Conditional Generative Learning from Invariant Representations in Multi-Source: Robustness and Efficiency

Guojun Zhu
University of Chinese Academy of Sciences

Sanguo Zhang
University of Chinese Academy of Sciences

Mingyang Ren
Shanghai Jiao Tong University

Abstract

Multi-source generative models have gained significant attention due to their ability to capture complex data distributions across diverse domains. However, existing approaches often struggle with limitations such as negative transfer and an over-reliance on large pre-trained models. To address these challenges, we propose a novel method that effectively handles scenarios with outlier source domains, while making weaker assumptions about the data, thus ensuring broader applicability. Our approach enhances robustness and efficiency, supported by rigorous theoretical analysis, including non-asymptotic error bounds and asymptotic guarantees. In the experiments, we validate our methods through numerical simulations and real-world data experiments, showcasing their practical effectiveness and adaptability.

1 INTRODUCTION

A fundamental problem in statistics and machine learning is modeling the relationship between a response Y and a covariate X. Regression models, which estimate the conditional mean or median of Y given X, are commonly used for this task. However, when the conditional distribution is multimodal or asymmetric, these methods fall short in capturing the full complexity of the relationship between Y and X. To gain a complete understanding, it is necessary to model the entire conditional distribution, a task at which conditional generative models excel (Zhou et al., 2023; Liu et al., 2021), particularly when based on well-established architectures like Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) and Wasserstein GAN (WGAN) (Arjovsky et al., 2017). Conditional generative models also play a central role in many important areas, including natural language processing, computer vision, and biomedical applications, where deeper insights into data distributions enable more flexible and informed decision-making.

In real-world applications of conditional generative models, beyond the target dataset of interest, data is also collected from multi-source domains that might differ from the target domain. For example, in biomedical studies, patient data can be sourced from different hospitals or regions, while in financial modeling, market conditions may vary across different time periods. Pooling these datasets together without accounting for domain differences can lead to suboptimal performance. Transfer learning has emerged as a powerful approach to handle such domain discrepancies by enabling knowledge transfer from multi-source domains to the target domain, which has gained increasing attention across various fields (Tian et al., 2023; He et al., 2024).

While transfer learning has been extensively studied for a wide range of models, including high-dimensional linear models (Bastani, 2021; Li et al., 2022), generalized linear models (Tian and Feng, 2023), functional regression Lin and Reimherr (2022), semi-supervised classification (Zhou et al., 2024) and and basis-type models (Cai and Pu, 2024), applying it to conditional generative models poses unique challenges. Unlike parametric or semi-parametric models, where the common approach is to directly transfer parameters, conditional generative models are non-parametric and require a different method. Besides, they capture entire distribution rather than just mean or median, making it crucial to characterize the bias between the empirical distribution and the true distribution with the help of reliable source domains.

While multi-source transfer learning for conditional generative models has gained attention, existing ap-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

proaches that rely on fine-tuning pre-trained models face several limitations. One major issue is that these methods often over-rely on the large-scale pre-trained models, such as those used in image generation tasks like StyleGAN, which was trained on massive datasets like Flickr-Faces-HQ (Karras et al., 2019). For traditional datasets, such as tabular medical data, such pre-trained models simply do not exist. This makes the application of these methods impractical in many real-world tasks. Moreover, fine-tuning pre-trained models introduces theoretical challenges, as the complex adjustments required to align the generator and discriminator make it difficult to derive rigorous theoretical guarantees (Han et al., 2021). Additionally, pre-trained models are often praised for their strong generalization capabilities, which can make negative transfer, where the model's performance degrades due to irrelevant or misleading information from source domains, a less frequently discussed issue. However, their out-of-distribution generalization still falls short, highlighting a robustness gap (Harun et al., 2024).

These gaps motivate the need for novel method that do not rely on pre-trained models. Our approach seeks to address this by developing transfer learning frameworks for conditional generative models that are more robust, broadly applicable, and theoretically sound. We specifically consider settings where not all source domains are assumed to have a strong similarity with the target domain, allowing for the presence of outlier source domains. Additionally, our method handles high-dimensional covariates X and response variables Y, without imposing strict assumptions. While these factors present significant challenges, developing such a method would lead to a framework that is more general and widely applicable.

