and $\mathcal{Y}$.

Monge’s Formulation. The optimal transport cost between $\mu$ and $\nu$ for ground cost $c(\cdot, \cdot)$ is

$$\text{Cost}(\mu, \nu) \overset{def}{=} \inf_{T_\# \mu = \nu} \int_{\mathcal{X}} c(x, T(x)) \, d\mu(x),$$

where the infimum is taken over all measurable maps $T : \mathcal{X} \to \mathcal{Y}$ pushing $\mu$ to $\nu$, see Figure 2. The map $T^*$ on which the infimum in (1) is attained is called the optimal transport.
map. Monge’s formulation does not allow splitting. For example, when \( \mu \) is a Dirac distribution and \( \nu \) is a non-Dirac distribution, the feasible set of equation (1) is empty.

**Kantorovich’s Relaxation.** Instead of asking to which particular point \( y \in \mathcal{Y} \) should all the probability mass of \( x \) be moved, Kantorovich (1948) asks how the mass of \( x \) should be distributed among all \( y \in \mathcal{Y} \). Formally, a transport coupling replaces a transport map; the OT cost is given by:

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

(2)

where the infimum is taken over all couplings \( \pi \in \Pi(\mu, \nu) \) of \( \mu \) and \( \nu \). The coupling \( \pi^* \) attaining the infimum of (2) is called the optimal transport plan. Unlike the formulation of (1), the formulation of (2) is well-posed, and with mild assumptions on spaces \( \mathcal{X}, \mathcal{Y} \) and ground cost \( c(\cdot, \cdot) \), the minimizer \( \pi^* \) of (2) always exists (Villani, 2008, Theorem 4.1). In particular, if \( \pi^* \) is deterministic, i.e., \( \pi^* = [\text{id}_\mathcal{X}, T^*]_{\#}\mu \) for some \( T^* : \mathcal{X} \to \mathcal{Y} \), then \( T^* \) minimizes (1).

**Duality.** The dual form of (2) is given by Kantorovich (1948):

\[
\text{Cost}(\mu, \nu) = \sup_{(u, v) \in L^1(\mu, \nu)} \left\{ \int_{\mathcal{X}} u(x) \, d\mu(x) + \int_{\mathcal{Y}} v(y) \, d\nu(y) : u(x) + v(y) \leq c(x, y) \right\},
\]

(3)

with \( u \in L^1(\mu) \), \( v \in L^1(\nu) \) called Kantorovich potentials. For \( u : \mathcal{X} \to \mathbb{R} \) and \( v : \mathcal{Y} \to \mathbb{R} \) define their \( c \)-transforms by

\[
\begin{align*}
\tilde{c}(y) &= \inf_{x \in \mathcal{X}} \{ c(x, y) - u(x) \} \quad \text{and} \\
\hat{c}(x) &= \inf_{y \in \mathcal{Y}} \{ c(x, y) - v(y) \}
\end{align*}
\]

respectively. Using \( c \)-transform, (3) is reformulated as (Villani, 2008, §5)

\[
\text{Cost}(\mu, \nu) = \sup_{v} \left\{ \int_{\mathcal{X}} \tilde{c}(x) \, d\mu(x) + \int_{\mathcal{Y}} v(y) \, d\nu(y) \right\} = \sup_{v} \left\{ \int_{\mathcal{X}} \hat{c}(x) \, d\mu(x) + \int_{\mathcal{Y}} v(y) \, d\nu(y) \right\} = \sup_{u} \left\{ \int_{\mathcal{X}} u(x) \, d\mu(x) + \int_{\mathcal{Y}} \hat{c}(y) \, d\nu(y) \right\}.
\]

(4)

**Primal-dual relationship.** For certain ground costs \( c(\cdot, \cdot) \), the primal solution \( T^* \) of (1) can be recovered from the dual solution \( u^* \) of (4). For example, if \( \mathcal{X} = \mathcal{Y} = \mathbb{R}^D \) and \( c(\cdot, \cdot) = h(x - y) \) with strictly convex \( h : \mathbb{R}^D \to \mathbb{R} \) and \( \mu \) is absolutely continuous supported on the compact set, then

\[
T^*(x) = x - (\nabla h)^{-1}(\nabla u^*(x)),
\]

(5)

see (Santambrogio, 2015, Theorem 1.17). For general costs, see (Villani, 2008, Theorem 10.28).

### 3 Optimal Transport in Generative Models

**Optimal transport cost as the loss (Figure 1a).** Starting with the works of Arjovsky & Bottou (2017); Arjovsky et al. (2017), the usage of OT cost as the loss has become a major way to apply OT for generative modeling. In this setting, given data distribution \( \nu \) and fake distribution \( \mu_\theta \), the goal is to minimize \( \text{Cost}(\mu_\theta, \nu) \) w.r.t. the parameters \( \theta \). Typically, \( \mu_\theta \) is a pushforward distribution of some given distribution, e.g., \( \mathcal{N}(0, I) \), via generator network \( G_\theta \).

The Wasserstein-1 distance \( \mathcal{W}_1 \), i.e., the transport cost for ground cost \( c(x, y) = \|x - y\| \), is the most practically prevalent example of such a loss. Models based on this loss are known as Wasserstein GANs (WGANs). They estimate \( \mathcal{W}_1(\mu_\theta, \nu) \) based on the dual form as given by (4). For \( \mathcal{W}_1 \), the optimal potentials \( u^*, v^* \) of (4) satisfy \( u^* = -v^* \) where \( u^* \) is a 1-Lipschitz function (Villani, 2008, Case 5.16). As a result, to compute \( \mathcal{W}_1 \), one needs to optimize the following simplified form:

\[
\mathcal{W}_1(\mu_\theta, \nu) = \inf_{\|u\|_{L^1} \leq 1} \left\{ \int_{\mathcal{X}} u(x) \, d\mu_\theta(x) - \int_{\mathcal{Y}} u(y) \, d\nu(y) \right\}.
\]

(6)

In WGANs, the potential \( u \) is called the discriminator. Optimization of (6) reduces constrained optimization of (2) with two potentials \( u, v \) to optimization of only one discriminator \( u \). In practice, enforcing the Lipschitz constraint on \( u \) is challenging. Most methods to do this are regularization-based, e.g., they use gradient penalty (Gulrajani et al., 2017) WGAN-GP and Lipschitz penalty (Petzka et al., 2018) WGAN-LP. Other methods enforce Lipschitz property via incorporating certain hard restrictions on the discriminator’s architecture (Anil et al., 2019; Tanmilian & Biau, 2021).

**General transport costs** (other than \( \mathcal{W}_1 \)) can also be used as the loss for generative models. They are less popular since they do not have a dual form reducing to a single potential function similar to (6) for \( \mathcal{W}_1 \). Consequently, the challenging estimation of the \( c \)-transform \( u^c \) is needed. To
avoid this. Sanjabi et al. (2018) consider the dual form of (3) with two potentials $u, v$ instead form (4) with one $u$ and softly enforce the condition $u(x) + v(y) \leq c(x, y)$ via entropy or quadratic regularization. Nhan Dam et al. (2019) use the dual form of (4) and amortized optimization to compute $u^*$ via an additional neural network. Both methods work for general $c(\cdot, \cdot)$, though the authors test them for $c(x, y) = ||x - y||$ only, i.e., $W_1$ distance. Mallasto et al. (2019) propose a fast way to approximate the c-transform and test the approach (WGAN-(q,p)) with several costs, in particular, the Wasserstein-2 distance ($W_2$), i.e., the transport cost for the quadratic ground cost $c(x, y) = \frac{1}{2}||x - y||^2$. Specifically for $W_2$, Liu et al. (2019) approximate the c-transform via a linear program (WGAN-QC).

A fruitful branch of OT-based losses for generative models comes from modified versions of OT cost, such as Sinkhorn (Genevay et al., 2018), sliced (Deshpande et al., 2018) and minibatch (Patras et al., 2019) OT distances. They typically have lower sample complexity than usual OT and can be accurately estimated from random mini-batches without using dual forms such as (3). In practice, these approaches usually learn the ground OT cost $c(\cdot, \cdot)$.

