Inverse Entropic Optimal Transport Solves Semi-supervised Learning via Data Likelihood Maximization

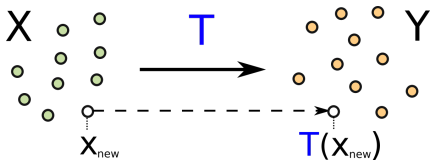
Mikhail Persiianov[1,2] **Arip Asadulaev**[3] **Nikita Andreev** **Nikita Starodubcev**[4]
Dmitry Baranchuk[4] **Anastasis Kratsios**[5,6] **Evgeny Burnaev**[1,7] **Alexander Korotin**[1,7]

[1]Applied AI Institute [2]Yandex School of Data Analysis [3]MBZUAI [4]Yandex Research
[5]Vector Institute [6]McMaster University [7]AXXX

Introduction to Semi-Supervised Domain Translation (SSDT)

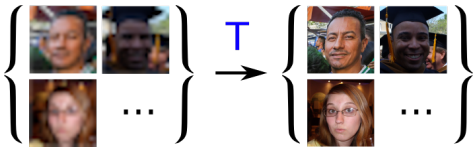
Motivation

The task: learn (from samples) a *translation* map between the two given data domains.

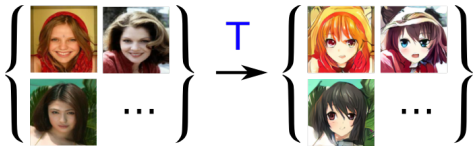


Important: the map should generalize to new data (similar to the train set).

Example 1: Image Super-Resolution

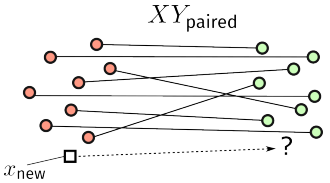


Example 2: Style Translation



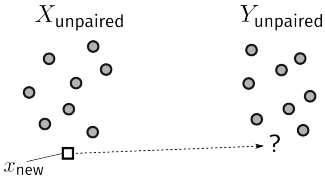
Problem Settings

Supervised



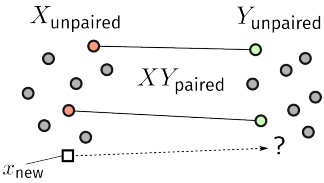
+ Relatively straightforward to solve.
– Paired data collection is expensive.

Unsupervised



+ Collecting unpaired data is straightforward.
– Ambiguity in solutions.

Semi-supervised (our focus)



+ Combines strengths of both sides.
– No principled learning algorithms.

Applications of Semi-Supervised Domain Translation

Figure 1: Day → Night translation for autonomous driving[3].

Figure 2: Sketch → Image translation for image editing[3].

Other examples:

- **Medical segmentation:** Few labeled MRI/CT scans + unlabeled target scans[1].
- **Remote sensing:** Land-cover or cloud detection with few labeled satellite images[2].
- **Cross-modality:** LiDAR → RGB / Night → Day for perception or surveillance[3].

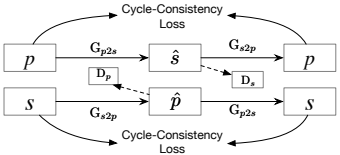
[1]Xiaofeng Liu et al. (2022). "Act: Semi-supervised domain-adaptive medical image segmentation with asymmetric co-training". In: *MICCAI*. Springer, pp. 66–76.

[2]Yadang Chen et al. (2022). "Semi-supervised contrastive learning for few-shot segmentation of remote sensing images". In: *Remote Sensing* 14.17, p. 4254.

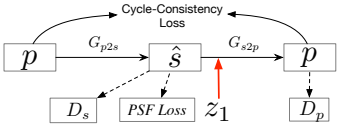
[3]Gaurav Parmar et al. (2024). "One-Step Image Translation with Text-to-Image Models". In: *arXiv e-prints*.

Limitations of Existing Approaches

- Overcomplicated heuristic loss term construction → many hyperparameters.
- No guarantees for convergence to ground-truth solution.



(a) Unpaired CycleGAN architecture



(b) Semi-Cycle-GAN architecture

Figure 3: Complicated objective[4].

