Neural Optimal Transport and Applications

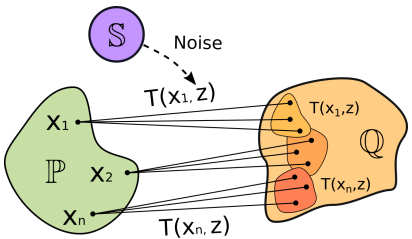
Evgeny Burnaev
Prof., Head of Skoltech Applied AI Center
Head of AIRI Research Group

Neural Optimal Transport

https://arxiv.org/abs/2201.12220

Alexander Korotin, Daniil Selikhanovych, Evgeny Burnaev

We present a novel neural-networks-based algorithm to compute optimal transport maps and plans for strong and weak transport costs.



Overview

Generative Modeling: Intro

What is Optimal Transport and
Why Do We Need It?

Background on OT

An Algorithm to Learn OT Plans

Experiments

Conclusion and extensions

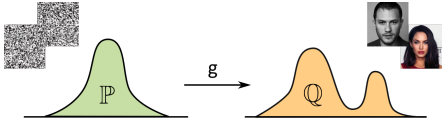
Generative Modeling: Intro

The Task of Generative Modeling

Map the given distribution $\mathbb{P}$ into the given distribution $\mathbb{Q}$.

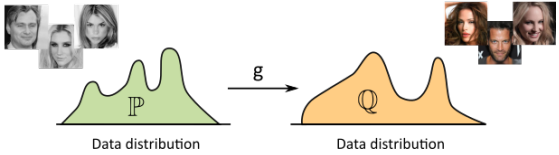
Case 1: noise $\rightarrow$ data

synthetic data generation/data manipulation

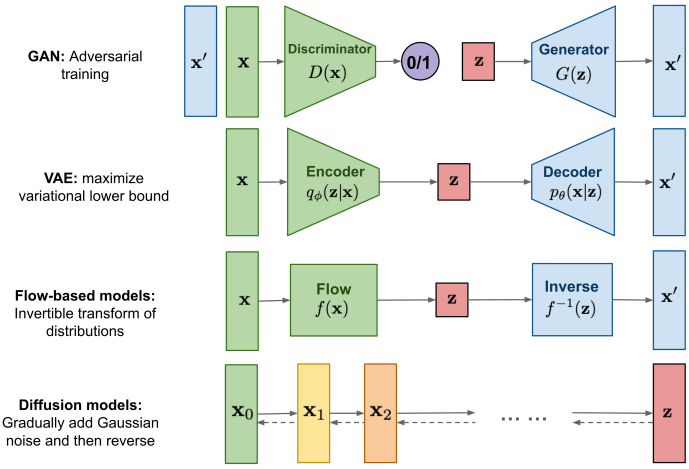


Case 2: data $\rightarrow$ data

unpaired style transfer, super-resolution, domain adapatation



Types of Generative Models[1]

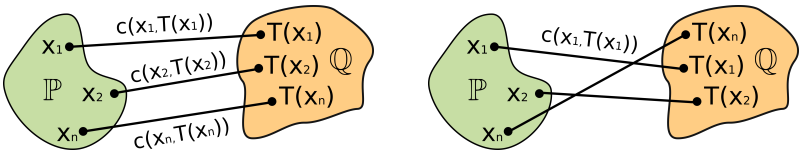


[1] https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

What is Optimal Transport and Why Do We Need It?

Monge's Formulation of Optimal Transport[2]

Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function, e.g., $c(x, y) = \frac{\|x-y\|^2}{2}$.



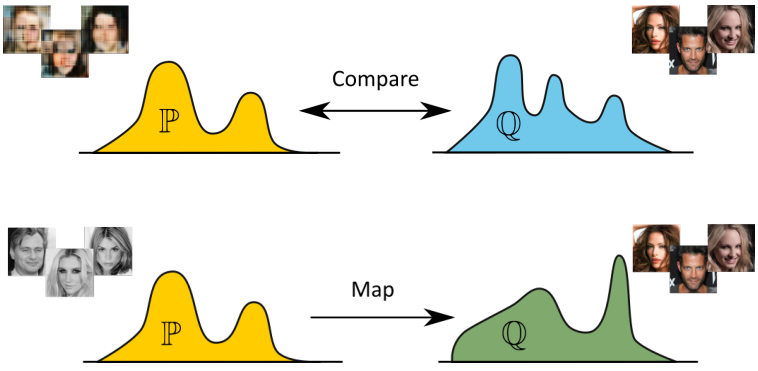
The optimal transport **cost** between measures $\mathbb{P}$ and $\mathbb{Q}$ is