The aforementioned methods use OT cost in the ambient space to train GANs. There also exist approaches using OT cost in the latent space. For example, Tolstikhin et al. (2017); Patrini et al. (2020) use OT cost between encoded data and a given distribution as an additional term to reconstruction loss for training an AE. As the result, AE’s latent distribution becomes close to the given one.

**Optimal transport map as the generative map** (Figure 1b). Methods to compute the OT map (plan) are less common in comparison to those computing the cost. Recovering the map from the primal form (1) or (2) usually yields complex optimization objectives containing several adversarial terms (Xie et al., 2019); Liu et al. (2021); Lu et al. (2020). Such procedures require careful hyperparameter choice. This needs to be addressed before using these methods in practice.

Primal-dual relationship (2) makes it possible to recover the OT map via solving the dual form (3). Dual-form based methods primarily consider $W_2$ cost due to its nice theoretical properties and relation to convex functions (Brenier, 1991). In the semi-discrete case ($\mu$ is continuous, $\nu$ is discrete), An et al. (2020a) and Lei et al. (2019) compute the dual potential and the OT map by using the Alexandrov theory and convex geometry. For the continuous case, Seguy et al. (2018) use the entropy (quadratic) regularization to recover the dual potentials and extract OT map from them via the barycentric projection. Taghvaei & Jalali (2019), Makkuva et al. (2020), Korotin et al. (2021a) employ input-convex neural networks (ICNNs, see Amos et al. (2017)) to parametrize potentials in the dual problem and recover OT maps by using their gradients.

![Figure 3: The existing most prevalent approach to use OT maps in generative models.](image)

The aforementioned dual form methods compute OT maps in latent spaces for problems such as domain adaptation and latent space mass transport, see Figure 3 OT maps in high-dimensional ambient spaces, e.g., natural images, are usually not considered. Recent evaluation of continuous OT methods for $W_2$ (Korotin et al. 2021b) reveals their crucial limitations, which negatively affect their scalability, such as poor expressiveness of ICNN architectures or bias due to regularization.

## 4 End-to-end solution to learn optimal maps

### 4.1 Equal dimensions of input and output distributions

In this section, we use $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$ and consider the Wasserstein-2 distance ($W_2$), i.e., the optimal transport for the quadratic ground cost $c(x, y) = \frac{1}{2}||x - y||^2$. We use the dual form (4) to derive a saddle point problem the solution of which yields the OT map $T^*$. We consider distributions $\mu, \nu$ with finite second moments. We assume that for distributions $\mu, \nu$ in view there exists a unique OT plan $\pi^*$ minimizing (3) and it is deterministic, i.e., $\pi^* = [id_{\mathbb{R}^D} , T^*] \# \mu$. Here $T^*$ is an OT map
which minimizes (1). Previous related works (Makkuva et al., 2020; Korotin et al., 2021a) assumed the absolute continuity of $\mu$, which implied the existence and uniqueness of $T^*$ (Brenier, 1991).

Let $\psi(y) \overset{\text{def}}{=} \frac{1}{2} \|y\|^2 - v(y)$, where $v$ is the potential of (4). Note that

$$v^c(x) = \inf_{y \in \mathbb{R}^D} \left\{ \frac{1}{2} \|x - y\|^2 - v(y) \right\} = \frac{1}{2} \|x\|^2 - \sup_{y \in \mathbb{R}^D} \{ \langle x, y \rangle - \psi(y) \} = \frac{1}{2} \|x\|^2 - \overline{\psi}(x).$$ (7)

Therefore, (4) is equivalent to

$$W_2^2(\mu, \nu) = \int_X \frac{\|x\|^2}{2} d\mu(x) + \int_Y \frac{\|y\|^2}{2} d\nu(y) + \sup_{\psi} \left\{ -\int_X \overline{\psi}(x) d\mu(x) - \int_Y \psi(y) d\nu(y) \right\} = \text{Constant}(\mu, \nu) - \inf_{\psi} \left\{ \int_X \psi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \right\}$$ (8)

$$\text{Constant}(\mu, \nu) - \inf_{\psi} \left\{ \int_X \sup_{y \in \mathbb{R}^D} \{ \langle x, y \rangle - \psi(y) \} d\mu(x) + \int_Y \psi(y) d\nu(y) \right\} = \text{Constant}(\mu, \nu) - \inf_{\psi} \left\{ \sup_T \int_X \{ \langle x, T(x) \rangle - \psi(T(x)) \} d\mu(x) + \int_Y \psi(y) d\nu(y) \right\}$$ (9)

$$\text{Constant}(\mu, \nu) - \inf_{\psi} \left\{ \sup_T \int_X \{ \langle x, T(x) \rangle - \psi(T(x)) \} d\mu(x) + \int_Y \psi(y) d\nu(y) \right\} = \text{Constant}(\mu, \nu) - \inf_{\psi} \left\{ \sup_T \int_X \{ \langle x, T(x) \rangle - \psi^* T(x) \} d\mu(x) + \int_Y \psi(y) d\nu(y) \right\}$$ (10)

where between lines (10) and (11) we replace the optimization over $y \in \mathbb{R}^D$ with the equivalent optimization over functions $T : \mathbb{R}^D \rightarrow \mathbb{R}^D$. The equivalence follows from the interchange between the integral and the supremum (Rockafellar, 1976 Theorem 3A). We also provide an independent proof of equivalence specializing Rockafellar’s interchange theorem in Appendix A.1. Thanks to the following lemma, we may solve saddle point problem (11) and obtain the OT map $T^*$ from its solution $(\psi^*, T^*)$.

**Lemma 4.1.** Let $T^*$ be the OT map from $\mu$ to $\nu$. Then, for every optimal potential $\psi^*$,

$$T^* \in \arg \sup_T \int_X \{ \langle x, T(x) \rangle - \psi^* T(x) \} d\mu(x).$$ (12)

We prove Lemma 4.1 in Appendix A.2. For general $\mu, \nu$ the $\arg \sup_T$ set for optimal $\psi^*$ might contain not only OT map $T^*$, but other functions as well. Working with real-world data in experiments (§5.2), we observe that despite this issue, optimization (11) still recovers $T^*$.

**Relation to previous works.** The use of the function $T$ to approximate the $c$-transform was proposed by Nhan Dam et al. (2019) to estimate the Wasserstein loss in WGANs. For $W_2$, the fact that $T^*$ is an OT map was used by Makkuva et al. (2020); Korotin et al. (2021a) who primarily assumed continuous $\mu$, $\nu$ and reduced (11) to convex $\psi$ and $T = \nabla \phi$ for convex $\phi$. Issues with non-uniqueness of solution of (12) were softened, but using ICNNs to parametrize $\psi$ became necessary. Korotin et al. (2021b) demonstrated that ICNNs negatively affect practical performance of OT and tested an unconstrained formulation similar to (11). As per the evaluation, it provided the best empirical performance (Korotin et al., 2021b §4.5). The method [MM:R] they consider parametrizes $\frac{1}{2} \| \cdot \|^2 - \psi(\cdot)$ by a neural network, while we directly parametrize $\psi(\cdot)$ by a neural network (§4.3).

Recent work by Fan et al. (2021) exploits formulation similar to (11) for general costs $c(\cdot, \cdot)$. While their formulation leads to a max-min scheme with general costs (Fan et al., 2021 Theorem 3), our approach gives rise to a min-max method for quadratic cost. In particular, we extend the formulation to learn OT maps between distributions in spaces with unequal dimensions, see the next subsection.

### 4.2 Unequal Dimensions of Input and Output Distributions

Consider the case when $\mathcal{X} = \mathbb{R}^H$ and $\mathcal{Y} = \mathbb{R}^D$ have different dimensions, i.e., $H \neq D$. In order to map the probability distribution $\mu$ to $\nu$, a straightforward solution is to embed $\mathcal{X}$ to $\mathcal{Y}$ via some $Q : \mathcal{X} \rightarrow \mathcal{Y}$ and then to fit the OT map between $Q \# \mu$ and $\nu$ for the quadratic cost on $\mathcal{Y} = \mathbb{R}^D$. In this case, the optimization objective becomes

$$\inf_{\psi} \sup_T \left\{ \int_X \{ \langle Q(x), T(Q(x)) \rangle - \psi(T(Q(x))) \} d\mu(x) + \int_Y \psi(y) d\nu(y) \right\}$$ (13)
with the optimal $T^*$ recovering the OT map from $Q_{#\mu}$ to $\nu$. For equal dimensions $H = D$ and the identity embedding $Q(x) \equiv x$, expression (13) reduces to optimization (11) up to a constant.

