# Neural Optimal Transport (NOT)

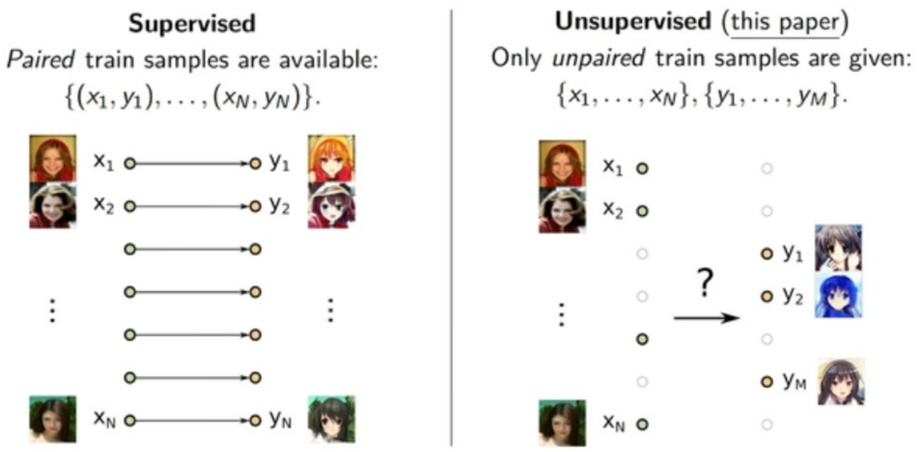
Alexander Korotin, Daniil Selikhanovych and Evgeny Burnaev

09/20/2024

Presented by Mengying Yan

# Motivation

- Domain translation
  - Unpaired (unsupervised) image-to-image translation




- Optimal transport
- Generative learning


(a) Celeba (female) $\rightarrow$ anime, outdoor $\rightarrow$ church, deterministic (one-to-one, $\mathbb{W}_2$).

# OT with neural networks

- OT cost as the loss to update generator in generative models
  - Only compute the OT cost
  - Example: Wasserstein GAN (Arjovsky et al., 2017)
- OT map/plan as the generative map
  - Most methods recover a non-stochastic (deterministic) plan --- which may not exist
  - Daniels et al. (2021) recover a stochastic plan, but is time consuming

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

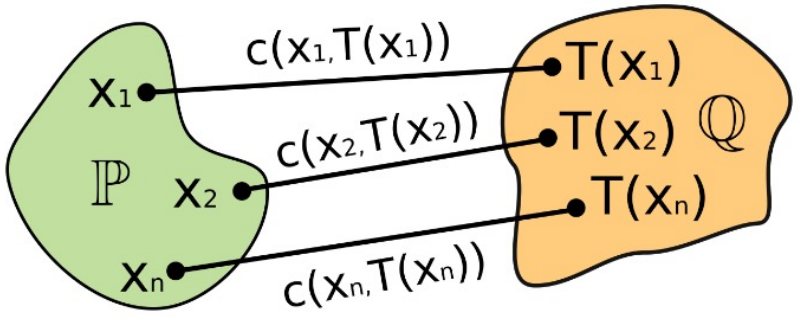
Grady Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport. Advances in Neural Information Processing Systems, 34, 2021.

# OT problem formulation

## Strong OT

## Monge's formulation

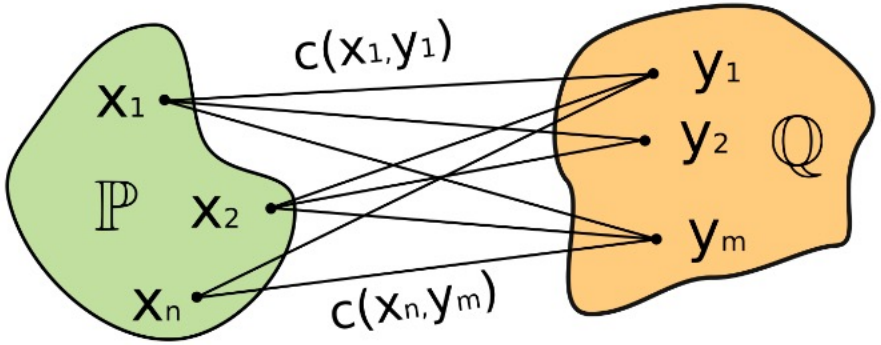
$$\mathrm{Cost}(\mathbb{P}, \mathbb{Q}) \stackrel{\mathrm{def}}{=} \inf_{T_\# \mathbb{P} = \mathbb{Q}} \int_{\mathcal{X}} c(x, T(x)) d\mathbb{P}(x)$$