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \inf_{T \sharp \mathbb{P} = \mathbb{Q}} \int_{\mathcal{X}} c\big(x, T(x)\big) d\mathbb{P}(x).$$

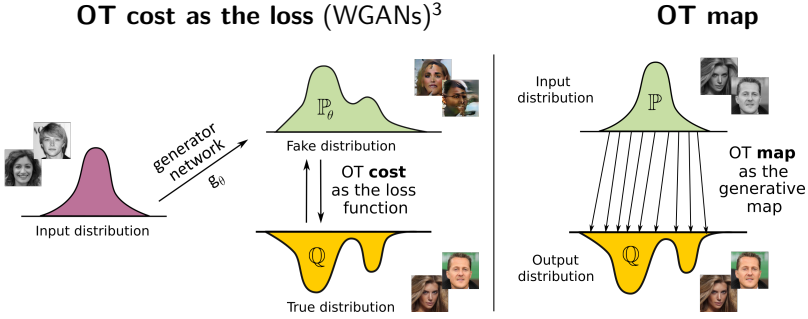
The map T^* attaining the minimum is called the optimal **transport map**.

[2] Cédric Villani (2008). *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media.

Optimal Transport in Machine Learning Tasks



Approaches to Use OT in Generative Models



Wasserstein GANs are not Optimal Transport![4]

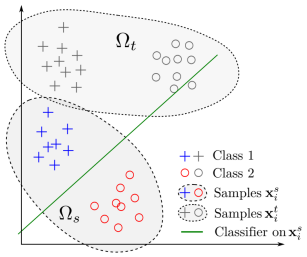
[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

[4] Alexander Korotin, Alexander Kolesov, and Evgeny Burnaev (2022). "Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?" In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=VtEEpi-dGlt.

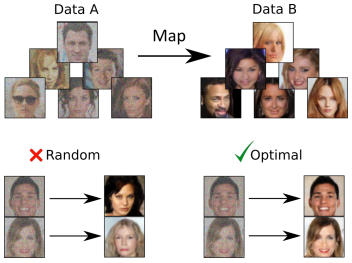
Why Do We Need Optimal Maps?

Practical tasks[5]

Domain Adaptation



Unpaired Learning



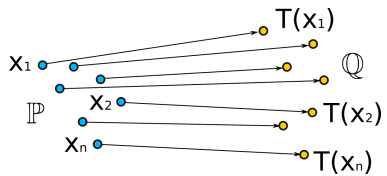
Theoretical side

rigorous OT theory, guarantees of performance.

[5] Nicolas Courty et al. (2016). "Optimal transport for domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.9, pp. 1853–1865.

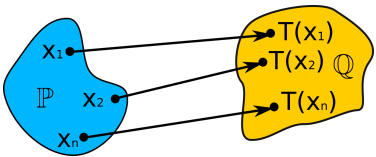
Types of Optimal Transport Methods

Discrete



+ Convex optimization;
+ Strong theoretical guarantees;
- Poor scalability;
- No out-of-support estimates;

Continuous (**Parametric**)



± **Neural networks**;
± Limited theoretical guarantees;
+ Good scalability;
+ Out-of-sample estimation

Development of Scalable Neural Methods

The presentation is based on the **following papers**:

[1] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2022b). "Neural Optimal Transport". In: *arXiv preprint arXiv:2201.12220*

[2] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2022a). "Kernel Neural Optimal Transport". In: *arXiv preprint arXiv:2205.15269*

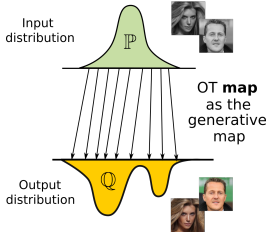
[3] Arip Asadulaev et al. (2022). "Neural Optimal Transport with General Cost Functionals". In: *arXiv preprint arXiv:2205.15403*