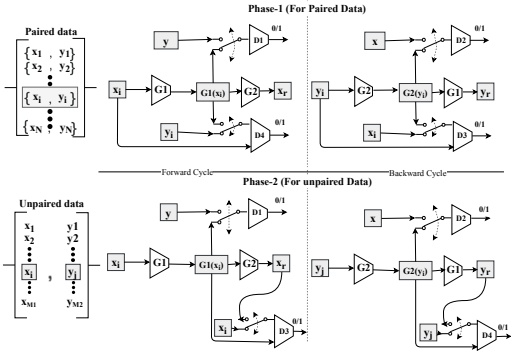


Figure 4: Complicated pipeline[5].

[4] Chaofeng Chen et al. (2023). "Semi-supervised Cycle-GAN for face photo-sketch translation in the wild". In: *Computer Vision and Image Understanding* 235, p. 103775

[5] Soumya Tripathy, Juho Kannala, and Esa Rahtu (2019). "Learning image-to-image translation using paired and unpaired training samples". In: *Computer Vision–ACCV 2018*. Springer, pp. 51–66

Our Proposed Energy-Based Approach for SSDT

Learning Setup for SSDT

Idea: Semi-supervised domain translation leverages both paired and unpaired data to improve performance and generalization. Paired data provides precise guidance, while unpaired data helps the model adapt to abundant or unlabeled samples.

Assumption: We have access to:

- Paired samples: $XY_{\text{paired}} \sim \pi^*$
- Additional unpaired samples: $X_{\text{unpaired}} \sim \pi_x^*$, $Y_{\text{unpaired}} \sim \pi_y^*$

Setup: The goal is to learn the true conditional mapping $\pi^*(\cdot \mid x)$ using all available data.

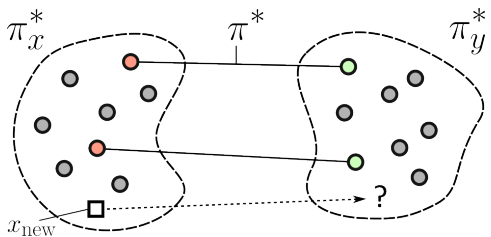


Figure 5: *Semi-supervised* domain translation setup (**our focus**).

Part I: Data Likelihood Maximization

Goal: Approximate the true distribution π^* by a parametric model π^θ via KL minimization:

$$\mathrm{KL}\left(\pi^* \| \pi^\theta\right) = \underbrace{\mathrm{KL}\left(\pi_x^* \| \pi_x^\theta\right)}_{\text{Marginal}} + \underbrace{\mathbb{E}_{x\sim\pi_x^*} \mathrm{KL}\left(\pi^*(\cdot|x) \| \pi^\theta(\cdot|x)\right)}_{\text{Conditional}}. \tag{1}$$

The conditional KL is crucial for domain translation since we often sample from marginal π_x^* and then sample from $\pi^\theta(\cdot|x)$. It can be rewritten as:

$$\mathbb{E}_{x\sim\pi_x^*}\mathbb{E}_{y\sim\pi^*(\cdot|x)}\left[\log \pi^*(y|x) - \log \pi^\theta(y|x)\right] = \underbrace{-\mathbb{E}_{x\sim\pi_x^*}\mathrm{H}\left(\pi^*(\cdot|x)\right)}_{=\text{const}} - \mathbb{E}_{x,y\sim\pi^*} \log \pi^\theta(y|x).$$

where the first term is independent of θ.

Resulting minimization objective equivalent to maximizing the conditional likelihood[6]:

$$\mathcal{L}(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_{x,y\sim\pi^*} \log \pi^\theta(y|x) \to \min_\theta. \tag{2}$$

Limitation: Requires fully paired data; unpaired samples cannot be used.

In this paper, we **remove** this restriction by leveraging both paired and unpaired data.

[6] George Papamakarios et al. (2021). "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57, pp. 1–64.

Part II: Exploiting Unpaired Data via Energy-Based Parameterization

To address the above-mentioned issue and utilize unpaired data, we first use **Gibbs-Boltzmann parametrization**[7]:

$$\pi^\theta(y|x) \stackrel{\text{def}}{=} \frac{\exp\left(-E^\theta(y|x)\right)}{Z^\theta(x)}, \; Z^\theta(x) \stackrel{\text{def}}{=} \int_{\mathcal{Y}} \exp\left(-E^\theta(y|x)\right) \mathrm{d}y, \tag{3}$$

where $E^\theta(\cdot|x) : \mathcal{Y} \to \mathbb{R}$ is *the Energy function*, and $Z^\theta(x)$ is the *normalization constant*. Substituting (3) into (2), we obtain:

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}_{x,y\sim\pi^*} E^\theta(y|x)}_{\text{Conditional, requires pairs } (x,y)\sim\pi^*} + \underbrace{\mathbb{E}_{x\sim\pi_x^*} \log Z^\theta(x)}_{\text{Marginal, requires } x\sim\pi_x^*} \to \min_\theta. \tag{4}$$

Observation: This objective already provides an opportunity to exploit the unpaired samples from the marginal distribution π_x^* to learn the conditional distributions $\pi^\theta(\cdot|x) \approx \pi^*(\cdot|x)$.