## Kantorovitch's relaxation

$$\mathrm{Cost}(\mathbb{P}, \mathbb{Q}) \stackrel{\mathrm{def}}{=} \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

- Allow mass splitting
- It is Wasserstein-p distance when $c(x, y) = \|x - y\|^p$
- Minimizer $\pi^*$ is the OT plan
- Linear



# OT problem formulation

## Weak OT

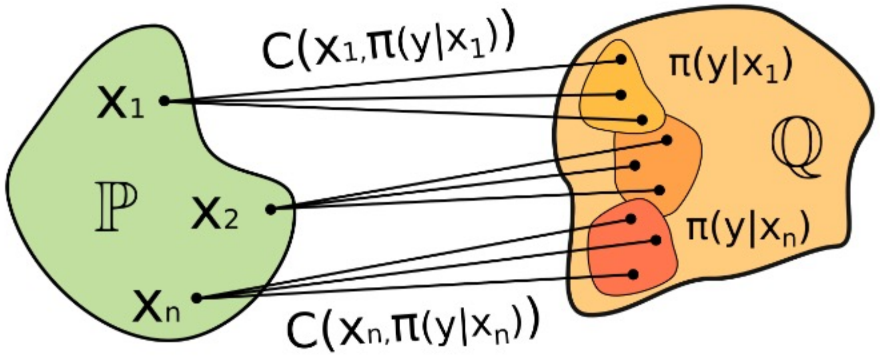
$$\mathrm{Cost}(\mathbb{P},\mathbb{Q}) \overset{\mathrm{def}}{=} \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}} C\big(x, \pi(\cdot|x)\big) d\pi(x)$$

- Mass splitting is allowed
- Transport cost is measured between a point and a distribution that is generated from this point
- Minimizer $\pi^*$ is called the OT plan
- Example of a weak OT cost ($\gamma$-weak quadratic cost):

