



# Sparse Domain Transfer via Elastic Net Regularization

Jingwei Zhang<sup>1</sup>  and Farzan Farnia<sup>1</sup> 

Department of Computer Science and Engineering, The Chinese University of Hong Kong  
{jwzhang22, farnia}@cse.cuhk.edu.hk

**Abstract.** Transportation of samples across different domains is a central task in several machine learning problems. A sensible requirement for domain transfer tasks in computer vision and language domains is the sparsity of the transportation map, i.e., the transfer algorithm aims to modify the least number of input features while transporting samples across the source and target domains. In this work, we propose *Elastic Net Optimal Transport (ENOT)* to address the sparse distribution transfer problem. The ENOT framework utilizes the  $L_1$ -norm and  $L_2$ -norm regularization mechanisms to find a sparse and stable transportation map between the source and target domains. To compute the ENOT transport map, we consider the dual formulation of the ENOT optimization task and prove that the sparsified gradient of the optimal potential function in the ENOT’s dual representation provides the ENOT transport map. Furthermore, we demonstrate the application of the ENOT framework to perform feature selection for sparse domain transfer. We present the numerical results of applying ENOT to several domain transfer problems for synthetic Gaussian mixtures and real image and text data. Our empirical results indicate the success of the ENOT framework in identifying a sparse domain transport map. Code is available at [github.com/buyeah1109/ENOT](https://github.com/buyeah1109/ENOT).

**Keywords:** Generative model · Domain transportation · Feature selection

## 1 Introduction

Deep neural networks (DNNs) have revolutionized the performance of computer vision models in domain transfer applications where the features of an input sample are altered to transfer the sample to a secondary domain [14, 26, 39]. The common goal of domain transfer algorithms is to transport an input data point to a target distribution by applying *minimal changes* to the input. Over recent years, domain transportation algorithms based on generative adversarial networks (GANs) including CycleGAN [38] and StyleGAN [16] have achieved empirical success in addressing the domain transfer task for image distributions. The success of these algorithms has inspired several studies of GAN-based domain transfer methodologies [6, 27, 37].

While the GAN-based methods have led to successful results in image-based domain transfer problems, their application demands significantly higher computational costs than standard GAN algorithms including only one generator/discriminator neural net pair to transfer a latent Gaussian vector to the data distribution. The extra computations in these domain transfer algorithms aim to ensure an invertible transfer map and thus limited modifications to an input sample. For example, the CycleGAN algorithm considers two pairs of generator/discriminator neural nets to impose a reversible transformation of an input image. However, the additional pair of neural nets in the CycleGAN setting will lead to a more challenging optimization task and higher training costs.

In this work, we focus on *sparse domain transfer problems* where the transfer of samples between source and target domains can be achieved by editing only a limited subset of input features. We note that the assumption of a sparse transport map applies to several real-world domain transfer problems, e.g. object translation, text revision, and gene editing problems. In the mentioned tasks, the sparsity level of the transfer map results in a meaningful measure of changes applied to an input sample. While sparse transportation maps are desired in many real-world domain transfer problems, the commonly-used GAN-based algorithms often lead to dense transfer maps editing a considerable fraction of input features.

To address sparse domain transfer tasks, we propose an optimal transport-based approach which takes advantage of the induced sparsity of the  $L_1$ -norm regularization and the stability properties of the  $L_2$ -norm regularization. Our proposed framework, which we call *Elastic Net Optimal Transport (ENOT)*, solves an optimal transport problem where the transportation cost follows from the elastic net function [40] combining standard Euclidean-norm-squared and  $L_1$ -norm cost functions. Therefore, the ENOT approach can be interpreted as a mechanism to regularize the standard optimal transport map toward sparser transportation functions. By tuning the coefficient of the  $L_1$ -norm regularization in ENOT's elastic net cost, the learner can adjust the sparsity level of the transportation map and explore the spectrum between the standard and fully  $L_1$ -norm-based optimal transport tasks.

To analyze the ENOT problem, we leverage optimal transport theory [33] and extend the duality results to the ENOT optimal transport setting. We prove a generalization of standard Brenier's theorem, highlighting the connection between the optimal potential function in the ENOT's dual problem with the optimal transport map transferring samples across domains. Our main theorem suggests that the composition of a soft-thresholding function with the gradient of the optimal potential function will perform sparse transportation across the domains. This result indicates that the ENOT framework offers a combination of the standard optimal transport problem with the squared-error cost function and the  $L_1$ -norm-based optimal transport problem which leads to a challenging optimization problem without the  $L_2$ -norm-based regularization in ENOT.

Furthermore, we utilize the ENOT framework to develop a *feature selection-based approach* to reduce the sparse domain transport task to a constraint-free

distribution transfer problem where an unconstrained transfer map is applied to only the selected feature subset. According to this variable selection-based approach, we break the domain transfer problem into two sub-problems: 1) ENOT-based variable selection choosing features to undergo modification for a given input sample, 2) Applying an unconstrained transportation map via standard GAN frameworks that transfers the input sample masked by the feature selection output to the target domain. By tuning the coefficient of the  $L_1$ -norm regularization in the ENOT’s elastic net cost, the learner can adjust the number of selected features prior to performing a constraint-free GAN-based distribution transfer.

Finally, we discuss the numerical results of the applications of the ENOT framework to sparse domain transfer problems from various areas including computer vision, computational biology, and natural language processing. Our empirical findings show that the feature selection-based domain transfer via ENOT can be easily adapted to different domains and achieves satisfactory results. We qualitatively evaluate ENOT’s application for feature identification in sparse domain transfer. The numerical results support the proposed methodology of sparse domain transfer via ENOT-based domain transportation and feature selection. The contributions of this work can be summarized as:

- Proposing a feature selection-based approach for the sparse domain transfer problem,
- Developing ENOT as an elastic net-based methodology to the sparse domain transfer and variable selection,
- Extending the theory of standard squared-error-based optimal transport task to the ENOT setting,
- Providing supportive numerical results for applying ENOT-based sparse domain transfer to various domain transfer tasks.