Instead of optimizing (13) over functions $G$, we consider to optimize directly over generative mappings $G : \mathcal{X} \rightarrow \mathcal{Y}$:

$$L(\psi, G) \overset{\text{def}}{=} \inf_{G} \sup_{\psi} \left\{ \int_{\mathcal{X}} \{ (Q(x), G(x)) - \psi(G(x)) \} d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) \right\} \quad (14)$$

Our following lemma establishes connections between (14) and OT with unequal dimensions:

**Lemma 4.2.** Assume that exists a unique OT plan between $Q_{#\mu}$ and $\nu$ and it is deterministic, i.e., $[id_{\mathcal{X}}, T^*]_{#\mu}(Q_{#\mu})$. Then $G^*(x) = T^* (Q(x))$ is the OT map between $\mu$ and $\nu$ for the $Q$-embedded quadratic cost $c(x, y) \equiv \frac{1}{2} \|Q(x) - y\|^2$. Moreover, for every optimal potential $\psi^*$ of problem (14),

$$G^* \in \arg\sup_{G} \int_{\mathcal{X}} \{ (Q(x), G(x)) - \psi^*(G(x)) \} d\mu(x). \quad (15)$$

We prove Lemma 4.2 in Appendix 4.3 and schematically present its idea in Figure 4.

Analogously to Lemma 4.1, it provides a way to compute the OT map $G^*$ for the $Q$-embedded quadratic cost between distributions $\mu$ and $\nu$ by solving the saddle point problem (14). Note the situation with non-uniqueness of $\arg\sup_{G} \psi^*$ is similar to 4.1.

**Relation to previous works.** In practice, learning OT maps directly between spaces of unequal dimensions was considered in the work by (Fan et al., 2021) §5.2 but only on toy examples. We demonstrate that our method works well in large-scale generative modeling tasks (§5.1). Theoretical properties of OT maps for embedded costs are studied, e.g., in (Pass, 2010; McCann & Pass, 2020).

### 4.3 Practical Aspects and Optimization Procedure

To optimize functional (14), we approximate $G : \mathbb{R}^H \rightarrow \mathbb{R}^D$ and $\psi : \mathbb{R}^D \rightarrow \mathbb{R}$ with neural networks $G_\theta$, $\psi_\omega$ and optimize their parameters via stochastic gradient descent-ascent (SGDA) by using minibatches from $\mu$, $\nu$. The practical optimization procedure is given in Algorithm 1 below. Following the usual practice in GANs, we add a small penalty (§3.3) on potential $\psi_\omega$ for better stability. The penalty is not included in Algorithm 1 to keep it simple.

**Relation to previous works.** WGAN by Arjovksy & Bottou (2017) uses $W_1$ as the loss to update the generator while we solve a different task — we fit the generator $G^*$ to be the OT map for $Q$-embedded quadratic cost. Despite this, our Algorithm 1 has similarities with WGAN’s training. The update of $\psi$ (line 4) coincides with discriminator’s update in WGAN. The update of generator $G$ (line 8) differs from WGAN’s update by the term $-(Q(\cdot), G_\theta(\cdot))$. Besides, in WGAN the optimization is $\inf_G \sup_D$. We have $\inf_\psi \sup_G$, i.e., the generator in our case is the solution of the inner problem.

### 4.4 Error Analysis

Given a pair $(\hat{\psi}, \hat{G})$ approximately solving (14), a natural question to ask is how good is the recovered OT map $\hat{G}$. In this subsection, we provide a bound on the difference between $G^*$ and $\hat{G}$ based on the duality gaps for solving outer and inner optimization problems.

In (8), and, as the result, in (10), (11), (13), (14), it is enough to consider optimization over convex functions $\psi$, see (Villani, 2008, Case 5.17). Our theorem below assumes the convexity of $\psi$ although it might not hold in practice since in practice $\psi$ is a neural network.
Algorithm 1: Learning the optimal transport map between unequal dimensions.

**Input:** Input distribution $\mu$ on $X = \mathbb{R}^{H}$; output distribution $\nu$ on $Y = \mathbb{R}^{D}$; generator network $G_{\theta} : \mathbb{R}^{H} \rightarrow \mathbb{R}^{D}$; potential network $\psi_{\omega} : \mathbb{R}^{D} \rightarrow \mathbb{R}$; number of iterations per network: $K_{G}, K_{\psi}$; embedding $Q : X \rightarrow Y$.

**Output:** Trained generator $G_{\theta}$ representing OT map from $\mu$ to $\nu$.

1 repeat
  for $k_{G} = 1$ to $K_{G}$ do
    Draw batch $\tilde{X} \sim \mu$ and $Y \sim \nu$;
    $L_{G} \leftarrow \frac{1}{|\tilde{X}|} \sum_{x \in \tilde{X}} \psi(G_{\theta}(x))$;
    Update $\theta$ by using $\frac{\partial L_{G}}{\partial \theta}$ to minimize $L_{G}$;
  for $k_{\psi} = 1$ to $K_{\psi}$ do
    Draw batch $\tilde{X} \sim \mu$;
    $L_{\psi} \leftarrow \frac{1}{|\tilde{X}|} \sum_{y \in \tilde{Y}} \psi_{\omega}(y) - \frac{1}{|\tilde{X}|} \sum_{x \in \tilde{X}} \psi(G_{\theta}(x))$;
    Update $\omega$ by using $\frac{\partial L_{\psi}}{\partial \omega}$ to minimize $L_{\psi}$;
  until not converged;
  return $G_{\theta}$.

**Theorem 4.3.** Assume that there exists a unique deterministic OT plan for $Q$-embedded quadratic cost between $\mu$ and $\nu$, i.e., $\pi^{*} = [id_{\mathbb{R}^{H}}, G^{*}]_{#}\mu$ for $G^{*} : \mathbb{R}^{H} \rightarrow \mathbb{R}^{D}$. Assume that $\psi$ is $\beta$-strongly convex ($\beta > 0$) and $\hat{G} : \mathbb{R}^{H} \rightarrow \mathbb{R}^{D}$. Define

$$
\epsilon_{1} = \sup_{G} L(\hat{G}, G) - L(\hat{G}, \hat{G}) \quad \text{and} \quad \epsilon_{2} = \inf_{G} \sup_{G} L(\psi, G) - \inf_{G} \sup_{G} L(\psi, G)
$$

Then the following bound holds true for the OT map $G^{*}$ from $\mu$ to $\nu$:

$$
\frac{\text{FID}(\hat{G}_{#}\mu, \nu)}{L^{2}} \leq 2 \cdot W_{2}^{2}(\hat{G}_{#}\mu, \nu) \leq \int_{X} \|\hat{G}(x) - G^{*}(x)\|^{2} d\mu(x) \leq \frac{2}{\beta}(\epsilon_{1}^{2} + \epsilon_{2}^{2}), \quad (16)
$$

where FID is the Fréchet inception distance [Heusel et al. 2017] and $L$ is the Lipschitz constant of the feature extractor of the pre-trained InceptionV3 neural network [Szegedy et al. 2016].

We prove Theorem 4.3 in Appendix A.4. The duality gaps upper bound $L^{2}(\mu)$ norm between computed $\hat{G}$ and true $G^{*}$ maps, and the $W_{2}$ between true $\nu$ and generated (fake) distribution $\hat{G}_{#}\mu$. Consequently, they upper bound FID between data $\nu$ and fake (generated) $\hat{G}_{#}\mu$ distributions.

**Relation to previous works.** Makkuva et al. [2020]; Korotin et al. [2021a] prove related bounds for $W_{2}$ with $\mu, \nu$ located on the spaces of the same dimension. Our result holds for different dimensions.

5 EXPERIMENTS

We evaluate our algorithm in generative modeling of the data distribution from a noise (5.1) and unpaired image restoration task (5.2). Technical details are given in Appendix B.4 Alternatively, in Appendix B.4 we test our method on toy 2D datasets and evaluate it on the Wasserstein-2 benchmark (Korotin et al. 2021b) in Appendix B.2. The code is in the supplementary material.