[4] Milena Gazdieva et al. (2022). "Unpaired
Image Super-Resolution with Optimal Transport Maps". In: *arXiv preprint arXiv:2202.01116*

[5] Litu Rout, Alexander Korotin, and Evgeny Burnaev
(2022). "Generative Modeling with Optimal Transport Maps". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=5JdLZg346Lw

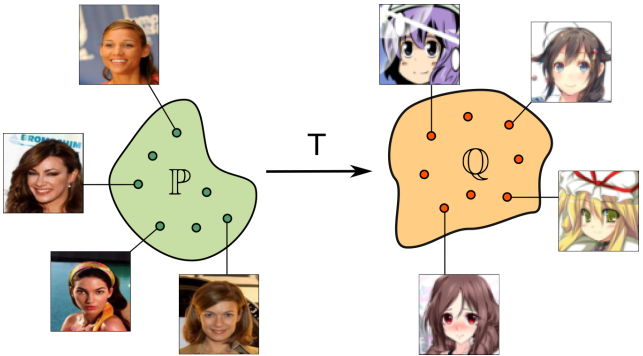
[5] Alexander Korotin,
Lingxiao Li, et al. (2021). "Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark". In: *Advances in Neural Information Processing Systems* 34

[6] Alexander Korotin, Vage Egiazarian,
et al. (2021). "Wasserstein-2 Generative Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=bEoxzW_EXsa



Example: Unpaired Image-to-Image Style Translation

Input: two unpaired datasets – empirical samples from $\mathbb{P}, \mathbb{Q}$.



Output: «*style translation map*» $T : \mathcal{X} \to \mathcal{Y}$ satisfying $T_{\#}\mathbb{P} = \mathbb{Q}$.

Preliminary Examples (128×128 images)

Deterministic (one-to-one)

Celeba (female) → anime

Outdoor → churches

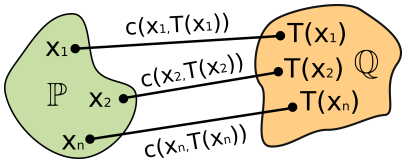
Stochastic (one-to-many)

Handbags → shoes

Background on OT

Issues of Monge's Formulation

Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function, e.g., $c(x, y) = \frac{\|x-y\|^2}{2}$.



The optimal transport **cost** between measures $\mathbb{P}$ and $\mathbb{Q}$ is

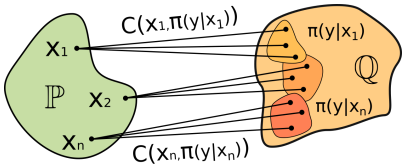
$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \inf_{T\sharp\mathbb{P}=\mathbb{Q}} \int_{\mathcal{X}} c\big(x, T(x)\big)\, d\mathbb{P}(x).$$

Issues of the formulation

- There might be no map T satisfying $T\sharp\mathbb{P} = \mathbb{Q}$ (no mass splitting).
  - The infimum might not be attained (no minimizer T^*).

Weak Formulation of Optimal Transport[6]

Let $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ be a **weak** cost function.



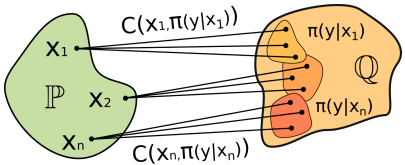
The optimal transport **cost** between measures $\mathbb{P}$ and $\mathbb{Q}$ is

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X}} C\big(x, \pi(\cdot|x)\big) \underbrace{d\pi(x)}_{d\mathbb{P}(x)}$$

where $\pi(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ denotes the conditional distribution of $y \in \mathcal{Y}$ conditioned on $x \in \mathcal{X}$.

[6] Nathael Gozlan et al. (2017). "Kantorovich duality for general transport costs and applications". In: *Journal of Functional Analysis* 273.11, pp. 3327–3405.

An Example: γ-Weak Wasserstein-2 ($\mathcal{W}_{2,\gamma}$)



We will focus on the γ-**weak quadratic cost** ($\gamma > 0$):

$$C\big(x, \pi(\cdot|x)\big) = \int_{\mathcal{Y}} \frac{1}{2}\|x - y\|^2 d\pi(y|x) - \frac{\gamma}{2}\mathrm{Var}\big(\pi(\cdot|x)\big),$$

where $\mathrm{Var}\big(\pi(\cdot|x)\big)$ denotes the **variance** of the $\pi(\cdot|x) \in \mathcal{P}(\mathcal{Y})$.