## 2 Related Work

**Sparsity and Optimal Transport Methods.** Several related works have studied various notions of sparsity in optimal transport frameworks. References [2, 4, 10, 22, 31] propose sparsity-based regularization of the transportation matrix in optimal transport problems. However, we note that the sparsity objective pursued in these works differs from the sparse domain transfer in our work: while the mentioned papers aim for a sparse transportation matrix to gain a sparse alignment of source and target samples, our proposed ENOT method focuses on the sparsity of the modified input features in the domain transfer. Meanwhile, [18] proposed neural optimal transport and leverage neural networks to model potential functions and conduct image transportation, and considered the standard Euclidean-norm-squared cost function for the transportation, which does not focus on sparse transportation.

[7] first introduces an optimal transport-based approach to the sparse domain transfer problem, which aim to precisely solve an entropic-regularized optimal transport problem over the empirical samples and then use a kernel-based

interpolation to generalize the solution to unseen data. The main difference between [7]’s analysis and ours is the focus on the *primal vs. dual* formulations of the sparse optimal transport problem. The analysis in [7] concentrates on primal formulation and finds the precise solution to the primal OT problem on the training data. On the other hand, our approach targets the dual formulation and is more similar to the Wasserstein GAN framework, which involves a potential neural net function and can be extended to large-scale image and text data. Overall, our neural net-based method for the ENOT optimal transport problem can be viewed as a complementary approach to the precise kernel-based framework in [7]. Also, the trained neural net can be used as an efficient-to-compute feature selection map, which we later used to introduce a *feature selection-based approach* to sparse domain transfer, a topic that has not been studied in [7], which is useful for large-scale image and text-related applications.

**Unsupervised Image to Image Translation (UI2I).** Several related works attempt to address image-based transportation problems. For the image style transfer task, CycleGAN [38] uses a cycle-consistent loss and two GANs to conduct cyclic unpaired transformation. DRIT++ [20] adopts encoders to obtain the latent representation of images and similar cross-cycle consistency loss. For the image colorization transfer task, Conditional GANs are leveraged to improve colorization performance [13]. [36] propose a real-time user-guided neural network colorization. Moreover, cyclic-loss [3, 29, 34, 38] and GANs [6, 27, 37] have been utilized to address UI2I. However, unlike ENOT, these related works do not focus on the sparsity of transfer maps.

**Sequence to Sequence Translation.** Sequence to sequence (Seq2Seq) neural net models are typically designed based on an encoder-decoder architecture. [15] propose the application of a convolutional neural network (CNN) as the encoder and a recurrent neural network (RNN) as the decoder. [30] utilize an RNN-based architecture for both the encoder and decoder neural nets. [32] propose a transformer based on multi-head self-attention. BART [21] offers a sequence-to-sequence pretraining solution and adopts a bidirectional encoder similar to BERT [8], and a decoder similar to GPT [28]. Unlike our proposed ENOT approach, the discussed methods usually result in a dense transportation map. We also note that Seq2Seq and GAN-based transfer methods are almost exclusively used for language and image distributions, respectively.

### 3 Preliminaries

Consider random vectors  $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$  with probability distributions  $P_X, P_Y$ , respectively. Given  $n$  independent samples  $\mathbf{x}_1, \dots, \mathbf{x}_n$  from  $P_X$  and  $m$  independent samples  $\mathbf{y}_1, \dots, \mathbf{y}_m$  from  $P_Y$ , the goal in the domain transfer problem is to learn a map  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^d$  transporting an input  $\mathbf{X}$  from distribution  $P_X$  to an output  $\psi(\mathbf{X})$  distributed as  $P_Y$ , i.e.,

$$\psi(\mathbf{X}) \stackrel{\text{dist}}{=} \mathbf{Y}.$$

Here,  $\stackrel{\text{dist}}{=}$  denotes identical probability distributions.

Without any constraint on the map  $\psi$ , there exist infinitely many transportation maps resulting in the required identical distributions. To uniquely characterize the transfer map, the optimal transport framework [33] seeks to find a map minimizing the expected transportation cost measured based on a cost function  $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ . According to this framework, the transportation map follows from the optimal coupling  $\Pi_{X,Y}$ , marginally distributed as  $P_X$  and  $P_Y$ , that is minimizing the expected transportation cost formulated as

$$\text{OT}_c(P_X, P_Y) := \inf_{\substack{\Pi_{X,Y} : \Pi_X = P_X \\ \Pi_Y = P_Y}} \mathbb{E}_{(X,Y) \sim \Pi} [c(\mathbf{X}, \mathbf{Y})].$$

Here,  $\text{OT}_c(P_X, P_Y)$  denotes the optimal transport cost between  $P_X, P_Y$ . It is well-known that under mild regularity conditions, a deterministic coupling mapping  $\mathbf{X}$  to a sample with distribution  $P_Y$  exists that solves the above problem. Also, the dual representation of the above optimization problem can be formulated via the Kantorovich duality [33] as

$$\sup_{\phi: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbb{E}[\phi(\mathbf{X})] - \mathbb{E}[\phi^c(\mathbf{Y})],$$

where  $\phi$  is the potential function and the  $c$ -transform  $\phi^c$  is defined as  $\phi^c(\mathbf{y}) := \sup_{\mathbf{y}' } \phi(\mathbf{y}') - c(\mathbf{y}, \mathbf{y}')$ .