5.1 MODELING DATA DISTRIBUTION FROM NOISE DISTRIBUTION

In this subsection, $\mu$ is a 192-dimensional normal noise and $\nu$ the high-dimensional data distribution.

Let the images from $\nu$ be of size $w \times h$ with $c$ channels. As the embedding $Q : X \rightarrow Y$ we use a naive upscaling of a noise. For $x \in \mathbb{R}^{192}$ we represent it as 3-channel $8 \times 8$ image and bicubically upscale it to the size $w \times h$ of data images from $\nu$. For grayscale images drawn from $\nu$, we stack $c$ copies over channel dimension.

We test our method on MNIST $32 \times 32$ [LeCun et al. 1998], CIFAR10 $32 \times 32$ [Krizhevsky et al. 2009], and CelebA $64 \times 64$ [Liu et al. 2015] image datasets. In Figure 5, we show random samples.
Figure 5: Randomly generated MNIST, CIFAR10, and CelebA samples by our method (OTM).

Table 1: Results on CIFAR10 dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Related Work</th>
<th>Inception ↑</th>
<th>FID ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVAE</td>
<td>Vahdat &amp; Kautz (2020)</td>
<td>-</td>
<td>51.71</td>
</tr>
<tr>
<td>PixelQN</td>
<td>Ostrovski et al. (2018)</td>
<td>5.29</td>
<td>49.46</td>
</tr>
<tr>
<td>EBM</td>
<td>De &amp; Mordvintsev (2019)</td>
<td>6.02</td>
<td>40.58</td>
</tr>
<tr>
<td>DCGAN</td>
<td>Kundur et al. (2010)</td>
<td>6.64±0.14</td>
<td>37.70</td>
</tr>
<tr>
<td>NCSN</td>
<td>Song &amp; Ermon (2019)</td>
<td>8.87±0.12</td>
<td>25.32</td>
</tr>
<tr>
<td>NCP-VAE</td>
<td>An et al. (2021)</td>
<td>-</td>
<td>24.08</td>
</tr>
<tr>
<td>WGAN</td>
<td>Ostrovski et al. (2018)</td>
<td>-</td>
<td>55.2</td>
</tr>
<tr>
<td>WGAN-GP</td>
<td>Chalapathy et al. (2017)</td>
<td>6.49±0.09</td>
<td>39.40</td>
</tr>
<tr>
<td>3P-WGAN</td>
<td>Nhan Dinh et al. (2019)</td>
<td>7.38±0.08</td>
<td>28.8</td>
</tr>
<tr>
<td>AE-OT</td>
<td>An et al. (2020a)</td>
<td>-</td>
<td>28.5</td>
</tr>
<tr>
<td>AE-OT-GAN</td>
<td>An et al. (2020b)</td>
<td>-</td>
<td>17.1</td>
</tr>
<tr>
<td>OTM</td>
<td>Ours</td>
<td>7.42±0.06</td>
<td>21.78</td>
</tr>
</tbody>
</table>

Table 2: Results on CelebA dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Related Work</th>
<th>FID ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCGAN</td>
<td>Kundur et al. (2010)</td>
<td>52.0</td>
</tr>
<tr>
<td>DRAGAN</td>
<td>Kottari et al. (2017)</td>
<td>42.3</td>
</tr>
<tr>
<td>BEGAN</td>
<td>Benson et al. (2017)</td>
<td>38.9</td>
</tr>
<tr>
<td>NVAE</td>
<td>Vahdat &amp; Kautz (2020)</td>
<td>13.4</td>
</tr>
<tr>
<td>NCP-VAE</td>
<td>An et al. (2021)</td>
<td>5.2</td>
</tr>
<tr>
<td>WGAN</td>
<td>Arjovsky et al. (2017)</td>
<td>41.3</td>
</tr>
<tr>
<td>WGAN-GP</td>
<td>Chalapathy et al. (2017)</td>
<td>30.0</td>
</tr>
<tr>
<td>WGAN-QC</td>
<td>Liu et al. (2019)</td>
<td>12.9</td>
</tr>
<tr>
<td>AE-OT</td>
<td>An et al. (2020a)</td>
<td>28.6</td>
</tr>
<tr>
<td>AE-OT-GAN</td>
<td>An et al. (2020b)</td>
<td>7.8</td>
</tr>
<tr>
<td>OTM</td>
<td>Ours</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Generated by our approach, namely **Optimal Transport Modeling** (OTM). To quantify the results, in Tables 1 and 2 we give the inception (Salimans et al., 2016) and FID (Heusel et al., 2017) scores of generated samples. Similar to (Song & Ermon, 2019, Appendix B.2), we compute them on 50K real and generated samples. Additionally, in Appendix B.4 we test our method on 128 × 128 CelebA faces. We provide qualitative results (images of generated faces) in Figure 11.

For comparison, we include the scores of existing generative models of three types: (1) OT map as the generative model; (2) OT cost as the loss; (3) not OT-based. Note that models of the first type compute OT in the latent space of an autoencoder in contrast to our approach. According to our evaluation, the performance of our method is better or comparable to existing alternatives.

### 5.2 Unpaired Image Restoration

In this subsection, we consider unpaired image restoration tasks on CelebA faces dataset. In this case, the input distribution $\mu$ consists of degraded images, while $\nu$ are clean images. In all the cases, embedding $Q$ is a straightforward identity embedding $Q(x) \equiv x$.

In image restoration, optimality of the restoration map is desired since the output (restored) image is expected to be close to the input (degraded) one minimizing the transport cost. Note that GANs do not seek for an optimal mapping. However, in practice, due to implicit inductive biases such as convolutional architectures, GANs still tend to fit low transport cost maps (Bézenac et al., 2021).

The experimental setup is shown in Figure 6. We split the dataset in 3 parts A, B, C containing 90K, 90K, 22K samples respectively. To each image we apply the degradation transform (decolorization, noising or occlusion) and obtain the degraded dataset containing of 3 respective parts A, B, C. For unpaired training we use part A of degraded and part B of clean images. For testing, we use parts C.

Figure 6: The training/testing scheme that we use for unpaired restoration tasks.

To quantify the results we compute FID of restored images w.r.t. clean images of part C. The scores for denoising, inpainting and colorization are given in Table 3, details of each experiment and qualitative results are given below.

As a baseline, we include WGAN-GP. For a fair comparison, we fit it using exactly the same hyperparameters as in our method OTM-GP. This is possible due to the similarities between our method and WGAN-GP’s training procedure, see discussion in 4.3. In OTM, there is no GP (B.3).
Denoising. To create noisy images, we add white normal noise with $\sigma = 0.3$ to each pixel. Figure 7 illustrates image denoising using our OTM approach on the test part of the dataset. We show additional qualitative results for varying $\sigma$ in Figure 15 of §B.4.

![Figure 7: OTM for image denoising on test C part of CelebA, 64 × 64.](image)

(a) Noisy samples.  
(b) Pushforward samples.  
(c) Original samples.

Colorization. To create grayscale images, we average the RGB values of each pixel. Figure 8 illustrates image colorization using OTM on the test part of the dataset.

![Figure 8: OTM for image colorization on test C part of CelebA, 64 × 64.](image)

(a) Grayscale samples.  
(b) Pushforward samples.  
(c) Original samples.

Inpainting. To create incomplete images, we replace the right half of each clean image with zeros. Figure 9 illustrates image inpainting using OTM on the test part of the dataset.

![Figure 9: OTM for image inpainting on test C part of CelebA, 64 × 64.](image)

(a) Occluded samples.  
(b) Pushforward samples.  
(c) Original samples.

6 Conclusion

Our method fits OT maps for the embedded quadratic transport cost between probability distributions. Unlike predecessors, it scales well to high dimensions producing applications of OT maps directly in ambient spaces, such as spaces of images. The performance is comparable to other existing generative models while the complexity of training is similar to that of popular WGANs.