Cost $C(x, \mu)$ is **convex** in μ since $\mathrm{Var}(\mu)$ is a concave functional of μ.

Dual Form of the Weak Optimal Transport

Primal form

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X}} C\big(x, \pi(\cdot|x)\big)\, d\pi(x).$$

Dual form[7]

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \sup_{f} \int_{\mathcal{X}} f^C(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} f(y) d\mathbb{Q}(y),$$

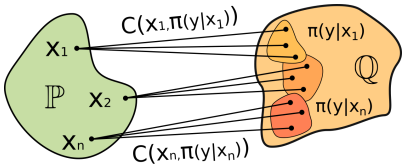
where $f : \mathcal{Y} \to \mathbb{R}$ are continuous, lower-bounded and not rapidly growing functions; f^C denotes the weak C-transform of f:

$$f^C(x) = \inf_{\mu \in \mathcal{P}(\mathcal{Y})} \left\{ C(x, \mu) - \int_{\mathcal{Y}} f(y) d\mu(y) \right\}.$$

[7] Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmun Pammer (2019). "Existence, duality, and cyclical monotonicity for weak transport costs". In: *Calculus of Variations and Partial Differential Equations* 58.6, pp. 1–28.

Dual Form of the Weak Optimal Transport

Primal form



Dual form

$$\sup_{f} \int_{\mathcal{X}} \inf_{\mu \in \mathcal{P}(\mathcal{Y})} \left\{ C(x,\mu) - \int_{\mathcal{Y}} f(y) d\mu(y) \right\} d\mathbb{P}(x) + \int_{\mathcal{Y}} f(y) d\mathbb{Q}(y),$$

Our goal

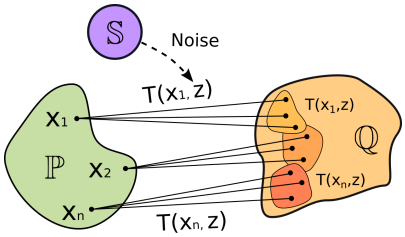
is to extract the **primal solution** (OT plan) π^* from the dual problem.

An Algorithm to Learn OT Plans

Stochastic Functions

We say that that $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is a **stochastic function**.

Stochastic functions can *implicitly*[8] represent transport plans $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$.



$$T(x, \cdot)_{\#}\mathbb{S} \sim \pi(\cdot|x).$$

If a stochastic function T^* represents some OT plan π^*,
we say that T^* is the **stochastic OT map**.

[8] Olav Kallenberg (1997). *Foundations of modern probability*. Vol. 2. Springer.

Reformulation of the Dual Problem

Lemma (Minimax reformulation of the dual problem)

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \sup_{f} \inf_{T} \mathcal{L}(f, T),$$

where the functional $\mathcal{L}$ is defined by

$$\mathcal{L}(f, T) \overset{def}{=} \int_{\mathcal{Y}} f(y) d\mathbb{Q}(y) + \int_{\mathcal{X}} \left(C\big(x, T(x, \cdot)_{\#}\mathbb{S}\big) - \int_{\mathcal{Z}} f\big(T(x, z)\big) d\mathbb{S}(z) \right) d\mathbb{P}(x).$$

Lemma (Stochastic OT maps solve the problem.)

For any potential f^* which attains the optimal value of the problem, and for any stochastic map T^* which realizes some OT plan π^*:

$$T^* \in \arg\inf_{T} \mathcal{L}(f^*, T).$$

The Algorithm: Preliminaries

$$\sup_{\omega}\inf_{\theta}\mathcal{L}(\omega,\theta)=\sup_{\omega}\inf_{\theta}\left[\int_{\mathcal{Y}}f_{\omega}(y)d\mathbb{Q}(y)+\right.$$
$$\left.\int_{\mathcal{X}}\left(C\big(x,T_{\theta}(x,\cdot)_{\#}\mathbb{S}\big)-\int_{\mathcal{Z}}f_{\omega}\big(T_{\theta}(x,z)\big)d\mathbb{S}(z)\right)d\mathbb{P}(x)\right].$$