[7] Yann LeCun et al. (2006). "A tutorial on energy-based learning". In: *Predicting structured data* 1.0.

Decoupling Cost and Potential in Energy Function

Goal: Incorporate independent samples $y \sim \pi_y^*$ into (4) to enable semi-supervised learning.

We propose, the following **energy function parameterization** for $\varepsilon > 0$:

$$E^\theta(y|x) \stackrel{\text{def}}{=} \frac{c^\theta(x,y) - f^\theta(y)}{\varepsilon}, \tag{5}$$

which decouples contributions from paired samples $(x, y) \sim \pi^*$ and unpaired samples $y \sim \pi_y^*$.

Substituting (5) into (4) and using $\mathbb{E}_{x,y\sim\pi^*} f^\theta(y) = \mathbb{E}_{y\sim\pi_y^*} f^\theta(y)$, we obtain our **final objective**:

$$\mathcal{L}(\theta) = \underbrace{\varepsilon^{-1}\mathbb{E}_{x,y\sim\pi^*}[c^\theta(x,y)]}_{\text{Joint, requires pairs } (x,y)\sim\pi^*} - \underbrace{\varepsilon^{-1}\mathbb{E}_{y\sim\pi_y^*} f^\theta(y)}_{\text{Marginal, requires } y\sim\pi_y^*} + \underbrace{\mathbb{E}_{x\sim\pi_x^*} \log Z^\theta(x)}_{\text{Marginal, requires } x\sim\pi_x^*} \to \min_\theta . \tag{6}$$

Question: How can we optimize this objective efficiently, given the typically intractable $Z^\theta(x)$?

Proposed Practical Parametrization: EBiEOT-GMM

Cost function and dual potential parameterization:

$$c^\theta(x,y) = -\varepsilon \log \sum_{m=1}^{M} v_m^\theta(x) \exp\left(\frac{\langle a_m^\theta(x), y\rangle}{\varepsilon}\right), \tag{7}$$

$$f^\theta(y) = \varepsilon \log \sum_{n=1}^{N} w_n^\theta \mathcal{N}(y \mid b_n^\theta, \varepsilon B_n^\theta), \tag{8}$$

where $v_m^\theta(x) : \mathbb{R}^{D_x} \to \mathbb{R}_+$ and $a_m^\theta(x) : \mathbb{R}^{D_x} \to \mathbb{R}^{D_y}$ are neural networks with learnable parameters θ_c and $\theta_f = \{w_n^\theta, b_n^\theta, B_n^\theta\}_{n=1}^N$ are learnable parameters.

Proposition (Tractable normalization constant)

Our parameterization of the cost c^θ and dual potential f^θ yields a closed-form $Z^\theta(x)$, enabling efficient end-to-end SGD[8] optimization of the loss (6)*, given by:*

$$Z^\theta(x) \overset{\text{def}}{=} \sum_{m=1}^{M}\sum_{n=1}^{N} z_{mn}(x), \text{ where } \quad z_{mn}(x) \overset{\text{def}}{=} w_n v_m(x) \exp\left(\frac{a_m(x)^\top B_n a_m(x) + 2b_n^\top a_m(x)}{2\varepsilon}\right).$$

[8] Diederik P Kingma (2014). "Adam: A method for stochastic optimization". In: *ICLR, 2014.*

Training and Inference

Training: Minimize the empirical counterpart of $\mathcal{L}(\theta)$ using SGD:

$$\mathcal{L}(\theta) \approx \widehat{\mathcal{L}}(\theta) \stackrel{\text{def}}{=} \varepsilon^{-1} \frac{1}{P} \sum_{p=1}^{P} c^{\theta}(x_p, y_p) - \varepsilon^{-1} \frac{1}{R} \sum_{r=1}^{R} f^{\theta}(y_r) + \frac{1}{Q} \sum_{q=1}^{Q} \log Z^{\theta}(x_q).$$

Proposition (Tractable conditional distributions)