Limitations. For distributions $\mu, \nu$ we assume the existence of the OT map between them. In practice, this might not hold for all real-world $\mu, \nu$. Working with equal dimensions, we focus on the quadratic ground cost $\frac{1}{2}\|x - y\|^2$. Nevertheless, our approach extends to other costs $c(\cdot, \cdot)$, see Fan et al. (2021). When the dimensions are unequal, we restrict our analysis to embedded quadratic cost $\frac{1}{2}\|Q(x) - y\|^2$ where $Q$ equalizes dimensions. Choosing the embedding $Q$ might not be straightforward in some practical problems, but our evaluation (§5.1) shows that even naive choices of $Q$ work well.

Potential impact and ethics. Real-world image restoration problems often do not have paired datasets limiting the application of supervised techniques. In these practical unpaired learning problems, we expect our optimal transport approach to improve the performance of the existing models. However, biases in data might lead to biases in the pushforward samples. This should be taken into account when using our method in practical problems.
Reproducibility. The PyTorch source code is provided at

https://github.com/LituRout/OptimalTransportModeling

The instructions to use the code are included in the README.md file.

7 ACKNOWLEDGMENT

This research was supported by the computational resources provided by Space Applications Centre (SAC), ISRO. The first author acknowledges the funding by HRD Grant No. 0303T50FM703/SAC/ISRO. Skoltech RAIC center was supported by the RF Government (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021).

REFERENCES


Leygonie Jacob, Jennifer She, Amjad Almahairi, Sai Rajeswar, and Aaron Courville. W2gan: Recovering an optimal transport map with a gan. 2018.


Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to

Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu. A geometric view of optimal

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost.
In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October
2019.

Shu Liu, Shaqun Jin, Yongxin Chen, Hongyuan Zha, and Haomin Zhou. Learning high dimensional


Guansong Lu, Zhiming Zhou, Yuxuan Song, Kan Ren, and Yong Yu. Guiding the one-to-one mapping
in cyclegan via optimal transport. In Proceedings of the AAAI Conference on Artificial

Guansong Lu, Zhiming Zhou, Jian Shen, Cheng Chen, Weinan Zhang, and Yong Yu. Large-scale op-
2020.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping

Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q, p)-wasserstein gans: Com-

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smol-

Robert J McCann and Brendan Pass. Optimal transportation between unequal dimensions. Archive

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for

Quan Hoang Nhan Dam, Trung Le, Tu Dinh Nguyen, Hung Bui, and Dinh Phung. Threeplayer
wasserstein gan via amortised duality. In Proc. of the 28th Int. Joint Conf. on Artificial Intelligence
(IJCAI), 2019.

Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative

Brendan Pass. Regularity of optimal transportation between spaces with different dimensions. arXiv

Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max
Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In Uncertainty in Artificial

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of wasserstein gans. In

convolutional generative adversarial networks. In 4th International Conference on Learning
Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track

R Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In Nonlinear


A PROOFS

A.1 PROOF OF EQUIVALENCE: EQUATION (10) AND (11)

Proof. Pick any $T : \mathcal{X} \to \mathcal{Y}$. For every point $x \in \mathcal{X}$ by the definition of the supremum we have

$$\langle x, T(x) \rangle - \psi(T(x)) \leq \sup_{y \in \mathcal{Y}} \{ \langle x, y \rangle - \psi(y) \}.$$

Integrating the expression w.r.t. $x \sim \mu$ yields

$$\int_{\mathcal{X}} \{ \langle x, T(x) \rangle - \psi(T(x)) \} d\mu(x) \leq \int_{\mathcal{X}} \sup_{y \in \mathcal{Y}} \{ \langle x, y \rangle - \psi(y) \} d\mu(x) = \mathcal{L}_1.$$

Since the inequality holds for all $T : \mathcal{X} \to \mathcal{Y}$, we conclude that

$$\mathcal{L}_2 = \sup_{T : \mathcal{X} \to \mathcal{Y}} \int_{\mathcal{X}} \{ \langle x, T(x) \rangle - \psi(T(x)) \} d\mu(x) \leq \int_{\mathcal{X}} \sup_{y \in \mathcal{Y}} \{ \langle x, y \rangle - \psi(y) \} d\mu(x) = \mathcal{L}_1,$$

i.e. $\mathcal{L}_2 \leq \mathcal{L}_1$. Now let us prove that the sup on the left side actually equals $\mathcal{L}_1$. To do this, we need to show that for every $\epsilon > 0$ there exists $T^* : \mathcal{X} \to \mathcal{Y}$ satisfying

$$\int_{\mathcal{X}} \{ \langle x, T^*(x) \rangle - \psi(T^*(x)) \} d\mu(x) \geq \mathcal{L}_1 - \epsilon.$$

First note that for every $x \in \mathcal{X}$ by the definition of the supremum there exists $y' = y'(x)$ which provides

$$\langle x, y'(x) \rangle - \psi(y'(x)) \geq \sup_{y \in \mathcal{Y}} \{ \langle x, y \rangle - \psi(y) \} - \epsilon.$$

We take $T^*(x) = y'(x)$ for all $x \in \mathcal{X}$ and integrate the previous inequality w.r.t. $x \sim \mu$. We obtain

$$\int_{\mathcal{X}} \{ \langle x, T^*(x) \rangle - \psi(T^*(x)) \} d\mu(x) \geq \int_{\mathcal{X}} \sup_{y \in \mathcal{Y}} \{ \langle x, y \rangle - \psi(y) \} d\mu(x) - \epsilon = \mathcal{L}_1 - \epsilon,$$

which is the desired inequality.

A.2 PROOF OF LEMMA 4.1

Proof. It is enough to prove that $\overline{\psi^*}(x) = \langle T^*(x), x \rangle - \psi^*(T^*(x))$ holds $\mu$-almost everywhere, i.e., $T^*(x) \in \text{arg} \sup_{y \in \mathbb{R}^D} \{ \langle x, y \rangle - \psi^*(y) \}$. Since $\nu = \mathbb{W}_2^2(\mu, \mu)$, we use (9) with $\psi \leftarrow \psi^*$ to derive

$$W_2^2(\mu, \nu) - \int_{\mathcal{X}} \frac{1}{2} \| x \|^2 d\mu(x) - \int_{\mathcal{Y}} \frac{1}{2} \| y \|^2 d\nu(y) =$$

$$- \int_{\mathcal{X}} \overline{\psi^*}(x) d\mu(x) - \int_{\mathcal{Y}} \psi^*(y) d\nu(y) = - \int_{\mathcal{X}} \overline{\psi^*}(x) d\mu(x) - \int_{\mathcal{Y}} \psi^*(T^*(x)) d\mu(x) =$$

$$- \int_{\mathcal{X}} \left[ \overline{\psi^*}(x) + \psi^*(T^*(x)) \right] d\mu(x) \leq - \int_{\mathcal{X}} \langle T^*(x), x \rangle d\mu(x) =$$

$$\int_{\mathcal{X}} \| x - T^*(x) \|^2 d\mu(x) - \int_{\mathcal{X}} \frac{1}{2} \| x \|^2 d\mu(x) - \int_{\mathcal{Y}} \frac{1}{2} \| T^*(x) \|^2 d\mu(x) =$$

$$W_2^2(\mu, \nu) - \int_{\mathcal{X}} \frac{1}{2} \| x \|^2 d\mu(x) - \int_{\mathcal{Y}} \frac{1}{2} \| y \|^2 d\nu(y).$$

As a result, inequality (17) becomes the equality, in particular, $\overline{\psi^*}(x) + \psi^*(T^*(x)) = \langle T^*(x), x \rangle$ holds $\mu$-almost everywhere.