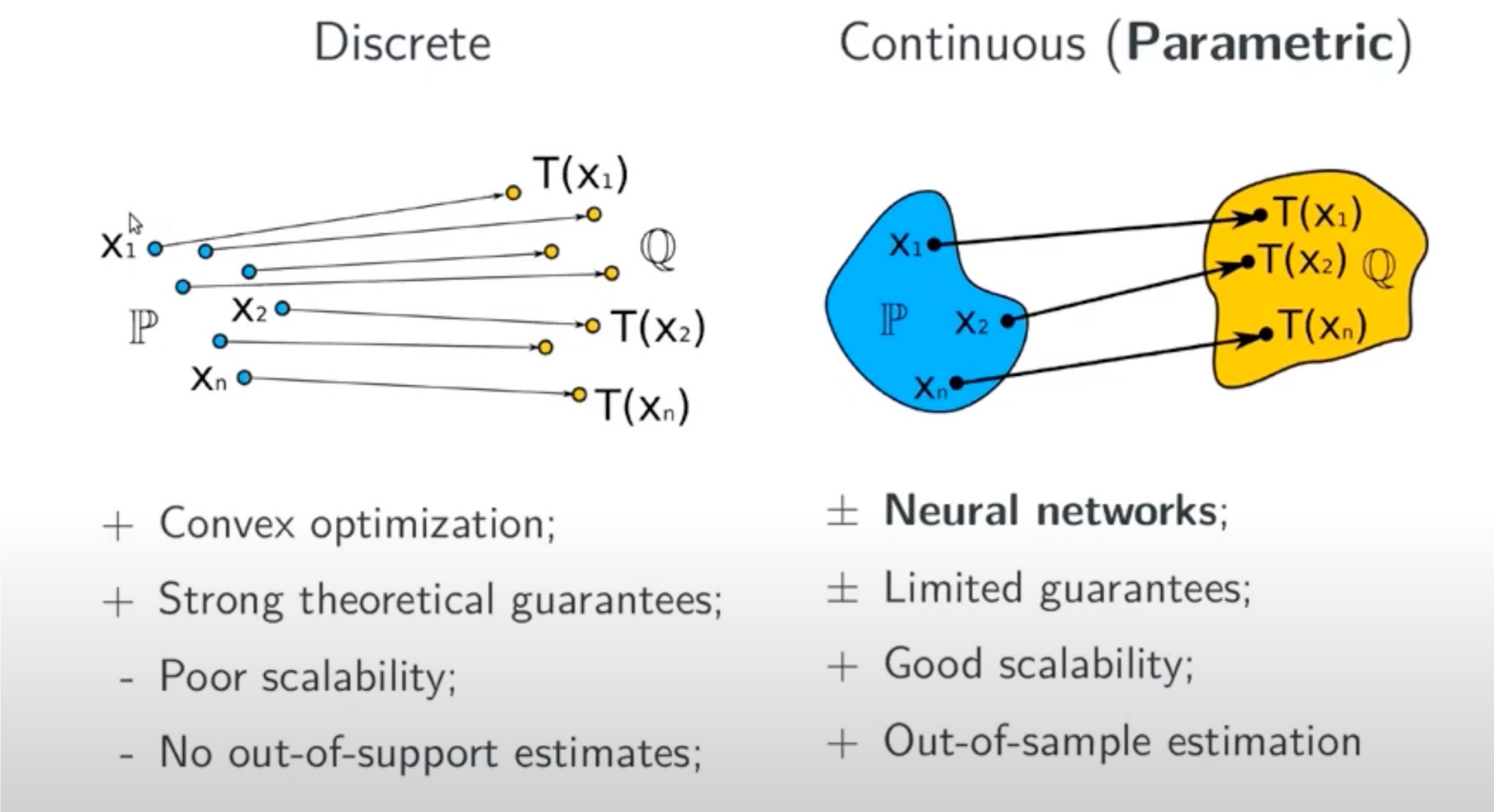
$$C(x,\mu) = \int_{\mathcal{Y}} \frac{1}{2}\|x-y\|^2 d\mu(y) - \frac{\gamma}{2}\mathrm{Var}(\mu)$$

Dissimilarity

Diversity (variance of generated distribution)