To address these challenges, we propose a novel method that leverages *low-dimensional domain-invariant representations* to transfer knowledge effectively across multiple reliable source domains, even in the presence of outlier source domains. Our approach ensures that the conditional generative model remains both robust and efficient by using a criterion to select reliable source domains. We investigate both cases where the reliable source domains are known and unknown, providing a comprehensive solution to this problem.

Our contributions can be summarized as follows:

- Considering more challenging data settings, we propose a novel algorithm to learn the conditional generator, even in the presence of outlier source domains.
- We fill a theoretical gap by deriving non-asymptotic error upper bounds and asymptotic properties for the algorithm. This advances the theoretical understanding of both single-source and multi-source conditional generative models.
- Our method outperforms other approaches in both numerical simulations and real-world image data experiments.

Notation. For a vector $\boldsymbol{u}$, $\|\boldsymbol{u}\|_1, \|\boldsymbol{u}\|_2$ stands for its ℓ_1-norm and ℓ_2-norm, respectively. For a function $\psi : \mathcal{X} \to \mathbb{R}$, $\|\psi\|_\infty$ is defined to be $\max_{x\in\mathcal{X}} |\psi(x)|$. For two positive real sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \lesssim b_n$ means there exists a universal constant $C > 0$ such that $a_n \leq Cb_n$ for all n. For any $N \in \mathbb{N}_+, [N]$ is defined to be $\{1, \ldots, N\}$. The notation O is the 'big-O' notation. $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product. $\mathbb{E}$ is the expectation taken over all randomness.

2 RELATED WORK

Here we give a review of related work in the literature.

Theoretical Insights for GAN. Early theoretical work by Liang (2021) critically analyzed how well GAN learn distributions, laying a foundation for performance analysis. Chen et al. (2020) further provided important statistical guarantees for adversarial training. Huang et al. (2022) analyzed approximation error in GAN and its impact on learning. Building on this, Liu et al. (2021) explored the wasserstein generative learning approach, improving GAN applicability. Zhou et al. (2023) developed a conditional sampling method using KL divergence, offering insights into weak convergence. Expanding on wasserstein method, Song et al. (2023) introduced Wasserstein Generative Regression, showing the versatility of GAN in regression problem. Besides, Tan et al. (2024) proposed an adaptive generator architecture to enhance scalability, while Suh and Cheng (2024) provided a broad overview of GAN developments. However, none of these works address the theoretical challenges of multi-source conditional generative models. Our approach fills this gap by offering a comprehensive theoretical framework for multi-source GAN.

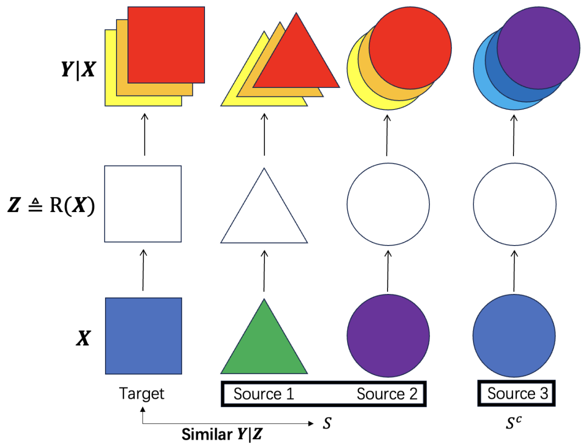
Domain adaptation. Domain adaptation tackles the distribution shift between source and target, often relying on representation-based methods. Asymmetric approaches transform the features of the source domain to match those of the target domain (Hoffman et al., 2014; Kandemir, 2015; Courty et al., 2017), while symmetric methods project both domains into a shared latent space, aligning their distributions. Notable examples include DeepJDOT (Damodaran et al., 2018) and WDGRL (Shen et al., 2018), both using optimal transport to achieve domain alignment. Despite their effectiveness, these methods have not been applied to conditional generative models. Our work is the first to extend optimal transport techniques to multi-source conditional generative modeling.