\qed
A.3 Proof of Lemma 4.2

Proof. Let \( QW_2^2 \) denote the \( Q \)-embedded quadratic cost. We use the change of variables formula to derive

\[
QW_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times Y} \frac{1}{2} \|Q(x) - y\|^2 d\pi(x, y) = \inf_{\pi \in \Pi(Q\#\mu, \nu)} \int_{X \times Y} \frac{1}{2} \|x - y\|^2 d\pi(x, y) = \mathcal{W}_2(Q\#\mu, \nu),
\]

i.e., computing the OT plan for \( QW_2^2(\mu, \nu) \) boils down to computing the OT plan for \( \mathcal{W}_2(Q\#\mu, \nu) \). It follows that \([\text{id}_{\mathbb{R}^H}, T^*(Q(x))]\#\mu = [\text{id}_{\mathbb{R}^H}, G^*\#\mu] \) is an OT plan for \( QW_2^2(\mu, \nu) \), and \( G^* \) is the OT map. Inclusion \([13]\) now follows from Lemma \([4.1]\) \( \square \)

A.4 Proof of Theorem 4.3

Proof. Pick any \( G' \in \text{arg sup}_G \mathcal{L}(\hat{\psi}, G) = \text{arg sup}_G \int_X \{ \langle Q(x), G(x) \rangle - \hat{\psi}(G(x)) \} \ d\mu(x) \) or, equivalently, for all \( x \in \mathbb{R}^H \), \( G'(x) \in \text{arg sup}_y \{ \langle Q(x), y \rangle - \hat{\psi}(y) \} \). Consequently, for all \( y \in \mathbb{R}^D \)

\[
\langle Q(x), G'(x) \rangle - \hat{\psi}(G'(x)) \geq \langle Q(x), y \rangle - \hat{\psi}(y),
\]

which after regrouping the terms yields

\[
\hat{\psi}(y) \geq \hat{\psi}(G'(x)) + \langle Q(x), y - G'(x) \rangle.
\]

This means that \( Q(x) \) is contained in the subgradient \( \partial \hat{\psi} \) at \( G'(x) \) for a convex \( \hat{\psi} \). Since \( \hat{\psi} \) is \( \beta \)-strongly convex, for points \( G(x), G'(x) \in \mathbb{R}^D \) and \( Q(x) \in \partial \hat{\psi}(G'(x)) \) we derive

\[
\hat{\psi}(G(x)) \geq \hat{\psi}(G'(x)) + \langle Q(x), G(x) - G'(x) \rangle + \frac{\beta}{2} \|G'(x) - G(x)\|^2.
\]

Regrouping the terms, this gives

\[
[\langle Q(x), G'(x) \rangle - \hat{\psi}(G'(x))] - [\langle Q(x), G(x) \rangle - \hat{\psi}(G(x))] \geq \frac{\beta}{2} \|G'(x) - G(x)\|^2.
\]

Integrating w.r.t. \( x \sim \mu \) yields

\[
\epsilon_1 = \mathcal{L}(\hat{\psi}, G') - \mathcal{L}(\hat{\psi}, G) \geq \beta \int_X \frac{1}{2} \|G'(x) - G(x)\|^2 d\mu(x) = \frac{\beta}{2} \cdot \|G - G'\|^2_{L^2(\mu)}.
\]

Let \( G^* \) be the OT map from \( \mu \) to \( \nu \). We use \( G^*\#\mu = \nu \) to derive

\[
\mathcal{L}(\hat{\psi}, G') = \int_X \{ \langle Q(x), G'(x) \rangle - \hat{\psi}(G'(x)) \} \ d\mu(x) + \int_Y \hat{\psi}(y) dv(y) = 
\int_X \{ \langle Q(x), G'(x) \rangle - \hat{\psi}(G'(x)) \} \ d\mu(x) + \int_X \hat{\psi}(G(x)) d\mu(x) =
\int_X \left\{ \frac{\langle Q(x), G'(x) \rangle - \hat{\psi}(G'(x)) + \hat{\psi}(G(x))}{\geq (Q(x), G'(x)) + \frac{\beta}{2} \|G' - G^*\|^2} \right\} d\mu(x) \geq
\int_X \langle Q(x), G^*(x) \rangle d\mu(x) + \beta \int_X \frac{1}{2} \|G' - G^*\|^2 d\mu(x).
\]

Let \( \psi^* \) be an optimal potential in \([14]\). Thanks to Lemma \([4.2]\) we have

\[
\inf \sup \mathcal{L}(\psi, G) = \mathcal{L}(\psi^*, G^*) =
\int_X \{ \langle Q(x), G^*(x) \rangle - \psi^*(G^*(x)) \} \ d\mu(x) + \int_Y \psi^*(y) dv(y) =
\]
\[
\int_X \{ \langle Q(x), G^*(x) \rangle - \psi^*(G^*(x)) \} \, d\mu(x) + \int_X \psi^*(G^*(x)) \, d\mu(x) = \\
\int_X \langle Q(x), G^*(x) \rangle \, d\mu(x)
\]

By combining (20) with (21), we obtain
\[
\epsilon_2 = \mathcal{L}(\hat{\psi}, G) - \mathcal{L}(\psi^*, G^*) \geq \beta \int_X \frac{1}{2} \|G^* - G\|^2 \, d\mu(x) = \frac{\beta}{2} \cdot \|G^* - G\|_2^2(\mu) \tag{22}
\]

The right-hand inequality of (16) follows from the triangle inequality combined with (19) and (22). The middle inequality of (16) follows from (Korotin et al., 2021a, Lemma A.2) and \( G_{\#} \mu = \nu \).

Now we prove the left-hand inequality of (16). Let \( \mathcal{I} \) be the feature extractor of the pre-trained InceptionV3 neural networks. FID score between generated (fake) \( \hat{\mathcal{I}} \) with betas \( \lambda = 01 \) and \( \psi \) with betas \( \lambda = 10 \), i.e. \( k_\psi = 1 \). The images are first cropped at the center with size 140 and then resized to \( 128 \times 128 \) for \( \mathcal{I} \).

We use the PyTorch framework. All the experiments are conducted on 2 × V100 GPUs. We compute inception and FID scores with the official implementation from OpenAI\(^1\) and TTUR\(^2\). The compared results are taken from the respective papers or publicly available source codes.

### B.1 General Training Details

**MNIST** [LeCun et al., 1998]. On MNIST, we use \( x \in \mathbb{R}^{192} \) and \( y \in \mathbb{R}^{32 \times 32} \). The batch size is 64, learning rate \( 2 \cdot 10^{-4} \), optimizer Adam [Kingma & Ba, 2014] with betas \((0, 0.9)\), gradient optimality coefficient \( \lambda = 10 \), and the number of training epochs \( T = 30 \). We observe stable training while updating \( \psi \) once in multiple \( G \) updates, i.e., \( k_G = 2 \) and \( k_\psi = 1 \).

**CIFAR10** [Krizhevsky et al., 2009]. We use all 50000 samples while training. The latent vector \( x \in \mathbb{R}^{192} \) and \( y \in \mathbb{R}^{128 \times 128} \), batch size 64, \( \lambda = 10 \), \( k_G = 1 \), \( k_\psi = 1 \), \( T = 1000 \), Adam optimizer with betas \((0, 0.9)\), and learning rate \( 2 \cdot 10^{-4} \) for \( G \) and \( 1 \cdot 10^{-3} \) for \( \psi \).

**CelebA** [Liu et al., 2015]. We use \( x \in \mathbb{R}^{192} \) and \( y \in \mathbb{R}^{64 \times 64 \times 3} \). The images are first cropped at the center with size 140 and then resized to \( 64 \times 64 \). We consider all 202599 samples. We use Adam with betas \((0, 0.9)\), \( T = 200 \), \( K_G = 2 \), \( K_\psi = 1 \) and learning rate \( 2 \cdot 10^{-4} \).

**Image restoration.** In the unpaired image restoration experiments, we use Adam optimizer with betas \((0, 0.9)\), \( K_G = 5 \), \( K_\psi = 1 \), \( \lambda = 0 \), learning rate \( 1 \cdot 10^{-4} \) and train for \( T = 300 \) epochs.

**CelebA128x128** [Liu et al., 2015]. On this dataset, we resize the cropped images as in CelebA to 128 × 128, i.e. \( y \in \mathbb{R}^{128 \times 128 \times 3} \). Here, \( K_G = 5 \), \( K_\psi = 1 \), \( \lambda = 0.01 \), learning rate \( 1 \cdot 10^{-4} \) and betas=(0.5, 0.999). The batch size is reduced to 16 so as to fit in the GPU memory.

**Anime128x128**. This dataset consists of 500000 high resolution images. We resize the cropped images as in CelebA to 128 × 128, i.e. \( y \in \mathbb{R}^{128 \times 128 \times 3} \). Here, \( K_G = 5 \), \( K_\psi = 1 \), \( \lambda = 0.01 \), learning rate \( 2 \cdot 10^{-4} \), batch size 16, and betas=(0, 0.9).