# Continuous optimal transport task

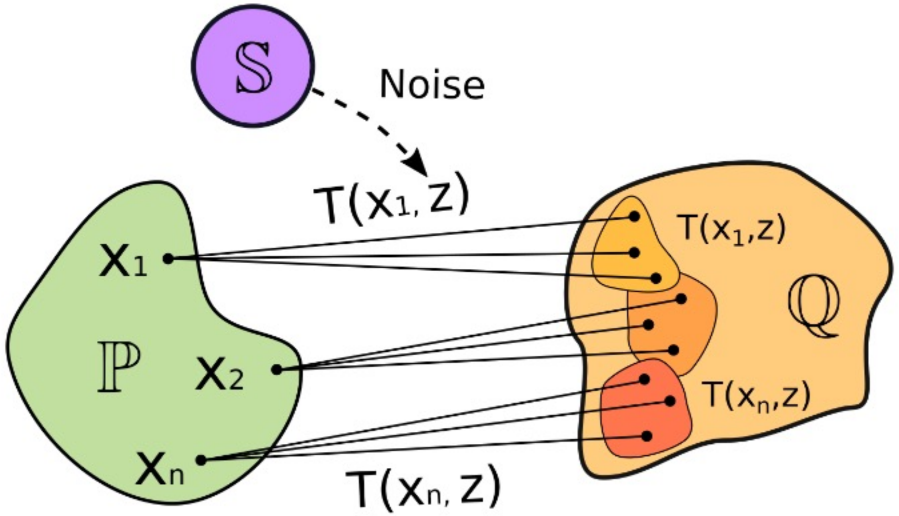



**This paper:**
Purpose a novel scalable algorithm to learn the deterministic and stochastic transport map for strong/weak costs with neural networks

Talk by Alexander Korotin at TII seminar (09 Aug 2022)

# Weak OT via stochastic functions

- $T: X \times Z \;\; \rightarrow Y$ is a stochastic function
- Z random noise
- If map T is independent of z, then the map is deterministic, o/w stochastic
- Stochastic functions can implicitly represent transport plans -- noise outsourcing



# Dual form of weak OT

## Primal form

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

Extract the primal solution $\pi^*$ (optimal plan) by from the dual problem

## Dual form

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

Potential function $f\colon \mathcal{Y} \to \mathbb{R}$

C-transform of f:

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

# Reformulation of the dual problem

1. Existence of transport maps (Lemma 1)
2. Reformulation of the C-transform (Lemma 2)
   - Replace the prob measure with the function that generates the prob measure

$$f^C(x) = \inf_t \left\{ C(x, t_\# \mathbb{S}) - \int_{\mathcal{Z}} f\big(t(z)\big) d\mathbb{S}(z) \right\}$$

3. Reformulate the integrated C-transform (Lemma 3)
   - Help represent the dual form as a saddle point (min-max)optimization problem

$$\int_{\mathcal{X}} f^C(x) d\mathbb{P}(x) = \inf_T \int_{\mathcal{X}} \left( C\big(x, T(x,\cdot)_\# \mathbb{S}\big) - \int_{\mathcal{Z}} f\big(T(x,z)\big) d\mathbb{S}(z) \right) d\mathbb{P}(x)$$

4. Maximin reformulation of the dual problem (Corollary 1)

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

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

# The key result

Stochastic OT maps solve the problem (Lemma 4)

- For any maximizer $f^*$and any stochastic map $T^*$ which realizes some optimal transport plan $\pi^*$, it holds that

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

- One may solve the saddle point problem and extract a stochastic OT map from its solution

# The algorithm

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

- We use ResNet[10] $f_{\omega} : \mathbb{R}^{3 \times W \times H} \to \mathbb{R}$;
- 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 $\mathbb{P}, \mathbb{Q}, \mathbb{S}$.

Talk by Alexander Korotin at TII seminar (09 Aug 2022)

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

**Input :** distributions $\mathbb{P}, \mathbb{Q}, \mathbb{S}$ accessible by samples; mapping network $T_\theta : \mathbb{R}^P \times \mathbb{R}^S \to \mathbb{R}^Q$; potential network $f_\omega : \mathbb{R}^Q \to \mathbb{R}$; number of inner iterations $K_T$; (weak) cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$; empirical estimator $\widehat{C}\big(x, T(x, Z)\big)$ for the cost;

**Output :** learned stochastic OT map $T_\theta$ representing an OT plan between distributions $\mathbb{P}, \mathbb{Q}$;

**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 $\omega$ by using $\frac{\partial \mathcal{L}_f}{\partial \omega}$;
- **for** $k_T = 1, 2, \dots, K_T$ **do**
  - Sample batch $X \sim \mathbb{P}$; 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 $\theta$ by using $\frac{\partial \mathcal{L}_T}{\partial \theta}$;

**until** *not converged*;

T: generator, f: discriminator. “NOT is NOT a WGAN”.