*Example 1.* In the special case of a norm cost  $c_1(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ , the result of Kantorovich duality can be written as

$$\text{OT}_{c_1}(P_X, P_Y) = \sup_{\phi: 1\text{-Lipschitz}} \mathbb{E}[\phi(\mathbf{X})] - \mathbb{E}[\phi(\mathbf{Y})]$$

where the potential function  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$  is constrained to be 1-Lipschitz with respect to the assigned norm  $\|\cdot\|$ , i.e., for every  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ :

$$|\phi(\mathbf{x}) - \phi(\mathbf{x}')| \leq \|\mathbf{x} - \mathbf{x}'\|.$$

*Example 2.* In the special case of the  $L_2$ -norm-squared cost  $c_2(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ , the result of Kantorovich duality can be written as

$$\sup_{\tilde{\phi}: \text{convex}} \mathbb{E} \left[ \frac{1}{2} \|\mathbf{X}\|_2^2 - \tilde{\phi}(\mathbf{X}) \right] + \mathbb{E} \left[ \frac{1}{2} \|\mathbf{Y}\|_2^2 - \tilde{\phi}^*(\mathbf{Y}) \right] \quad (1)$$

where the potential function  $\phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}\|_2^2 - \tilde{\phi}(\mathbf{x})$  is constrained to be the subtraction of a convex function  $\tilde{\phi}$  from  $\frac{1}{2}\|\mathbf{x}\|_2^2$ , and  $\tilde{\phi}^*$  is the Fenchel conjugate defined as

$$\tilde{\phi}^*(\mathbf{x}) := \sup_{\mathbf{x}'} \mathbf{x}'^\top \mathbf{x} - \tilde{\phi}(\mathbf{x}').$$

The Brenier theorem reveals that in the setting of Example 2, the gradient of the optimal solution  $\tilde{\phi}$  provides the unique monotone (gradient of a convex function) map transporting samples between the two domains:

**Theorem 1 (Brenier’s Theorem, [33]).** *Suppose that  $P_X, P_Y$  are absolutely continuous with respect to one another. Then, the gradient of the solution  $\tilde{\phi}^*$  to equation 1 is the unique monotone map for transferring  $P_X$  to  $P_Y$ , that is*

$$\nabla \tilde{\phi}^*(\mathbf{X}) \stackrel{\text{dist}}{=} \mathbf{Y}.$$

In the following sections, we aim to define and analyze optimal transport costs that can capture the sparsity of the transportation map, i.e. the number of non-zero coordinates of  $\mathbf{y} - \mathbf{x}$ .

## 4 Elastic Net Regularization for Sparse Optimal Transport

In this work, we aim to address the sparse domain transfer problem where the transfer map  $\psi$  between distributions  $P_X, P_Y$  alters the fewest possible coordinates in the  $d$ -dimensional feature vector  $\mathbf{X} = [X^{(1)}, \dots, X^{(d)}]$ . To apply the optimal transport framework, a proper cost function is the cardinality (number of non-zero elements  $\text{card}(\mathbf{z}) = \sum_{i=1}^d \mathbf{1}[z_i \neq 0]$ ) of the difference between the original and transported samples:

$$c_{\text{sparse}}(\mathbf{x}, \mathbf{y}) = \text{card}(\mathbf{x} - \mathbf{y}).$$

Since the cardinality function lacks continuity and convexity, the resulting optimal transport problem will be computationally difficult. A common convex proxy for the cardinality function is the  $L_1$ -norm where we simply use  $c_{L_1}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1$ . While the primal optimal transport problem could be solved for the empirical samples with the  $L_1$ -norm cost, the domain transfer map requires solving the optimization problem for the data distribution which would be complex in the primal case. Therefore, we focus on the dual optimization problem to the optimal transport task. However, solving the dual optimization problem of the  $L_1$ -norm cost requires optimizing over the  $L_1$ -norm-based 1-Lipschitz functions which would be challenging.

To handle the computational complexity of the dual optimization problem with  $L_1$ -norm cost function, we propose to apply the *elastic net* [40] cost function with coefficients  $0 \leq \alpha \leq 1$  and  $\lambda > 0$ :

$$c_{\text{EN}}^{\alpha, \lambda}(\mathbf{x}, \mathbf{y}) = \lambda(1 - \alpha)\|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\alpha\|\mathbf{x} - \mathbf{y}\|_1. \quad (2)$$

Using the above cost function, we propose the *Elastic Net-based Optimal Transport (ENOT)* as the optimal transport method formulated with the cost function in equation 2. For the dual formulation of the ENOT problem, we can apply the Kantorovich duality to obtain the following optimization task:

$$\max_{\phi: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbb{E}[\phi(\mathbf{X})] - \mathbb{E}[\phi_{\text{EN}}^{\alpha, \lambda}(\mathbf{Y})] \quad (3)$$

where the elastic-net-based  $c$ -transform can be written as follows:

$$\phi_{\text{EN}}^{\alpha, \lambda}(\mathbf{y}) := \max_{\boldsymbol{\delta} \in \mathbb{R}^d} \phi(\mathbf{y} + \boldsymbol{\delta}) - \lambda(1 - \alpha)\|\boldsymbol{\delta}\|_2^2 - \lambda\alpha\|\boldsymbol{\delta}\|_1.$$

**Theorem 2.** Consider the ENOT dual problem in equation 3. Then, there exists an optimal potential function  $\phi^*$  for this problem which satisfies the following weakly-concavity property: for every  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  and real value  $\gamma \in [0, 1]$ :

$$\begin{aligned} \phi^*(\gamma\mathbf{x} + (1 - \gamma)\mathbf{y}) &\geq \gamma\phi^*(\mathbf{x}) + (1 - \gamma)\phi^*(\mathbf{y}) \\ &\quad - \lambda\gamma(1 - \gamma)(1 - \alpha)\|\mathbf{x} - \mathbf{y}\|_2^2 - \lambda\alpha\|\mathbf{x} - \mathbf{y}\|_1. \end{aligned}$$

*Proof.* We defer the proof to the Appendix.