- We use ResNet[9] $f_{\omega}:\mathbb{R}^{3\times W\times H}\to\mathbb{R}$;
- We use UNet $T_{\theta}:\mathbb{R}^{(3+1)\times H\times W}\to\mathbb{R}^{3\times W\times H}$.
  - The noise simply as an additional input channel (RGB**Z**);
  - We use a Gaussian noise $\mathbb{S}$ of $\dim=W\times H$ with axis-wise $\sigma=0.1$.
- We solve the saddle point problem with the **stochastic gradient ascent-descent** by using random batches from $\mathbb{P},\mathbb{Q},\mathbb{S}$.

[9] Huidong Liu, Xianfeng Gu, and Dimitris Samaras (2019). "Wasserstein gan with quadratic transport cost". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4832–4841.

The Algorithm: Neural Optimal Transport (NOT)

Algorithm 1: Neural optimal transport (**NOT**)

repeat

Sample batches $Y \sim \mathbb{Q}$, $X \sim \mathbb{P}$;

For each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;

$\mathcal{L}_f \leftarrow \frac{1}{|X|} \sum_{x \in X} \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega\big(T_\theta(x, z)\big) - \frac{1}{|Y|} \sum_{y \in Y} f_\omega(y)$;

Update ω by using $\frac{\partial \mathcal{L}_f}{\partial \omega}$;

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

Sample batches $X \sim \mathbb{P}$;

For each $x \in X$ sample batch $Z_x \sim \mathbb{S}$;

$\mathcal{L}_T \leftarrow \frac{1}{|X|} \sum_{x \in X} \Big[\widehat{C}\big(x, T_\theta(x, Z_x)\big) - \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega\big(T_\theta(x, z)\big)\Big]$;

Update θ by using $\frac{\partial \mathcal{L}_T}{\partial \theta}$;

until *not converged*;

Unbiased Empirical Estimator for the γ-Weak Quadratic Cost

Recall the γ-weak quadratic cost

$$C\big(x, \pi(\cdot|x)\big) = \int_{\mathcal{Y}} \frac{1}{2}\|x - y\|^2 d\pi(y|x) - \frac{\gamma}{2}\mathrm{Var}\big(\pi(\cdot|x)\big),$$

We use the estimator $\widehat{C}$ for the γ-weak quadratic cost $C\big(x, T(x,\cdot)_{\#}\mathbb{S}\big)$:

$$\widehat{C}\big(x, T(x, Z)\big) = \frac{1}{2|Z|}\sum_{z\in Z}\|x - T(x,z)\|^2 - \frac{\gamma}{2}\hat{\sigma}^2,$$

where $Z \sim \mathbb{S}$ is a random batch and $\hat{\sigma}^2$ is the (corrected) batch variance

$$\hat{\sigma}^2 = \frac{1}{|Z|-1}\sum_{z\in Z}\Big\| T(x,z) - \frac{1}{|Z|}\sum_{z\in Z} T(x,z)\Big\|^2.$$

Neural Optimal Transport is NOT a GAN

NOT is NOT a WGAN. It solves a different problem.

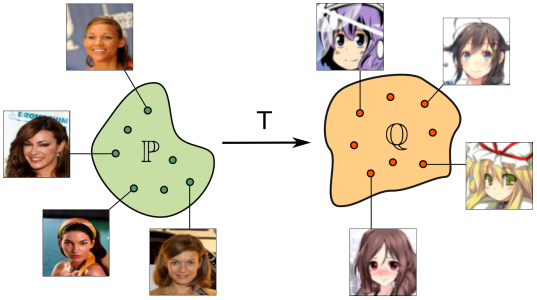
	Neural Optimal Transport [NOT]	**Wasserstein Generative Adversarial Nets** [WGAN]
The idea	OT plan as the generative model	OT cost as the loss for generator
Minimax optimization objective	$\max_{f} \min_{T} \mathcal{L}(f, T)$	$\min_{T} \max_{f} \mathcal{L}(T, f)$
Transport map T (generator)	T^* solves the inner problem; it is an OT map from $\mathbb{P}$ to $\mathbb{Q}$	T^* solves the outer problem; it is an arbitrary map from $\mathbb{P}$ to $\mathbb{Q}$
Potential f (discriminator)	Unconstrained f	Constrained $f \in \text{Lip}_1$

Note that, in general, swapping min and max is prohibited, i.e., $\max_f \min_T(\cdot) \neq \min_T \max_f(\cdot)$, for example,

$$1 = \min_{x} \max_{y} \sin(x + y) \neq \max_{y} \min_{x} \sin(x + y) = -1.$$

Experiments

General Details



Datasets: various 64×64 and 128×128 RGB images datasets.