# Estimator for the $\gamma$-weak quadratic cost

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

Unbiased Monte-Carlo estimator

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

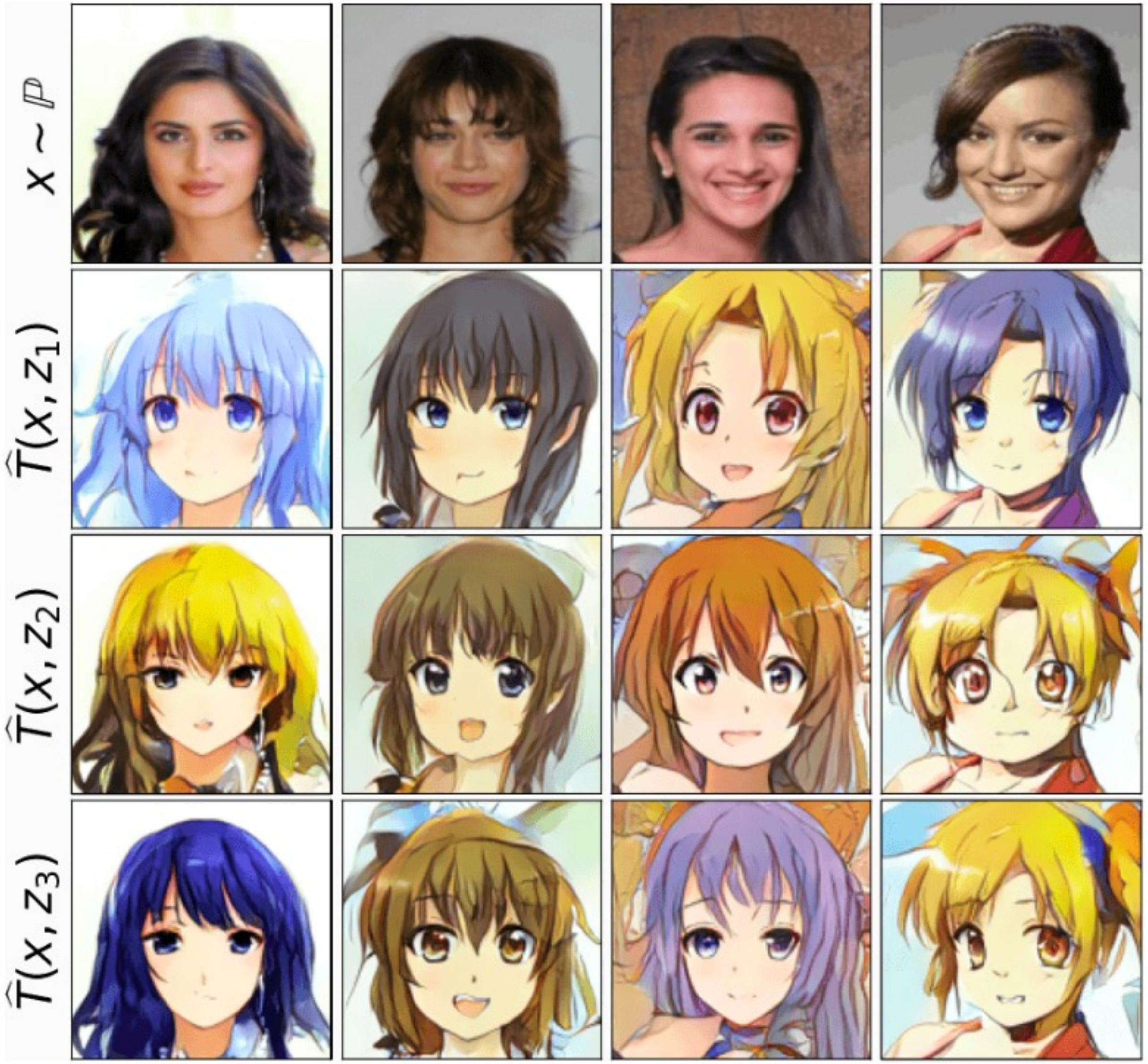
$\sigma^2$ is batch variance

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

# Results

One-to many translation with optimal plans

- $\gamma$-weak quadratic cost
- Stochastic




(a) Celeba (female) $\rightarrow$ anime, $128 \times 128$ ($\mathcal{W}_{2,\frac{2}{3}}$).