The above result shows the existence of an optimal potential function possessing a weakly-concave structure defined based on an elastic net function. Our next result reveals the extension of the Brenier’s theorem to the elastic net cost function. In this extension, we use  $\text{ST}_\gamma$  to denote the soft-thresholding operator defined for a scalar input as

$$\text{ST}_\gamma(z) := \begin{cases} z + \gamma & \text{if } z \leq -\gamma \\ 0 & \text{if } -\gamma < z < \gamma \\ z - \gamma & \text{if } \gamma \leq z. \end{cases}$$

For a vector input  $\mathbf{z} \in \mathbb{R}^d$ , we define the soft-thresholding map as the coordinate-wise application of the scalar soft-thresholding function, i.e.,

$$\forall i \in \{1, \dots, d\} : \quad \text{ST}_\gamma(\mathbf{z})_i = \text{ST}_\gamma(z_i)$$

**Theorem 3.** Consider the dual ENOT problem in equation 3. Then, given the optimal potential function  $\phi^*$  the following will provide the optimal transport map transferring samples across domains:

$$\mathbf{X} - \text{ST}_{\frac{\alpha}{2(1-\alpha)}}\left(\frac{1}{2\lambda(1-\alpha)}\nabla\phi^*(\mathbf{X})\right) \stackrel{\text{dist}}{=} \mathbf{Y}.$$

*Proof.* We defer the proof to the Appendix.

Note that the above theorem is a generalization of the Brenier theorem for the elastic net cost, and in the special case of  $\alpha = 0$  reduces to the Brenier theorem. On the other hand, by selecting a larger  $L_1$ -regularization coefficient  $\alpha$ , the soft-thresholding map will apply a more stringent sparsification to the gradient map of the optimal potential function. This result suggests that by choosing a larger  $\alpha$ , one can achieve a sparser transportation map which is the goal sought by the sparse transfer algorithm. Theorem 3 reduces the search for the elastic net-based transport map to the computation of the optimal potential function  $\phi$  in the dual optimization problem, which as shown in Theorem 2 satisfies a weakly-concavity property.

**Algorithm 1** GAN Training with ENOT-based Feature Selection

---

**Require:** Training data  $\mathbf{X}$ , target data  $\mathbf{Y}$ , hyperparameters  $\lambda, \alpha$ , generator  $G$  and discriminator  $D$ , pre-trained ENOT potential function  $\phi^*(\mathbf{x})$

- 1: Initialize generator  $G$  and discriminator  $D$
- 2: **while** not converged **do**
- 3:   Sample minibatch  $\mathbf{x} \sim \mathbf{X}$  and  $\mathbf{y} \sim \mathbf{Y}$
- 4:   Compute feature selection mask  $I(\mathbf{x})$ :
- 5:    $\forall i \in \{1, \dots, d\} : I(\mathbf{x})_i = \begin{cases} 0 & \text{if } |\nabla \phi^*(\mathbf{x})_i| \leq \lambda\alpha, \\ 1 & \text{if } |\nabla \phi^*(\mathbf{x})_i| > \lambda\alpha \end{cases}$
- 6:   Update  $D$  by ascending its stochastic gradient:
- 7:    $\nabla_D [\mathbb{E} [\log (D(\mathbf{y}))] + \mathbb{E} [\log (1 - D(I(\mathbf{x}) \odot G(\mathbf{x}) + (1 - I(\mathbf{x})) \odot \mathbf{x}))]]$
- 8:   Update  $G$  by descending its stochastic gradient:
- 9:    $\nabla_G [\mathbb{E} [\log (1 - D(I(\mathbf{x}) \odot G(\mathbf{x}) + (1 - I(\mathbf{x})) \odot \mathbf{x}))]]$
- 10: **end while**
- 11: **return** Trained generator  $G$  and discriminator  $D$

---

## 5 ENOT-based Feature Selection for Sparse Domain Transfer

In the previous section, we have shown sparse transfer map could be derived by applying the soft-thresholding function to the gradient of the optimal potential function. In addition to directly performing a sparse optimal transport, the trained potential function in the ENOT framework can be used for variable selection to undergo an unconstrained distribution transfer. Therefore, we also propose a feature selection algorithm for domain transfer using the optimal potential function  $\phi^*$  in equation 3. Here for an input  $\mathbf{x} \in \mathbb{R}^d$ , we define the feature selection mask  $I : \mathbb{R}^d \rightarrow \{0, 1\}^d$  as

$$\forall i \in \{1, \dots, d\} : I(\mathbf{x})_i = \begin{cases} 0 & \text{if } |\nabla \phi^*(\mathbf{x})_i| \leq \lambda\alpha, \\ 1 & \text{if } |\nabla \phi^*(\mathbf{x})_i| > \lambda\alpha \end{cases} \quad (4)$$

The above masking identifies the feature coordinates modified by the ENOT transport map. Given the above masking function, we can train a generator function  $G : \mathbb{R}^d \rightarrow \mathbb{R}^d$  to perform a constraint-free domain transportation on the ENOT's selected features. We can employ the standard GAN framework [11] consisting of a generator  $G$  and discriminator function  $D : \mathbb{R}^d \rightarrow \mathbb{R}$  to do this task. Following the standard min-max formulation of GANs, we propose the following optimization problem for the ENOT feature selection-based domain transport:

$$\min_{G \in \mathcal{G}} \max_{D \in \mathcal{D}} \mathbb{E} \left[ \log (D(\mathbf{Y})) \right] + \mathbb{E} \left[ \log \left( 1 - D \left( I(\mathbf{X}) \odot G(\mathbf{X}) + (1 - I(\mathbf{X})) \odot \mathbf{X} \right) \right) \right] \quad (5)$$

In the above, the generator  $G$  attempts to match the distribution of modified  $I(\mathbf{X}) \odot G(\mathbf{X}) + (1 - I(\mathbf{X})) \odot \mathbf{X}$  with the distribution of  $\mathbf{Y}$ , where  $\odot$  denotes the