**Toy datasets.** The dimension is \( D = H = 2 \), total number of samples is 10000. We use the batch size 400, \( \lambda = 0.1 \), \( K_\psi = 1 \), \( K_G = 16 \), and

\(^1\)IS: https://github.com/openai/improved-gan/tree/master/inception_score
\(^2\)FID: https://github.com/bioinf-jku/TTUR
\(^3\)Anime: https://www.kaggle.com/reitanaka/alignedanimefaces
\( T = 100 \). The optimizer is Adam with betas \((0.5, 0.99)\) and learning rate \(1 \cdot 10^{-3}\). We use the following datasets: Gaussian to mixture of Gaussians \(^{[5]}\) two moons (sklearn.datasets.make_moons), circles (sklearn.datasets.make_circles), gaussian to S-curve (sklearn.datasets.make_s_curve), and gaussian to swiss roll (sklearn.datasets.make_swiss_roll).

**Wasserstein-2 benchmark (Appendix B.2).** The dimension is \( D = H = 64 \times 64 \times 3 \). We use the following datasets: Gaussian to mixture of Gaussians \(^{[4]}\), two moons \((\text{sklearn.datasets.make_moons})\), circles \((\text{sklearn.datasets.make_circles})\), gaussian to S-curve \((\text{sklearn.datasets.make_s_curve})\), and gaussian to swiss roll \((\text{sklearn.datasets.make_swiss_roll})\).

**B.2 Evaluation on the Continuous Wasserstein-2 Benchmark**

To empirically show that the method recovers the optimal transport maps well on equal dimensions, we evaluate it on the recent continuous Wasserstein-2 benchmark by Korotin et al. \((2021b)\). The benchmark provides a number of artificial test pairs \((\mu, \nu)\) of continuous probability distributions with analytically known OT map \( T^* \) between them.

<table>
<thead>
<tr>
<th>Method</th>
<th>( L^2)-UVP</th>
<th>( % )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM:R</td>
<td>1.4%</td>
<td></td>
</tr>
<tr>
<td>OTM (ours)</td>
<td>1.32%</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: \( L^2\)-UVP metric of the recovered transport map on the "Early" images benchmark pair.

For evaluation, we use the "Early" images benchmark pair \((D = 12288)\), see \([\text{Korotin et al., 2021b, M4.1}]\) for details. We adopt the \( L^2\)-unexplained percentage metric \((\text{Korotin et al., 2021a, M5.1})\) to quantify the recovered OT map \( \hat{T} \):

\[
L^2\text{-UVP}(\hat{T}) = 100 \cdot \frac{\|\hat{T} - T^*\|^2}{\text{Var}(\nu)}\%.
\]

For our method the \( L^2\)-UVP metric is only \( \approx 1\% \), see Table 4. This is comparable to the best \([\text{MM:R}]\) method which the authors evaluate on their benchmark. The qualitative results are given in Figure 10.

Figure 10: Qualitative results of OTM on the "Early" images benchmark pair \((\mu, \nu)\) by \([\text{Korotin et al. (2021b)}]\). The 1st line shows samples \( x \sim \mu \), the 2nd line shows fitted OT map \( \hat{T}(x) \), and the 3rd line shows the corresponding optimal map \( T^*(x) \sim \nu \).

**B.3 Further Discussion and Evaluation**

**Generative modeling.** In the experiments, we use the gradient penalty on \( \psi \) for better stability of optimization. The penalty is intended to make the gradient norm of the optimal WGAN critic equal to 1 \((\text{Gulrajani et al., 2017, Corollary 1})\). This condition does not necessarily hold for optimal \( \psi^* \) in our case and consequently might introduce bias to optimization.

To address this issue, we additionally tested an alternative regularization which we call the gradient optimality. For every optimal potential \( \psi^* \) and map \( G^* \) of problem \((14)\), we get from Lemma 4.2

\[
\nabla_G \left\{ \mathbb{E}_{x \sim \mu} \left[ (Q(x), G^*(x)) - \psi^*(G^*(x)) \right] \right\} = \mathbb{E}_{x \sim \mu} [Q(x)] - \mathbb{E}_{x \sim \mu} [\nabla \psi^*(G^*(x))].
\]

Since \( \mu \) is normal noise distribution and \( Q(x) \) is naive upscaling \((\text{M5.1})\), the above expression simplifies to \( \mathbb{E}_{x \sim \mu} \nabla \psi^*(G^*(x)) = 0 \). Based on this property, we establish the following regularizer \( \lambda \| \mathbb{E}_{x \sim \mu} \nabla \psi(G(x)) \| \) for \( \lambda > 0 \) and add this to \( L_\psi \) in our Algorithm 1.

While gradient penalty considers expectation of norm, gradient optimality considers norm of expectation. The gradient optimality is always non-negative and vanishes at the optimal point.

\[^{[5]}\text{https://github.com/AmirTag/OT-ICNN}^{[6]}\]
We conduct additional experiments with the gradient optimality and compare FID scores for different $\lambda$ in Table 5. It leads to an improvement of FID score from the earlier 7.7 with the gradient penalty to the current 6.5 with the gradient optimality on CelebA (Table 2).

**Unpaired restoration.** In the unpaired restoration experiments (§5.2), we test OTM with the gradient penalty to make a fair comparison with the baseline WGAN-GP: We find OTM without regularization, i.e., $\lambda = 0$ works better than OTM-GP (Table 3). In practice, more $G$ updates for a single $\psi$ update works fairly well (8B.1).

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>FID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>16.91</td>
</tr>
<tr>
<td>0.01</td>
<td>16.22</td>
</tr>
<tr>
<td>0.1</td>
<td>16.70</td>
</tr>
<tr>
<td>1.0</td>
<td>10.01</td>
</tr>
<tr>
<td>10</td>
<td>6.50</td>
</tr>
</tbody>
</table>

Table 5: Ablation study of gradient optimality in OTM.

### B.4 Additional Qualitative Results

OTM works with both the grayscale and color embeddings of noise in the ambient space.

**CelebA128x128.** Figure 11 shows the grayscale embedding $Q(x)$, the recovered transport map $\hat{G}(x)$, and independently drawn real samples $y \sim \nu$.

![OTM between 128-dimensional noise and CelebA, 128 × 128.](image)

Figure 11: OTM between 128-dimensional noise and CelebA, 128 × 128. The 1st line shows the grayscale embedding $Q(x)$ (repeating bicubic upscaling of a noise, 16 × 8), the 2nd line shows corresponding generated samples, and the 3rd line shows random samples from the dataset.

**Anime128x128.** Figure 12 shows the color embedding $Q(x)$, the recovered transport map $\hat{G}(x)$, and independently drawn real samples $y \sim \nu$.

![OTM between 192-dimensional noise and Anime, 128 × 128.](image)

Figure 12: OTM between 192-dimensional noise and Anime, 128 × 128. The 1st line shows the color embedding $Q(x)$ (bicubic upscaling of a noise, 3 × 8 × 8), the 2nd line shows corresponding generated samples, and the 3rd line shows random samples from the dataset.

The extended qualitative results with color embedding on MNIST, CIFAR10, and CelebA are shown in Figure 13a, Figure 13b, and Figure 13c respectively. Table 6 shows quantitative results on MNIST. The color embedding $Q$ is bicubic upscaling of a noise in $\mathbb{R}^{3×8×8}$. The samples are...
Table 6: Results on MNIST dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Related Work</th>
<th>FID ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>Kingma &amp; Welling (2013)</td>
<td>23.8 ± 0.6</td>
</tr>
<tr>
<td>LSGAN</td>
<td>Ma et al. (2017)</td>
<td>7.8 ± 0.6</td>
</tr>
<tr>
<td>BEGAN</td>
<td>Berthelot et al. (2017)</td>
<td>13.1 ± 1.0</td>
</tr>
<tr>
<td>WGAN</td>
<td>Arjovsky et al. (2017)</td>
<td>6.7 ± 0.4</td>
</tr>
<tr>
<td>SIG</td>
<td>Dai &amp; Seljak (2021)</td>
<td>4.5</td>
</tr>
<tr>
<td>AE-OT</td>
<td>An et al. 2020a</td>
<td>6.2 ± 0.2</td>
</tr>
<tr>
<td>AE-OT-GAN</td>
<td>An et al. 2020b</td>
<td>3.2</td>
</tr>
<tr>
<td>OTM</td>
<td>Ours</td>
<td>2.4</td>
</tr>
</tbody>
</table>

generated randomly (uncurated) by fitted optimal transport maps between noise and ambient space, e.g., spaces of high-dimensional images. Figure 14 illustrates latent space interpolation between the generated samples. Figure 15 shows denoising of images with varying levels of $\sigma = 0.1, 0.2, 0.3, 0.4$ by the model trained with $\sigma = 0.3$.