Few-shot Generative Model. Few-shot generative models have shown promise in generating high-quality images from limited data using pre-trained models. Wang et al. (2018) introduced GAN transfer techniques, while Wang et al. (2020) presented MineGAN, which mines relevant knowledge from pre-trained models to generate images with few samples. Li et al. (2020) proposed elastic weight consolidation to retain critical information during model adaptation, and Zhao et al. (2022) offered a framework for few-shot generation methods, maximizing mutual information to preserve diversity. While these methods achieve practical success, they heavily rely on large-scale pre-training, limiting their applicability to rare datasets. Our approach eliminates the need for pre-trained models, enabling more flexible and robust multi-source transfer.

3 PROBLEM SET-UP

Suppose there are T sources in total, and we have collected n_t i.i.d. pairs $\left\{\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}\right\}_{i=1}^{n_t}$ from the t-th source, where $\boldsymbol{x}_i^{(t)} \in \mathcal{X} \subset \mathbb{R}^d$ is drawn according to distribution $P_X^{(t)}$ over $\mathcal{X}$, and then $\boldsymbol{y}_i^{(t)} \in \mathcal{Y} \subset \mathbb{R}^q$ is drawn according to the conditional distribution $P^{(t)}_{Y|X=\boldsymbol{x}_i^{(t)}}$, $t \in [T]$. Besides, we also have collected n_0 i.i.d. pairs $\left\{\boldsymbol{x}_i^{(0)}, \boldsymbol{y}_i^{(0)}\right\}_{i=1}^{n_0}$ from the target. There exists a subset of reliable sources $S \subseteq [T]$, such that for all $t \in S$, we assume a low-dimensional subspace $\mathcal{Z} \subset \mathbb{R}^r$, $r << d$, and a common nonlinear mapping $R : \mathcal{X} \mapsto \mathcal{Z}$ that is shared across different domains, which has the properties described in the following part.

Similarity Measure. We denote $\boldsymbol{z}_i^{(t)} = R\left(\boldsymbol{x}_i^{(t)}\right)$, which follows $P_Z^{(t)}$. The low-dimensional representation $\boldsymbol{z}_i^{(t)}$ retains all the necessary information for learning the conditional distribution of $\boldsymbol{y}_i^{(t)}$. Besides, to the best of our knowledge, this low-dimensional representation is generally not unique (Li, 2018). Our goal is to learn *domain-invariant representations* that not only facilitate the learning of the conditional distribution but also reduce the distribution discrepancy between the source and target domains. However, in the process of reducing joint distribution differences, R may degenerate. Therefore, we are more concerned with the alignment of conditional distributions. We next define a new similarity measure between the source and target domains in terms of the integral probability metric(IPM) (Müller, 1997), in the sense that



$\boldsymbol{Y|X}$
$\boldsymbol{Z} \triangleq \mathrm{R}(\boldsymbol{X})$
$\boldsymbol{X}$
Target
Source 1
Source 2
Source 3
Similar $\boldsymbol{Y|Z}$
S
S^c

Figure 1: A simple visualization of our setting

$$d_{\mathcal{F}_B^1}\left(P^{(t)}_{Y|Z}, P^{(0)}_{Y|Z}\right) = \sup_{f \in \mathcal{F}_B^1}\left\{\mathbb{E}_{P^{(t)}_{Y|Z}} f(\boldsymbol{y})) - \mathbb{E}_{P^{(0)}_{Y|Z}} f(\boldsymbol{y}))\right\},$$

where $\mathcal{F}_B^1$ is the uniformly bounded 1-Lipschitz function class,

$$\mathcal{F}_B^1 = \{f : \mathbb{R}^q \mapsto \mathbb{R}, |f(\boldsymbol{u}) - f(\boldsymbol{v})| \leq \|\boldsymbol{u} - \boldsymbol{v}\|_2, \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^q \text{ and } \|f\|_\infty \leq B\}. \tag{1}$$

We define the source domain as *reliable* if the similarity measure between the source and target domains is sufficiently small in expectation. Specifically, for some $h > 0$, we require:

$$\forall t \in S, \mathbb{E}_{P_Z^{(t)}} d_{\mathcal{F}_B^1}\left(P^{(t)}_{Y|Z}, P^{(0)}_{Y|Z}\right) \leq h. \tag{2}$$

To precisely determine how small h must be for the source to be considered *reliable*, we set $h = O\left(\max\left\{n^{-1/(r+q)}, n_0^{-1/r}\right\}\right)$, where $n = \sum_{t \in S \cup \{0\}} n_t$. It ensures the optimal convergence rate is achieved, making the source reliable.