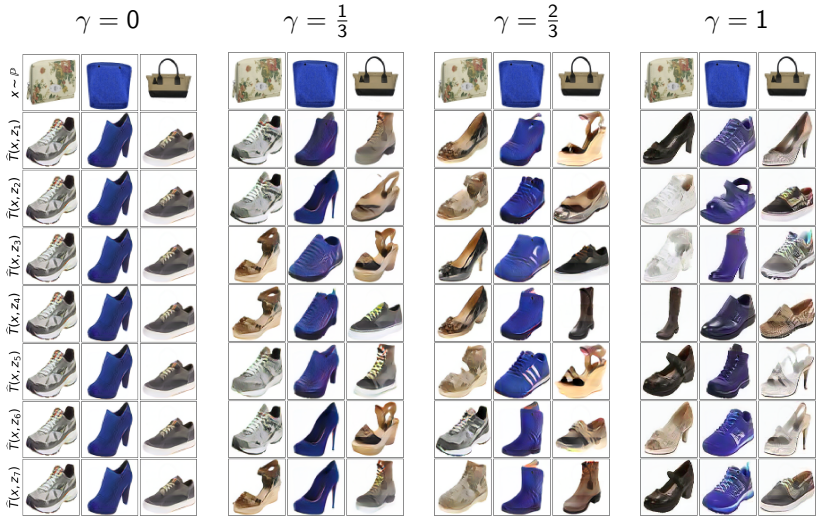
Transport cost: γ-weak quadratic cost ($\gamma \geq 0$).

Train-test split. 90% – train; the rest 10% – test.

Computational resources: 1 to 4 Tesla V100 GPUs,
convergence in 1-3 days (depending on the particular experiment).

Preliminary Experiments

HANDBAGS → SHOES, 64×64



Deterministic maps - Qualitative Results (Part 1)

HANDBAGS $\rightarrow$ SHOES, 128×128.

SHOES $\rightarrow$ HANDBAGS, 128×128.

Deterministic maps - Qualitative Results (Part 2)

CELEBA (FEMALE) $\rightarrow$ ANIME, 128×128.

OUTDOOR $\rightarrow$ CHURCH, 128×128.

Deterministic maps - Comparison with Other Methods

Method	DiscoGAN	Cycle GAN	NOT (ours)
FID↓	22.42	16.00	**13.77**

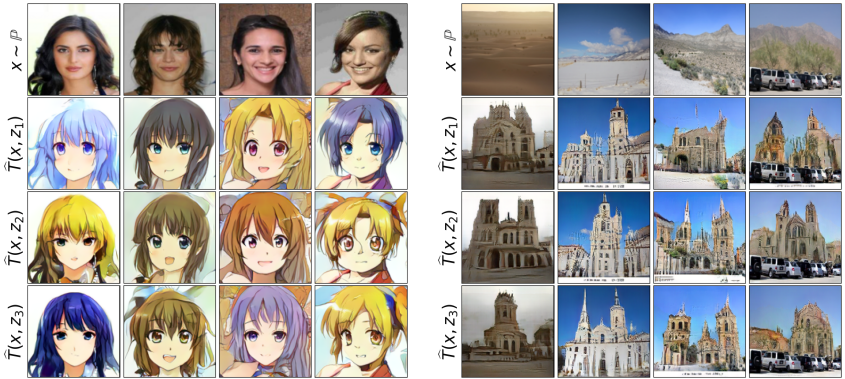
NOT (ours)