(a) $\gamma = 0$ (b) $\gamma = \frac{1}{3}$ (c) $\gamma = \frac{2}{3}$ (d) $\gamma = 1$

Figure 7: Stochastic *Handbags* $\rightarrow$ *shoes* translation with the $\gamma$-weak quadratic cost for various $\gamma$.

# Comparison – simpler model

| **Type** | **One-to-one** | | | **One-to-many** | | |
|---|---|---|---|---|---|---|
| **Method** | Disco GAN | Cycle GAN | NOT (ours) | AugCycle GAN | MUNIT | NOT (ours) |
| Hyperparameters of optimization objectives | None | Weights of cycle and identity losses $\lambda_{cyc}, \lambda_{id}$ | None | Weights of cycle losses $\gamma_1, \gamma_2$ | Weights of reconstruction losses $\lambda_x, \lambda_c, \lambda_s$ | Diversity control parameter $\gamma$ |
| Total number of hyperparameters | **0** | 2 | **0** | 2 | 3 | **1** |
| Networks | 2 generators, $2\times29.2$M<br><br>2 discriminators $2\times0.7$M | 2 generators $2\times11.4$M<br><br>2 discriminators $2\times2.8$M | 1 transport 9.7M,<br><br>1 potential 22.9M [32.4M$^*$] | 2 generators $2\times1.1$M,<br><br>2 discriminators $2\times2.8$M,<br><br>2 encoders $2\times1.4$M | 2 generators $2\times15.0$M,<br><br>2 discriminators $2\times8.3$M | 1 transport map 9.7M,<br><br>1 potential 22.9M [32.4M$^*$] |
| Total number of networks and parameters | 4 networks 59.8M | 4 networks 28.2M | **2 networks** 32.6M [42.1M$^*$] | 6 networks 7.0M | 4 networks 46.6M | **2 networks** 32.6M [42.1M$^*$] |

Table 2: Comparison of the number of hyperparameters of the optimization objectives, the number of networks and their parameters for the considered unpaired translation methods for $64\times64$ images.

# Comparison – smaller FID

FID (Fréchet inception distance): compares the distribution of generated images with the distribution of a set of real images

| **Type** | **One-to-one** | | | **One-to-many** | | |
|---|---|---|---|---|---|---|
| **Method** | Disco GAN | Cycle GAN | NOT (ours) | AugCycle GAN | MUNIT | NOT (ours) |
| Handbags $\rightarrow$ shoes ($64\times 64$) | 22.42 | 16.00 | **13.77** | 18.84 $\pm$ 0.11 | 15.76 $\pm$ 0.11 | **13.44** $\pm$ 0.12 |
| Celeba male $\rightarrow$ female ($64\times 64$) | 35.64 | 17.74 | **13.23** | 12.94 $\pm$0.08 | 17.07 $\pm$0.11 | **11.96** $\pm$0.07 |
| Outdoor $\rightarrow$ church ($128\times 128$) | 75.36 | 46.39 | **25.5** | 51.42 $\pm$0.12 | 31.42 $\pm$0.16 | **25.97** $\pm$0.14 |

# Comments

- This paper proposed a neural network based algorithm to solve stochastic transport plan
- GAN alternative
- It is worth reading and implementing 👍
- Finding the right cost may be the key

# If time allows…

Go through the practical example:

- https://github.com/iamalexkorotin/NeuralOptimalTransport/blob/main/seminars/NOT_seminar_weak_solutions.ipynb

# Resources:

GitHub repo:
https://github.com/iamalexkorotin/NeuralOptimalTransport

Short presentation:

https://iclr.cc/virtual/2023/oral/12644

Longer presentation

https://www.tii.ae/seminar/aidrc-seminar-series-alexander-korotin