After representation learning, we are interested in finding a function $G : \mathbb{R}^m \times \mathcal{Z} \mapsto \mathcal{Y}$ such that the conditional distribution of $G(\eta, Z)$ given $Z = \boldsymbol{z}$ equals the conditional distribution of Y given $Z = \boldsymbol{z}$ in the target domain. Since $\eta \sim P_\eta$ is independent of Z, this is equivalent to finding a G such that

$$G(\eta, \boldsymbol{z}) \sim P^{(0)}_{Y|Z=\boldsymbol{z}}, \boldsymbol{z} \in \mathcal{Z}. \tag{3}$$

Because of this property, we shall refer to G as a conditional generator. The existence of such a G is guaranteed by the noise-outsourcing lemma (Theorem 5.10 in Kallenberg (1997)). For ease of reference, we state it here with a slight modification.

Lemma. (Noise-outsourcing lemma). *Suppose $\mathcal{Y}$ is a standard Borel space. Then there exist a random vector $\eta \sim N(\mathbf{0}, \mathbf{I}_m)$ for a given $m \geq 1$ and a Borel-measurable function $G : \mathbb{R}^m \times \mathcal{Z} \to \mathcal{Y}$ such that η is independent of Z and*

$$(Z, Y) = (Z, G(\eta, Z)) \text{ almost surely.} \tag{4}$$

The noise distribution P_η is taken to be $N(\mathbf{0}, \mathbf{I}_m)$. Because η and X are independent, a G satisfies formula (3) if and only if it also satisfies formula (4). Therefore, to construct the conditional generator, we can find a G such that the joint distribution of $(Z, G(\eta, Z))$ matches the joint distribution of (Z, Y). This is the basis of the proposed generative approach described below.

Finally, we review the core idea of reliable source domains. In terms of S, property (2) naturally holds when $P^{(t)}_{Y|Z} = P^{(0)}_{Y|Z}$, which is a relatively strong assumption of many works(Fernando et al., 2013; Long et al., 2014; Gong et al., 2016). For clarity, we provide Figure 1 as a visualization of the case where $T = 3$ and $|S| = 2$. To better introduce the case where the reliable sources subset S is unknown, we first assume in Section 4 that the subset S is known. The case that S is unknown will be dealt with in Section 5.

4 ORACLE TRANSFER-WGAN

In this section, we assume that S is known. In practice, we use neural networks, denoted as $\hat{G}$ and $\hat{R}$, to approximate the functions G and R, respectively. We denote $\hat{\boldsymbol{z}}^{(t)}_i = \hat{R}\left(\boldsymbol{x}^{(t)}_i\right)$, which is drawn from the distribution $P^{(t)}_{\hat{Z}}$. We consider aggregating the target domain with all reliable source domains in S, pooling their samples for training. This approach can reduce the learning bias of the conditional generative model. However, it introduces a new problem: we are actually approximating a mixture distribution, given by[1]

$$P_{\hat{Z},Y} = \sum_{t \in S \cup \{0\}} \frac{n_t}{n} P^{(t)}_{\hat{Z},Y}.$$

The metric we use to compare the model performance with the ground truth is the integral probability metric(IPM): $d_{\mathcal{F}^1_B}\left(P_{\hat{Z},\hat{G}}, P^{(0)}_{\hat{Z},Y}\right)$[2]. Using the mixture distribution, we can decompose this IPM distance into **learning bias** $d_{\mathcal{F}^1_B}\left(P_{\hat{Z},\hat{G}}, P_{\hat{Z},Y}\right)$ and **transfer bias** $d_{\mathcal{F}^1_B}\left(P_{\hat{Z},Y}, P^{(0)}_{\hat{Z},Y}\right)$. Transfer learning provides a way to address the *curse of dimensionality* in learning bias by utilizing data from multiple sources, but it unexpectedly introduces transfer bias. Therefore, our objective is to balance learning bias and transfer bias while taking advantage of the properties of domain-invariant representations.