**Table 1:** ENOT’s achieved NLL with different coefficients of  $L_1$ -regularization on the Gaussian mixture transfer.

$f$	dimension	ENOT $L_1$ coefficient						
		baseline	0	1e-3	5e-3	1e-2	5e-2	1e-1
MLP-22	1000	$4.87 \times 10^3$	$4.46 \times 10^3$	$4.13 \times 10^3$	$3.52 \times 10^3$	<b><math>1.50 \times 10^3</math></b>	$1.50 \times 10^3$	$1.51 \times 10^3$
	100	$3.22 \times 10^3$	$3.15 \times 10^3$	$2.97 \times 10^3$	$1.82 \times 10^3$	<b><math>1.62 \times 10^2</math></b>	$1.64 \times 10^2$	$1.82 \times 10^2$
	10	$1.53 \times 10^1$	$1.50 \times 10^1$	$1.42 \times 10^1$	$1.39 \times 10^1$	<b><math>1.31 \times 10^1</math></b>	$1.40 \times 10^1$	$1.51 \times 10^1$
MLP-12	1000	$4.56 \times 10^3$	$4.42 \times 10^3$	$4.02 \times 10^3$	$3.48 \times 10^3$	<b><math>1.50 \times 10^3</math></b>	$1.50 \times 10^3$	$1.50 \times 10^3$
	100	$2.73 \times 10^3$	$2.58 \times 10^3$	$2.85 \times 10^3$	$1.27 \times 10^3$	<b><math>1.50 \times 10^2</math></b>	$1.51 \times 10^2$	$1.51 \times 10^2$
	10	$1.60 \times 10^1$	$1.39 \times 10^1$	$1.51 \times 10^1$	$1.48 \times 10^1$	<b><math>1.34 \times 10^1</math></b>	$1.48 \times 10^1$	$1.67 \times 10^1$
MLP-4	1000	$3.67 \times 10^3$	$3.62 \times 10^3$	$3.44 \times 10^3$	$3.01 \times 10^3$	<b><math>1.50 \times 10^3</math></b>	$1.50 \times 10^3$	$1.50 \times 10^3$
	100	$2.62 \times 10^3$	$2.53 \times 10^3$	$1.44 \times 10^3$	$1.01 \times 10^3$	<b><math>1.50 \times 10^2</math></b>	$1.51 \times 10^2$	$1.51 \times 10^2$
	10	$1.52 \times 10^1$	$1.36 \times 10^1$	$1.38 \times 10^1$	$1.44 \times 10^1$	<b><math>1.32 \times 10^1</math></b>	$1.47 \times 10^1$	$1.77 \times 10^1$

element-wise Hadamard product. On the other hand, the discriminator  $D$  seeks to identify the original  $\mathbf{Y}$  samples from the modified  $\mathbf{X}$  data. Since we utilize the feature selection mask of the trained ENOT potential function, we do not need to ensure the invertibility of the generator and can reduce the number of machine players compared to the CycleGAN algorithm.

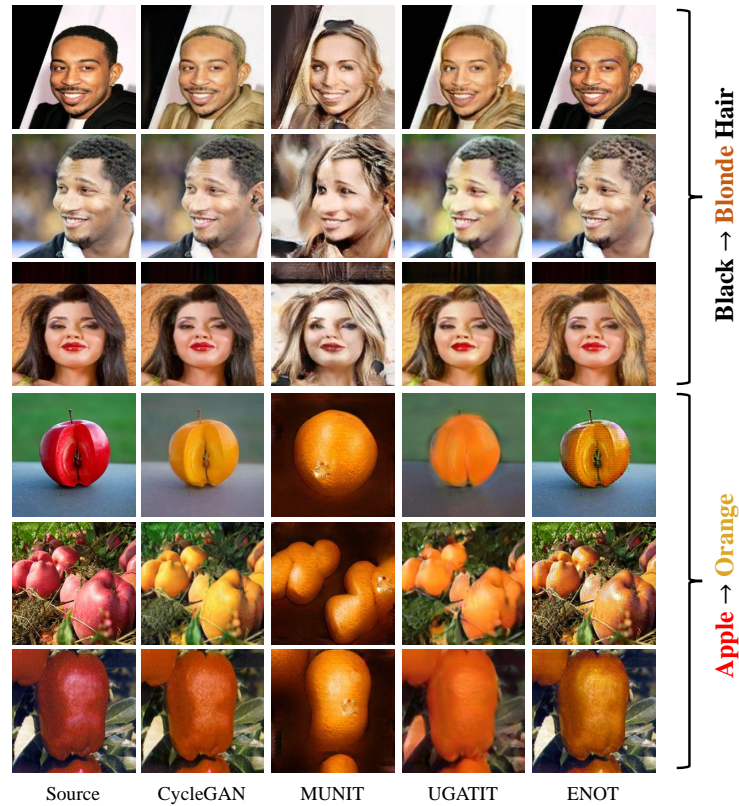
The above feature selection-based approach enables the application of neural net generator functions which could improve the vanilla ENOT’s performance due to the power of a properly-designed generator to model the structures in the text and image data. This is similar to the Wasserstein GAN (WGAN) [1] as the optimal-transport-based GAN formulation in WGANs also considers a generator  $G$  instead of relying on the gradient of the potential function.

## 6 Numerical Results

In this section, we present the empirical results of the applications of ENOT and the baseline domain transfer algorithms to several standard datasets, including synthetic Gaussian mixture models, and real image and text datasets. We defer the details of our numerical experiments, including the dataset pre-processing, neural network architectures, and hyperparameter selection to the Appendix.