DiscoGAN

CycleGAN

Stochastic Maps - Qualitative Results

128× 128 images



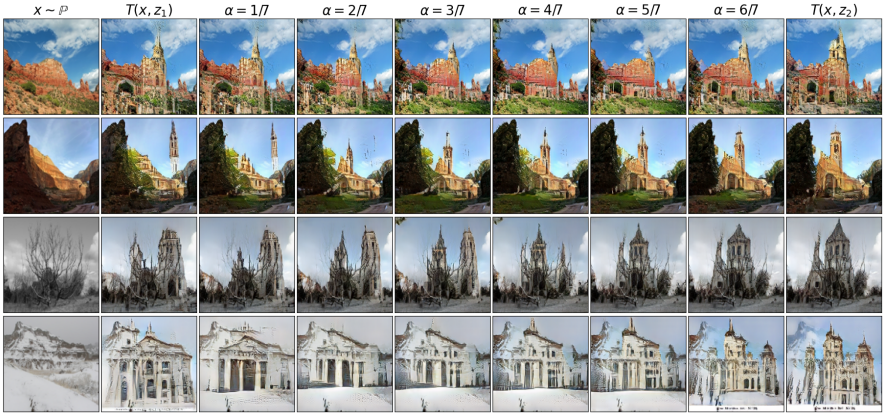
Celeba (female) → anime

Outdoor → church

Stochastic Maps - Qualitative Results

Outdoor → church, 128×128

Stochastic Maps - Qualitative Results



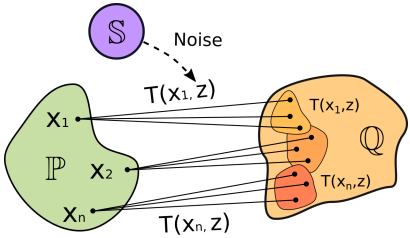
Outdoor → church (Interpolation)

Stochastic Maps – Comparison with Other Methods

Method	AugCycleGAN	MUNIT	NOT (ours)
FID↓	18.84 ± 0.11	15.76 ± 0.11	**13.44** ± 0.12

Conclusion and extensions

Conclusion



Neural Optimal Transport (NOT)

1. is a way to use learn OT plans as a generative mapping;
2. is built on rigorous OT theory;
3. solves a saddle point problem (not the GAN's one);
4. easily controls stochasticity (one-to-many mappings);
5. applicable to unpaired learning problems and beyond;

References i

- Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.
- Asadulaev, Arip et al. (2022). "Neural Optimal Transport with General Cost Functionals". In: *arXiv preprint arXiv:2205.15403*.
- Backhoff-Veraguas, Julio, Mathias Beiglböck, and Gudmun Pammer (2019). "Existence, duality, and cyclical monotonicity for weak transport costs". In: *Calculus of Variations and Partial Differential Equations* 58.6, pp. 1–28.
- Courty, Nicolas et al. (2016). "Optimal transport for domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.9, pp. 1853–1865.
- Gazdieva, Milena et al. (2022). "Unpaired Image Super-Resolution with Optimal Transport Maps". In: *arXiv preprint arXiv:2202.01116*.

References ii

- Gozlan, Nathael et al. (2017). "Kantorovich duality for general transport costs and applications". In: *Journal of Functional Analysis* 273.11, pp. 3327–3405.
- Kallenberg, Olav (1997). *Foundations of modern probability*. Vol. 2. Springer.
- Korotin, Alexander, Vage Egiazarian, et al. (2021). "Wasserstein-2 Generative Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=bEoxzW_EXsa.
- Korotin, Alexander, Alexander Kolesov, and Evgeny Burnaev (2022). "Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?" In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=VtEEpi-dGlt.
- Korotin, Alexander, Lingxiao Li, et al. (2021). "Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark". In: *Advances in Neural Information Processing Systems* 34.

References iii

- Korotin, Alexander, Daniil Selikhanovych, and Evgeny Burnaev (2022a). "Kernel Neural Optimal Transport". In: *arXiv preprint arXiv:2205.15269*.
- — (2022b). "Neural Optimal Transport". In: *arXiv preprint arXiv:2201.12220*.
- Liu, Huidong, Xianfeng Gu, and Dimitris Samaras (2019). "Wasserstein gan with quadratic transport cost". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4832–4841.
- Rout, Litu, Alexander Korotin, and Evgeny Burnaev (2022). "Generative Modeling with Optimal Transport Maps". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=5JdLZg346Lw.
- Villani, Cédric (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.