Building on the work of (Liu et al., 2021), we adopt the WGAN architecture. However, since it relies on minimax optimization, it is prone to instability during training. Alternating the training of $\hat{R}$ and $\hat{G}$ simultaneously tends to exacerbate this instability in practice. To enhance the stability of WGAN training, we split the process into two stages, as motivated by (Wang et al., 2024). In the first stage, the primary objective is to identify domain-invariant representations, with regularization applied to minimize distributional differences. In the second stage, using only these identified representations, we conduct a refined estimation.

Stage 1. To improve the stability of the discriminator's training, this stage uses $(\boldsymbol{x}^{(t)}_i, \boldsymbol{y}^{(t)}_i)$ as the input samples for the discriminator. We have $d_{\mathcal{F}^1_B}\left(P_{X,\hat{G}}, P_{X,Y}\right) \leq W_1\left(P_{X,\hat{G}}, P_{X,Y}\right)$, where W_1 is the 1-Wasserstein distance, the Kantorovich-Rubinstein theorem shows that the dual form of the 1-Wasserstein distance can be written as a form of integral probability metric(IPM) (Villani et al., 2009),

$$\begin{aligned} W_1\left(P_{X,\hat{G}}, P_{X,Y}\right) = \sup_{D \in \mathcal{F}^1_{\text{Lip}}} \{ & \mathbb{E}_{P_X P_\eta} D(X, G(\eta, R(X))) \\ & - \mathbb{E}_{P_{X,Y}} D(X, Y) \}, \end{aligned}$$

$$\mathcal{F}^1_{\text{Lip}} = \left\{ f : \mathbb{R}^{d+q} \to \mathbb{R}, \frac{|f(\boldsymbol{u}) - f(\boldsymbol{v})|}{\|\boldsymbol{u} - \boldsymbol{v}\|_2} \leq 1, \forall \boldsymbol{u}, \boldsymbol{v} \right\}.$$

Thus, finding the conditional generator and the representation can be formulated as a minimax problem,

$$\underset{G,R}{\operatorname{argmin}} \underset{D \in \mathcal{F}^1_{\text{Lip}}}{\operatorname{argmax}} \mathcal{L}_1(R, G, D; S),$$

which we incorporate regularization into the objective function, based on the original form of the 1-Wasserstein distance between the source domains and the target domain.

$$\begin{aligned} \mathcal{L}_1(R, G, D; S) = & \mathbb{E}_{P_X P_\eta} D(X, G(\eta, R(X))) \\ & - \mathbb{E}_{P_{X,Y}} D(X, Y) \\ & + \sum_{t \in S} \lambda_t \inf_\gamma \int \|(R(X^{(t)}), Y^{(t)}) - (R(X^{(0)}), Y^{(0)})\|_1 d\gamma, \end{aligned}$$

where λ_t represents weights for different source domains, $\gamma \in \Pi\left(P^{(t)}_{X,Y}, P^{(0)}_{X,Y}\right)$ describes the space of joint probability distributions with marginals $P^{(t)}_{X,Y}$ and $P^{(0)}_{X,Y}$. We avoid using the dual form for regularization because we do not want to introduce additional neural networks.

[1] For convenience, we denote that any distribution notation without the domain index (t) refers to a mixture distribution.

[2] We omit the argument $\hat{G}(\eta, \hat{Z})$ and refer to it as $\hat{G}$.

Let $\eta_i^{(t)}$ be independently generated from P_η. The empirical version of $\mathcal{L}_1(R, G, D; S)$ is

$$\widehat{\mathcal{L}}_1(R, G, D; S) = \frac{1}{n}\left[\sum_{\substack{t \in S \cup \{0\} \\ i=1}}^{n_t} D\left((\boldsymbol{x}_i^{(t)}, G\left(\eta_i, R\left(\boldsymbol{x}_i^{(t)}\right)\right)\right)\right.$$
$$\left. -D\left(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}\right)\right] + \sum_{t \in S} \lambda_t \min_{\gamma \in \Pi\left(P_{X,Y}^{n_t}, P_{X,Y}^{n_0}\right)} \left\langle \gamma, \mathbf{C}_R^{(t)} \right\rangle_F,$$

where $P_{X,Y}^{n_t}, P_{X,Y}^{n_0}$ are the empirical distributions and $\mathbf{C}_R^{(t)} = (\mathbf{C}_{R,ij}^{(t)})_{i,j=1}^{n_t,n_0}$ is a cost matrix $\in \mathbb{R}^{n_t \times n_0}$,

$$\mathbf{C}_{R,ij}^{(t)} = \|(R(\boldsymbol{x}_i^{(t)}), \boldsymbol{y}_i^{(t)}) - (R(\boldsymbol{x}_j^{(0)}), \boldsymbol{y}_j^{(0)})\|_1.$$