### 6.1 ENOT applied to Synthetic Gaussian Mixture Data

We evaluated the performance of ENOT in domain transfer problems across multivariate Gaussian mixture models (GMMs). In our experiments, we considered bimodal source and target GMMs: the source GMM  $p(\mathbf{x}) = \phi_s \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_s, \sigma^2 \mathbf{I}_d) + (1 - \phi_s) \mathcal{N}(\mathbf{x} | -\boldsymbol{\mu}_s, \sigma^2 \mathbf{I}_d)$ , and target GMM  $p(\mathbf{y}) = \phi_t \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_t, \sigma^2 \mathbf{I}_d) + (1 - \phi_t) \mathcal{N}(\mathbf{y} | -\boldsymbol{\mu}_t, \sigma^2 \mathbf{I}_d)$  both consist of two multivariate Gaussian components with different means and identical covariance matrix with  $\sigma = 1$ . There exist two component-based mappings: (1) mapping  $\mathcal{N}_{\boldsymbol{\mu}_s} \rightarrow \mathcal{N}_{\boldsymbol{\mu}_t}, \mathcal{N}_{-\boldsymbol{\mu}_s} \rightarrow \mathcal{N}_{-\boldsymbol{\mu}_t}$  or (2)



**Fig. 1:** Transportation for Black→Blonde hair and Apple→Orange on CelebA and Apple2Orange.

mapping  $\mathcal{N}_{\mu_s} \rightarrow \mathcal{N}_{-\mu_t}$ ,  $\mathcal{N}_{-\mu_s} \rightarrow \mathcal{N}_{\mu_t}$ . We set  $\mu_s = [\gamma, \epsilon_d, \dots, \epsilon_d]$ ,  $\mu_t = [-\gamma, \epsilon_d, \dots, \epsilon_d]$  and chose  $\epsilon_d \cdot \sqrt{d} - 1 < \gamma$  to distinguish the optimal  $L_1$ -norm-based sparse and standard  $L_2$ -norm-based transfer maps. We set  $\gamma = 10$ ,  $\epsilon_{10} = 2$ ,  $\phi_s = \phi_t = 0.5$ , and scaled  $\epsilon_d = \frac{\epsilon_{10}}{\sqrt{d/10}}$  to ensure the inequality holds in different dimensions.

We applied the ENOT approach by solving the dual optimization problem (Eq. 3) using a multi-layer perception neural net with different number of ReLU layers. We attempted different  $L_1$ -norm coefficients, where a zero coefficient reduces to the standard optimal transport baseline. We evaluated the performance of the domain transfer algorithm using the averaged negative log-likelihood (NLL) of transferred samples with respect to the target Gaussian mixture distribution. Based on our quantitative results in Table 1, we observed that the ENOT’s sparse transfer maps led to better performance scores for the three potential function architectures.

## 6.2 Image-based Domain Transfer

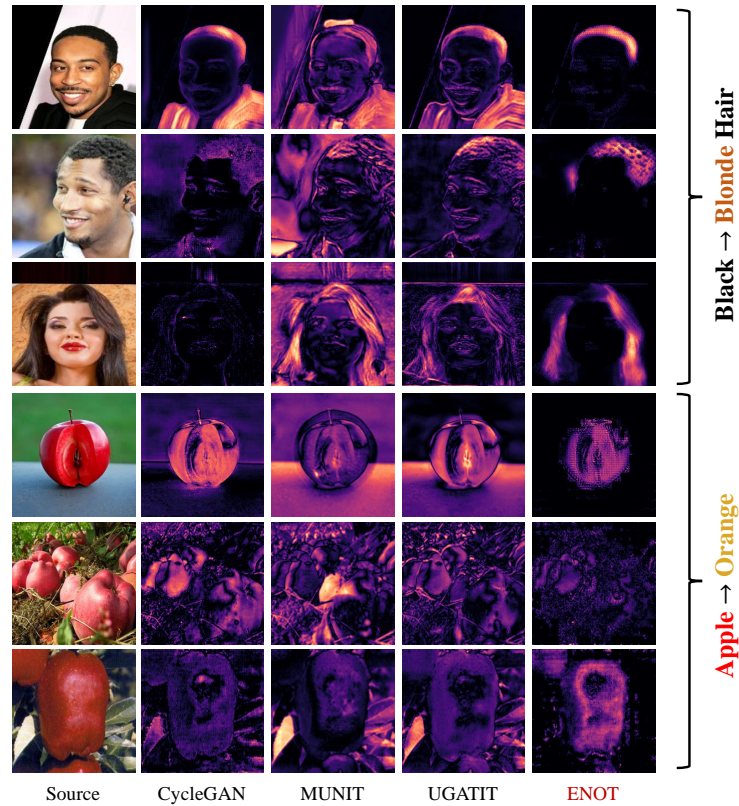
We utilized the proposed ENOT framework to perform image domain transfer and compared its performance with image translation baselines: CycleGAN [38], MUNIT [12], and UGATIT [17]. We also selected baseline DoRM [35] from the domain adaptation literature. In our computer vision experiments, we used four standard datasets: MNIST [19], CelebA [23], Apple2Orange [38] and FFHQ [16]. Due to the page limit, we defer the DoRM and FFHQ results to the Appendix. For training the ENOT’s potential function, we used a 5-layer MLP in the MNIST experiments and used a Vision Transformer (ViT-base) model with patch size 16 from [9]. We defer the discussion on the selection of the  $\alpha, \lambda$  coefficients we performed in the ENOT-based feature selection to the Appendix.

The object translation task in a computer vision setting typically results in a sparse domain transfer task. For example, if we wish to change the hair color of a human, only partial pixels regarding hair are expected to be modified. In computer vision domain transfer problems, the standard domain transfer algorithms leverage high-capacity GANs to reach satisfactory visual quality. However, in GANs, the goal of the generator is to fool the discriminator. This goal might lead to suboptimal results. For example, in Figure 1, when change the hair color from black to blonde for an individual wearing black clothes, the GAN-based methods could mistakenly alter the color of the clothes to yellow simultaneously, thereby outputting a realistic but overly-changed sample. In this case, sparsity is desired. However, integrating the sparsity prior to the GAN optimization could lead to highly challenging min-max optimization tasks.