From our parametrization of the cost function (7) *and dual potential* (8) *it follows that the* $\pi^{\theta}(\cdot|x)$ *are Gaussian mixtures:*

$$\pi^{\theta}(y|x) = \frac{1}{Z^{\theta}(x)} \sum_{m=1}^{M} \sum_{n=1}^{N} z_{mn}(x) \mathcal{N}(y \mid d_{mn}(x), \varepsilon B_n), \tag{9}$$

where $d_{mn}(x) \stackrel{\text{def}}{=} b_n + B_n a_m(x)$ *and* $z_{mn}(x)$ *is from Proposition 1.*

Inference: Conditional distributions $\pi^{\theta}(\cdot|x)$ are Gaussian mixtures (9), enabling efficient sampling of y given x.

2D Translation: Gaussian → Swiss Roll. P=128, Q=R=1024

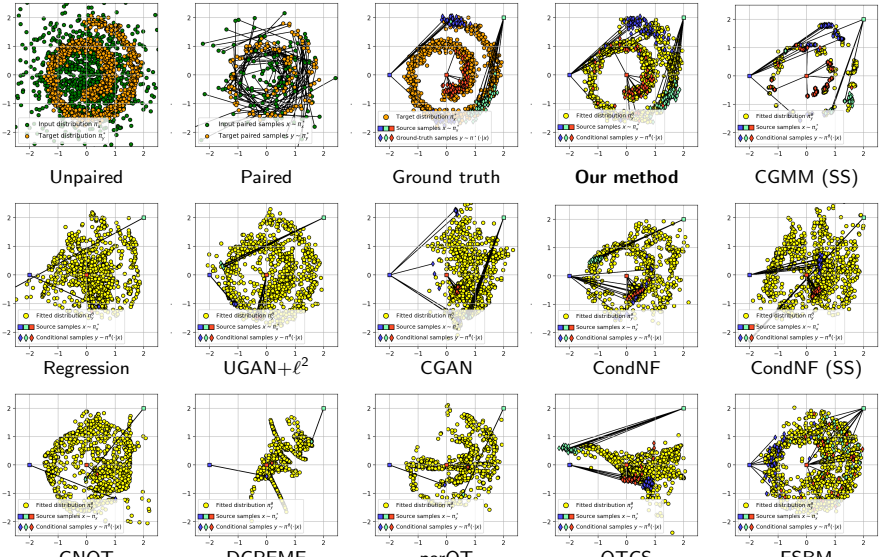


Image Translation via ALAE[9](setup from FSBM paper[10])

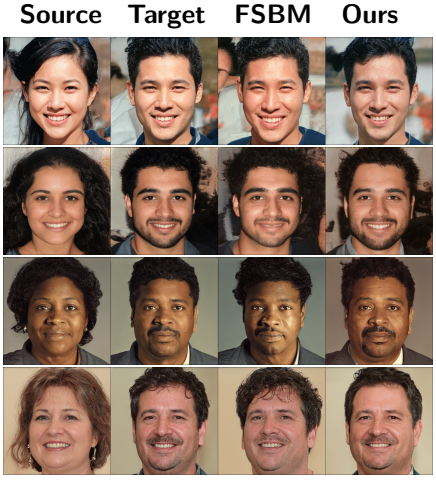


Figure 6: *Woman-to-Man* translation task.

Figure 7: *Old-to-Young* translation task.

[9] Panagiotis Theodoropoulos et al. (2025). "Feedback Schrödinger Bridge Matching". In: *ICLR, 2025.*

[10] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto (2020). "Adversarial latent autoencoders". In: *CVPR, 2020.*

Conclusion and Extensions

Discussion and Conclusion

Contribution:

- First ever simple, non-minimax objective integrating paired and unpaired data.
- Supports two parameterizations for continuous $y \in \mathbb{R}^{D_y}$: GMM and neural networks.

Future Directions:

- Extension to discrete targets $y \in \mathbb{K}^{D_y}$ is a promising direction.
- Established connection to iEOT may foster development of advanced semi-supervised methods based on **EBMs**[11], **diffusion schrödinger bridges**, **flow matching**.

Limitations:

- Relies on GMM parameterization, which affects scalability.
- Fully neural parameterizations for cost and potential relies on MCMC sampling.

Thank You!

Source code available at:

`https://github.com/MuXauJl11110/EBiEOT`

[11] Yaxuan Zhu et al. (2024). "Learning Energy-Based Models by Cooperative Diffusion Recovery Likelihood". In: *ICLR, 2024*.