Although this avoids introducing additional neural networks, it still requires solving an optimization problem with γ. Efficient computational schemes have been proposed with stochastic versions using the dual formulation of the problem Genevay et al. (2016); Seguy et al. (2017), allowing for the tackling of small to medium-sized problems.

We use a feedforward neural network $G_{\boldsymbol{\theta}_1}$ with parameter $\boldsymbol{\theta}_1$ for estimating the conditional generator G in Stage 1, a second network $D_{\boldsymbol{\phi}_1}$ with parameter $\boldsymbol{\phi}_1$ for estimating the discriminator D in Stage 1 and a third network $R_{\boldsymbol{\omega}}$ with parameter $\boldsymbol{\omega}$ for estimating the representation R. We estimate $\boldsymbol{\theta}_1$, $\boldsymbol{\phi}_1$ and $\boldsymbol{\omega}$ by solving the minimax problem,

$$(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\phi}}_1) = \underset{\boldsymbol{\omega}, \boldsymbol{\theta}_1}{\operatorname{argmin}} \underset{\boldsymbol{\phi}_1}{\operatorname{argmax}} \widehat{\mathcal{L}}_1\left(R_{\boldsymbol{\omega}}, G_{\boldsymbol{\theta}_1}, D_{\boldsymbol{\phi}_1}; S\right).$$

The estimated representation is $\hat{R} = R_{\hat{\omega}}$ which will be used in Stage 2. At this stage, the estimated $G_{\boldsymbol{\theta}_1}$ and $D_{\boldsymbol{\phi}_1}$ are not the final results.

Stage 2. To further refine estimation, network retraining is conducted using the identified representation $\hat{R}$. This stage uses $(\hat{R}(\boldsymbol{x}_i^{(t)}), \boldsymbol{y}_i^{(t)})$ as the input samples for the discriminator which is different from the stage 1. We consider the minimax problem with no regularization:

$$\underset{G}{\operatorname{argmin}} \underset{D \in \mathcal{F}_{\text{Lip}}^1}{\operatorname{argmax}} \mathcal{L}_2(G, D; S),$$

where

$$\begin{aligned}\mathcal{L}_2(G, D; S) =& \mathbb{E}_{P_X P_\eta} D(\hat{R}(X), G(\eta, \hat{R}(X))) \\ &- \mathbb{E}_{P_{X,Y}} D(\hat{R}(X), Y).\end{aligned}$$

The empirical version of $\mathcal{L}_2(G, D; S)$ is

$$\widehat{\mathcal{L}}_2(G, D; S) = \frac{1}{n}\left[\sum_{\substack{t \in S \cup \{0\} \\ i=1}}^{n_t} D\left((\hat{R}(\boldsymbol{x}_i^{(t)}), G\left(\eta_i, \hat{R}\left(\boldsymbol{x}_i^{(t)}\right)\right)\right)\right.$$
$$\left. -D\left(\hat{R}(\boldsymbol{x}_i^{(t)}), \boldsymbol{y}_i^{(t)}\right)\right].$$

We use a feedforward neural network $G_{\boldsymbol{\theta}}$ with parameter $\boldsymbol{\theta}$ for estimating the conditional generator G, a second network $D_{\boldsymbol{\phi}}$ with parameter $\boldsymbol{\phi}$ for estimating the discriminator D. We estimate $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ by solving the minimax problem,

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \widehat{\mathcal{L}}_2\left(G_{\boldsymbol{\theta}}, D_{\boldsymbol{\phi}}; S\right).$$

The estimated conditional generator and discriminator are $\hat{G} = G_{\hat{\theta}}, \hat{D} = D_{\hat{\phi}}$. Due to space limitations, the detailed algorithm implementation and the **non-asymptotic error bound** can be found in the Appendix.