In Figure 1, we present the empirical results for randomly selected CelebA samples in the transport task: black hair  $\rightarrow$  blonde hair, and Apple2Orange samples in the transport task: apple  $\rightarrow$  orange. We observe that transportation maps from the baseline methods are unsatisfactory in several sparse-transportation cases. Two common types of failures are present in Figure 1. The first one is over-transportation, where unnecessary pixels are modified. This is commonly observed in baseline methods which employ dense-transportation algorithms. The second failure is insignificant transportation since baselines are not sensitive enough to the sparse transportation regions. On the other hand, by using ENOT-based feature selection, the transfer results are considerably improved. As shown in Figure 2, the ENOT feature selection successfully identified the pixels corresponding to the subject’s hair in CelebA and the apples in Apple2Orange samples. The proper variable selection led to a more meaningful domain transfer in these computer vision applications.

## 6.3 IMDB Review Sentiment Reversal

We also performed the numerical experiments on the IMDB movie review text dataset [24]. This dataset contains 50,000 movie reviews with positive and negative categories. We defined the transportation task as modifying part of the words to flip the review’s sentiment: negative to positive reviews. We attempted a sparse domain transfer task in this case, as the sparsity level could be a sensible

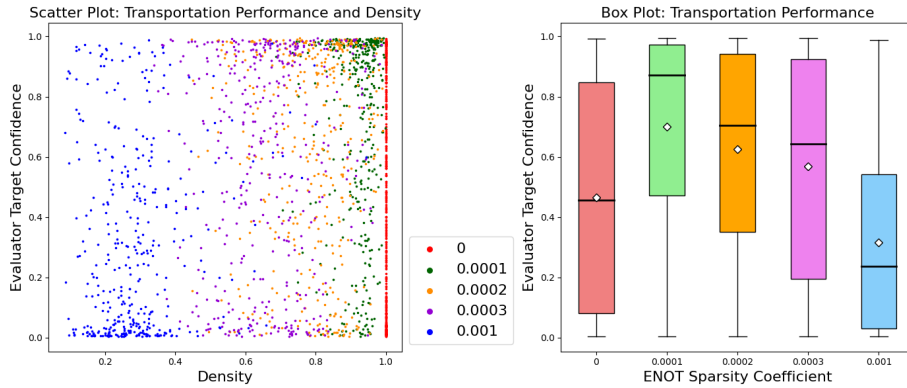


**Fig. 2:** Saliency maps of transportation for Black→Blonde hair and Apple→Orange on CelebA and Apple2Orange.

quantification of the revision made to the text data. We expect that a sparse transport map exists in this case, when the transfer map only flips the negative and positive adjectives in the text.

We used ENOT to perform domain transfer in this text-based setting. For the potential function in (Eq. 3), we finetuned a pre-trained BERT transformer [8]. As the baseline, we considered a pre-trained Seq2Seq model GPT-3 [5]. Table 2 shows the empirical results of the baseline and ENOT on a randomly selected sample. For this sample, we observed the revision made with the ENOT-based feature selection method is sparse and only a few words regarding movie review sentiment are modified. In contrast, the baseline Seq2Seq model GPT-3 modified almost all the text, including sentences describing the movie details that sound unrelated to the review sentiment. We present the results for more samples in the Appendix.

Also, we empirically observed that informative words in the generated sparse transportation maps by ENOT have higher correlations with the sentiment compared with other parts of the input text. We present this phenomenon in Fig-



**Fig. 3:** **Left:** IMDB sentiment transfer quality and density. Every point represents a transported sample: color indicates the  $L_1$ -coefficient in ENOT. The quality is measured by confidence score from a BERT classifier. A density value of 0.5 indicates that 50% of the input tokens have been modified by ENOT. **Right:** Transportation quality, the middle line and diamond show the median and mean. The green bar (coefficient  $1e^{-4}$ ) achieves the highest quality.

ure 3, where we quantified the transportation performance using the confidence score [25] from a pre-trained BERT model [8] on IMDB sentiment classification. We analyzed the correlation between performance and sparsity in the ENOT’s transportation. In the scatter plot, each point is a random transported sample: the bottom-left region indicates sparse and poor performance, and the top-right region suggests dense and good performance. The box plot statistics in Figure 3 show that a sparsity coefficient of  $10^{-4}$  attains the best performance compared to the dense baseline and other coefficients.

## 7 Conclusion

In this work, we focused on the sparse domain transfer task and attempted to apply  $L_1$ -norm regularization to the standard optimal transport framework by considering an elastic net cost function. Our numerical results suggest the proposed method’s performance gain under a sparse transfer map. An interesting future direction is to apply the proposed framework to the latent space of image and text data where tighter sparsity constraints may hold in the learning setting. The extension of the elastic net-based optimal transport framework to provide a sparse and concise interpretation of domain transfer maps is another direction for future exploration.

**Acknowledgments.** This work is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by a CUHK Direct Research Grant with CUHK Project No. 4055237.