(a) MNIST, 32 × 32, grayscale  
(b) CIFAR10, 32 × 32, RGB  
(c) CelebA, 64 × 64, RGB

Figure 13: Randomly generated MNIST, CIFAR10, and CelebA samples by our method (OTM).

Figure 14: OTM for latent space interpolation on CelebA, 64 × 64. Extended samples.

**Toy datasets.** Figure 16 shows the results of our method and related approaches (43) on a toy 2D dataset. Figure 17 shows the results of our method applied to other toy datasets.

### B.5 Neural Network Architectures

This section contains architectures on **CIFAR10** (Table 7), **CelebA** (Table 8), **CelebA 128 × 128** (Figure 11) generation, **image restoration** tasks (Table 9), evaluation on the toy 2D datasets and **Wasserstein-2 images benchmark** (Table 4).

In the unpaired restoration tasks (452), we use UNet architecture for transport map $G$ and convolutional architecture for potential $\psi$. Similarly to (Song & Ermon 2019), we use BatchNormilation (BN) and InstanceNormalization+ (INorm+) layers. In the ResNet architectures, we use the ResidualBlock of NCSN (Song & Ermon 2019).

In the toy 2D examples, we use a simple multi-layer perceptron with 3 hidden layers consisting of 128 neurons each and LeakyReLU activation. The final layer is linear without any activation.
Figure 15: OTM for image denoising for varying levels of noise on test C part of CelebA, 64 × 64.

Figure 16: Mapping between a Gaussian and a Mixture of 8 Gaussians in 2D by various methods. The colors green, blue, and peru represent input, pushforward, and output samples respectively.

The transport map $G$ and potential $\psi$ architectures on MNIST $32 \times 32$, CelebA $128 \times 128$, and Anime $128 \times 128$ are the generator and discriminator architectures of WGAN-QC [Liu et al. (2019)] respectively.

In the evaluation on the Wasserstein-2 benchmark, we use publicly available UNet5 architecture for transport map $T$ and WGAN-QC discriminator’s architecture for $\psi$ [Liu et al. 2019]. These neural network architectures are the same as the authors of the benchmark use.

5https://github.com/milesial/Pytorch-UNet
Figure 17: OTM on toy datasets, $D = 2$. Here, the colors green, blue, and peru represent input, pushforward, and output samples respectively.

Table 7: Architectures for generation task on CIFAR10, $32 \times 32$.

\[
G(\cdot)
\]
\begin{align*}
\text{Noise}: \ x & \in \mathbb{R}^{128} \\
\text{Linear, Reshape, output shape}: & \ [128 \times 4 \times 4] \\
\text{ResidualBlock Up, output shape}: & \ [128 \times 8 \times 8] \\
\text{ResidualBlock Up, output shape}: & \ [128 \times 16 \times 16] \\
\text{ResidualBlock Up, output shape}: & \ [128 \times 32 \times 32] \\
\text{Conv, Tanh, output shape}: & \ [3 \times 32 \times 32]
\end{align*}

\[
\psi(\cdot)
\]
\begin{align*}
\text{Target}: \ y & \in \mathbb{R}^{3 \times 32 \times 32} \\
\text{ResidualBlock Down, output shape}: & \ [128 \times 16 \times 16] \\
\text{ResidualBlock Down, output shape}: & \ [128 \times 8 \times 8] \\
\text{ResidualBlock, output shape}: & \ [128 \times 8 \times 8] \\
\text{ResidualBlock, output shape}: & \ [128 \times 8 \times 8] \\
\text{ReLU, Global sum pooling, output shape}: & \ [128 \times 1 \times 1] \\
\text{Linear, output shape}: & \ [1]
\end{align*}
Table 8: Architectures for generation task on Celeba, $64 \times 64$.

<table>
<thead>
<tr>
<th>$G(.)$</th>
<th>Noise: $x \in \mathbb{R}^{128}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvTranspose, BN, LeakyReLU, output shape: $[256 \times 1 \times 1]$</td>
<td></td>
</tr>
<tr>
<td>ConvTranspose, BN, LeakyReLU, output shape: $[512 \times 4 \times 4]$</td>
<td></td>
</tr>
<tr>
<td>Conv, PixelShuffle, BN, LeakyReLU, output shape: $[512 \times 8 \times 8]$</td>
<td></td>
</tr>
<tr>
<td>Conv, PixelShuffle, BN, LeakyReLU, output shape: $[512 \times 16 \times 16]$</td>
<td></td>
</tr>
<tr>
<td>Conv, PixelShuffle, BN, LeakyReLU, output shape: $[512 \times 32 \times 32]$</td>
<td></td>
</tr>
<tr>
<td>ConvTranspose, Tanh, output shape: $[3 \times 64 \times 64]$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\psi(.)$</th>
<th>Target: $y \in \mathbb{R}^{3 \times 64 \times 64}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv, output shape: $[128 \times 64 \times 64]$</td>
<td></td>
</tr>
<tr>
<td>ResidualBlock Down, output shape: $[256 \times 32 \times 32]$</td>
<td></td>
</tr>
<tr>
<td>ResidualBlock Down, output shape: $[256 \times 16 \times 16]$</td>
<td></td>
</tr>
<tr>
<td>ResidualBlock Down, output shape: $[256 \times 8 \times 8]$</td>
<td></td>
</tr>
<tr>
<td>ResidualBlock Down, output shape: $[128 \times 4 \times 4]$</td>
<td></td>
</tr>
<tr>
<td>Conv, output shape: $[1]$</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Architectures for restoration tasks on CelebA, $64 \times 64$.

<table>
<thead>
<tr>
<th>$G(.)$</th>
<th>Input: $x \in \mathbb{R}^{3 \times 64 \times 64}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv, BN, LeakyReLU, output shape: $[256 \times 64 \times 64]$</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 32 \times 32]$, x1</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 16 \times 16]$, x2</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 8 \times 8]$, x3</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 4 \times 4]$, x4</td>
<td></td>
</tr>
<tr>
<td>Nearest Neighbour Upsample, Conv, BN, ReLU, output shape: $[256 \times 8 \times 8]$, y3</td>
<td></td>
</tr>
<tr>
<td>Add (y3, x3), output shape: $[256 \times 8 \times 8]$, y3</td>
<td></td>
</tr>
<tr>
<td>Nearest Neighbour Upsample, Conv, BN, ReLU, output shape: $[256 \times 16 \times 16]$, y2</td>
<td></td>
</tr>
<tr>
<td>Add (y2, x2), output shape: $[256 \times 16 \times 16]$, y2</td>
<td></td>
</tr>
<tr>
<td>Nearest Neighbour Upsample, Conv, BN, ReLU, output shape: $[256 \times 32 \times 32]$, y1</td>
<td></td>
</tr>
<tr>
<td>Add (y1, x1), output shape: $[256 \times 32 \times 32]$, y1</td>
<td></td>
</tr>
<tr>
<td>Nearest Neighbour Upsample, Conv, BN, ReLU, output shape: $[256 \times 64 \times 64]$, y</td>
<td></td>
</tr>
<tr>
<td>Add (y, x), output shape: $[256 \times 64 \times 64]$, y</td>
<td></td>
</tr>
<tr>
<td>ConvTranspose, Tanh, output shape: $[3 \times 64 \times 64]$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\psi(.)$</th>
<th>Target: $y \in \mathbb{R}^{3 \times 64 \times 64}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 32 \times 32]$</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 16 \times 16]$</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 8 \times 8]$</td>
<td></td>
</tr>
<tr>
<td>Conv, LeakyReLU, AvgPool, output shape: $[256 \times 4 \times 4]$</td>
<td></td>
</tr>
<tr>
<td>Linear, output shape: $[1]$</td>
<td></td>
</tr>
</tbody>
</table>