Remark. As illustrated in the Appendix, compared to the results in Liu et al. (2021), where the convergence rate without utilizing other source domains is $n_0^{-1/(d+q)}$, we improve this to $n^{-1/(r+q)} + n_0^{-1/r}$ by incorporating representation learning and leveraging information from the source domains when S is known. When both d and q are high-dimensional, $r \ll d$ signifies that the representation dimension is much smaller than the original data dimension, allowing for a more **efficient** convergence.

5 SELECTED TRANSFER-WGAN

In the previous section, we introduced an oracle algorithm based on a known subset S of reliable source domains. This leads to an intriguing and practically significant question: Can we develop a data-driven, adaptive selection criterion to estimate the subset $\hat{S}$?

Recall that our previous definition of S,

$$\forall t \in S, \mathbb{E}_{P_{\hat{Z}}^{(t)}} d_{\mathcal{F}_B^1}\left(P_{Y|\hat{Z}}^{(t)}, P_{Y|\hat{Z}}^{(0)}\right) \leq h.$$

Estimating the conditional distributions $P_{Y|\hat{Z}=\hat{\boldsymbol{z}}}^{(t)}$ and $P_{Y|\hat{Z}=\hat{\boldsymbol{z}}}^{(0)}$ can be challenging due to the differences in covariates $\boldsymbol{x}$. This often results in an insufficient number of samples for $\hat{Z} = \hat{\boldsymbol{z}}$, leading to significant biases compared to the ground truth. Therefore, using this distance directly for selection is not feasible.

To address this issue, we aim to utilize the joint distribution as a bridge. By doing so, we can constrain the 1-Wasserstein distance of the joint distribution instead of conditional distribution. This approach allows for a more feasible and practical method of selection. Then we can follow the oracle method mentioned in Section 4 by using $\hat{S}$ as a substitute for unknown S.

5.1 Selection Criterion

At first, we are unaware of which source domain is intrinsically similar to the target domain with the low-dimensional representation. Therefore, in the initial step, we should train a *full model* using all the source domains. We also observe that representation learning, compared to the subsequent conditional generative model learning, is less affected by outlier source domains. This phenomenon has also been studied and confirmed by Ortego et al. (2021). Thus, we can utilize the representation neural network $\tilde{R} = R_{\tilde{\boldsymbol{\omega}}}$ and representations $\tilde{\boldsymbol{z}}_i^{(t)} = \tilde{R}\left(\boldsymbol{x}_i^{(t)}\right)$ obtained from the full model to construct our selection criterion.

Stage 1. In this stage, our goal is to estimate $\hat{S}$ using the representation $\tilde{R}$ trained in the full model. To avoid confusion with the previous content, we present the training loss function of the full model here:

$$(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\omega}}) = \underset{\boldsymbol{\theta},\boldsymbol{\omega}}{\operatorname{argmin}} \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \widehat{\mathcal{L}}_1\left(R_{\boldsymbol{\omega}}, G_{\boldsymbol{\theta}}, D_{\boldsymbol{\phi}}; [T]\right).$$

Then, for some constant $C > 0$, we estimate $\hat{S}$ as:

$$\left\{t : W_1(P^{n_t}_{\tilde{Z},Y}, P^{n_0}_{\tilde{Z},Y}) \le C\left(\max\left\{n^{-1/(r+q)}, n_0^{-1/r}\right\}\right)\right\},$$

where $P^{n_t}_{\tilde{Z},Y}, P^{n_0}_{\tilde{Z},Y}$ are empirical distributions.

Stage 2. In this stage, our goal is to estimate $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\omega}}$ using $\hat{S}$. We simply replace S in the Oracle Transfer-WGAN with $\hat{S}$. The detailed algorithm implementation can be referenced in the Appendix.

5.2 Theoretical results

In this section, we summarize the key theoretical results. Due to space constraints, Assumptions 2-6 and the conditions of the theorems have been moved to the Appendix. Additionally, we choose $h = O\left(\max\left\{n^{-1/(r+q)}, n_0^{-1/r}\right\}\right)$. To clearly define a reliable source domain, we assume that the outlier source domains are sufficiently distant from the target domain.