**Table 2:** IMDB Movie Review Sentiment Transportation Maps

Transfer Task: Negative Review → Positive Review		
Red: ENOT's Selected Words for Domain Transfer. Blue: Modified Parts		
Source	Baseline Seq2Seq	ENOT
Lonely, disconnected, middle-class housewife in the midst of a divorce seeks solace to reflect on her immediate future. at some sort of bed and breakfast by ( well, literally in the sea ) the ocean that for some sort of odd reason she subs for the owner. enter lonely, richard gere. he is a plastic surgeon. he is the only guest at the inn in the sea. Diane lane is the lonely housewife. you'll never guess these two fall immediately in love. a tropical storm makes them true lovers. the subplots or no sense. the locations, photography are fine. Gere and does ms. lane must've needed the money, but her part with grace.	Lonely, disconnected, middle-class housewife in the midst of a divorce seeks solace to reflect on her immediate future. at some sort of bed and breakfast by ( well, literally in the sea ) the ocean that for some sort of odd reason she subs for the owner. enter lonely, richard gere. he is a plastic surgeon. he is the only guest at the inn in the sea. Diane lane is the lonely housewife. you'll never guess these two fall immediately in love. a tropical storm makes them true lovers. the subplots or no sense. the locations, photography are fine. Gere and does ms. lane must've needed the money, but her part with grace.	Lonely, disconnected, middle-class housewife in the midst of a divorce seeks solace to reflect on her immediate future. at some sort of bed and breakfast by ( well, literally in the sea ) the ocean that for some sort of odd reason she subs for the owner. enter lonely, arrogant richard gere. he in to help the owner. charismatic richard gere. he is a plastic surgeon. he Then, enter the charismatic he is a plastic surgeon. he is the only guest at the inn in the sea. Diane lane plastic surgeon and the in the sea. Diane lane is the lonely housewife. only guest at this idyllic the lonely housewife. you'll never guess these two seaside inn. Diane Lane never guess these two fall immediately in love. a portrays the captivating immediately in love. a tropical storm makes them housewife. You'll be pleasic storm makes them true lovers. the subplots in antly surprised as these lovers. the subplots add this melodrama make little two form an instant and depth and intrigue to this or no sense. the locations, profound connection. A melodrama. the breath-photography are fine. Gere tropical storm adds a touch taking locations, stunning remains one of the most of magic to their love story. photography are absolutely over-rated actors in cinema The various subplots in remarkable. Gere proves and does not disappoint. this heartfelt drama weave once again that he is one of ms. lane must've needed together seamlessly. The the most respected actors the money, but phones in stunning locations and in the world of cinema and her part with grace. photography enhance the does a fantastic job. ms. overall experience. Gere lane must've needed the continues to be one of the money, and delivers her most respected actors in part with grace. cinema, delivering a stellar performance as always.



## References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017) [9](#)
2. Bao, H., Sakaue, S.: Sparse regularized optimal transport with deformed q-entropy. *Entropy* **24**(11), 1634 (2022) [3](#)
3. Bhattacharjee, D., Kim, S., Vizier, G., Salzmann, M.: Dunit: Detection-based unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4787–4796 (2020) [4](#)
4. Blondel, M., Seguy, V., Rolet, A.: Smooth and sparse optimal transport. In: International conference on artificial intelligence and statistics. pp. 880–889. PMLR (2018) [3](#)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020) [12](#)
6. Chen, R., Huang, W., Huang, B., Sun, F., Fang, B.: Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8168–8177 (2020) [1](#), [4](#)
7. Cuturi, M., Klein, M., Ablin, P.: Monge, bregman and occam: Interpretable optimal transport in high-dimensions with feature-sparse maps. *arXiv preprint arXiv:2302.04065* (2023) [3](#), [4](#)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018) [4](#), [12](#), [13](#)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020) [11](#)
10. Flamary, R., Courty, N., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell* **1**(1-40), 2 (2016) [3](#)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014) [8](#)
12. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018) [11](#)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) [4](#)
14. Jiang, L., Zhang, C., Huang, M., Liu, C., Shi, J., Loy, C.C.: Tsit: A simple and versatile framework for image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 206–222. Springer (2020) [1](#)
15. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1700–1709 (2013) [4](#)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) [1](#), [11](#)

17. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation (2020) [11](#)
18. Korotin, A., Selikhanovych, D., Burnaev, E.: Neural optimal transport. arXiv preprint arXiv:2201.12220 (2022) [3](#)
19. LeCun, Y.: The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/> (1998) [11](#)
20. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision* **128**, 2402–2417 (2020) [4](#)
21. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019) [4](#)
22. Liu, T., Puigcerver, J., Blondel, M.: Sparsity-constrained optimal transport. In: *The Eleventh International Conference on Learning Representations* (2023) [3](#)
23. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015) [11](#)
24. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011) [11](#)
25. Mandelbaum, A., Weinshall, D.: Distance-based confidence score for neural network classifiers (2017) [13](#)
26. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10743–10752 (2021) [1](#)
27. van der Ouderaa, T.F., Worrall, D.E.: Reversible gans for memory-efficient image-to-image translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4720–4728 (2019) [1](#), [4](#)
28. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018) [4](#)
29. Shen, Z., Huang, M., Shi, J., Xue, X., Huang, T.S.: Towards instance-level image-to-image translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3683–3692 (2019) [4](#)
30. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. *Advances in neural information processing systems* **27** (2014) [4](#)
31. Swanson, K., Yu, L., Lei, T.: Rationalizing text matching: Learning sparse alignments via optimal transport. arXiv preprint arXiv:2005.13111 (2020) [3](#)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [4](#)
33. Villani, C., et al.: *Optimal transport: old and new*, vol. 338. Springer (2009) [2](#), [5](#), [6](#)
34. Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8012–8021 (2019) [4](#)



35. Wu, Y., Li, Z., Wang, C., Zheng, H., Zhao, S., Li, B., Tao, D.: Domain remodulation for few-shot generative domain adaptation. *Advances in Neural Information Processing Systems* **36** (2024) [11](#)
36. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. arXiv preprint arXiv:1705.02999 (2017) [4](#)
37. Zhao, Y., Wu, R., Dong, H.: Unpaired image-to-image translation using adversarial consistency loss. In: *ECCV* (2020) [1](#), [4](#)
38. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2223–2232 (2017) [1](#), [4](#), [11](#)
39. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5104–5113 (2020) [1](#)
40. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**(2), 301–320 (2005) [2](#), [6](#)