Assumption 1. The similarity measure between the outlier sources and the target domain is assumed to be of a much larger order than h. Specifically, we assume

$$\forall t \in S^c, \mathbb{E}_{P^{(t)}_{\tilde{Z}}} d_{\mathcal{F}^1_B}\left(P^{(t)}_{Y|\tilde{Z}}, P^{(0)}_{Y|\tilde{Z}}\right) = O(h^\alpha), \alpha > 1.$$

From this, we can derive the following two theorems:

Theorem 5.1 *Suppose that $P_{\tilde{Z},Y}, P_{\hat{Z},Y}$ are supported on $[-U,U]^{r+q}$ for some $U > 0$ and satisfies Assumptions 1, 5-6 provided in Appendix, we have*

$$P(\hat{S} = S) \to 1, \text{when } n_t, n_0 \to +\infty.$$

Theorem 5.2 *Suppose that $P_{\tilde{Z},Y}, P_{\hat{Z},Y}$ are supported on $[-U,U]^{r+q}$ for some $U > 0$ and satisfies Assumptions 1, 3-6 provied in Appendix, we have*

$$\mathbb{E}_{\hat{G}}\mathbb{E}_{P^{(0)}_{\hat{Z}}} d_{\mathcal{F}^1_B}\left(P_{\hat{G}|\hat{Z}}, P^{(0)}_{Y|\hat{Z}}\right) \precsim n^{-1/(r+q)} + n_0^{-1/(r+q)}.$$

Remark 1. As a result, although the first term in the upper bound is dominated by the second term, and the efficiency improvement is only from $n_0^{-1/(d+q)}$ to $n_0^{-1/(r+q)}$, this is due to the fact that when S is unknown, we must also consider the bias introduced by the estimate $\hat{S}$ in the presence of limited samples. This represents a trade-off where **efficiency is sacrificed to ensure robustness**. Since $r \ll d$, there is still a significant improvement.

Remark 2. Since we only know the order of h, the corresponding constant remains unknown. In practical applications, we can adjust the constant after determining the order and use it as a threshold. Additionally, we can sort $W_1\left(P^{n_t}_{\tilde{Z},Y}, P^{n_0}_{\tilde{Z},Y}\right)$ for $t \in [T]$, where smaller values indicate more reliable source domains for transfer, allowing for a more **robust** selection.

6 EXPERIMENTS

In this section, we present the key settings and results of three different experiments, with additional details provided in the appendix.

6.1 Numerical simulation

We focus on the problem of estimating the conditional mean and standard deviation in nonparametric conditional density models. Since our approach is the first to eliminate the need for a pretrained model, and no pretrained models are available for this task, we compare the proposed Selected Transfer-WGAN method (referred to as **STWGAN** in Table 1) with two baselines: **Target-Only**, a method trained exclusively on the target domain without representation learning, and **Pool**, an ablation variant where $\lambda_t = 0$. We have placed additional method comparisons and experimental results in the Appendix. We simulated data from the following three models:

Model 1 (M1). A nonlinear model:

$$Y = X_1 + \exp\left(X_2 + X_3/3\right) + \sin\left(X_4 + X_5\right) + \varepsilon,$$

where $\varepsilon \sim N(0, X_1^2)$.

Model 2 (M2). A model with a multiplicative error:

$$Y = (2 + X_1^2/3 + X_2^2 + X_3^2 + X_4^2 + X_5^2)/3 \times \varepsilon,$$

where $\varepsilon \sim N(X_3, 1)$.

Model 3 (M3). A mixture model:

$$\begin{aligned} Y =& \mathbb{I}_{\{U \le 1/3\}} N\left(-3 - X_1/3 - X_2^2, 0.25\right) \\ &+ \mathbb{I}_{\{U > 1/3\}} N\left(3 + X_1/3 + X_2^2, 1\right), \end{aligned}$$

where $U \sim \text{Unif}(0,1)$ and is independent of X.

In each model, the covariate vector X is generated from $N(\boldsymbol{\mu}^{(t)}, \mathbf{I}_{100})$ in the t-th domain. So the ambient dimension of X is 100, but (M1) and (M2) only depend on the first 5 components of X and (M3) only depends on the first 2 components of X. To further demonstrate the robustness and efficiency of our approach, we consider 5 different source domains, with the corresponding values of $\boldsymbol{\mu}^{(t)}$ provided in appendix. Additionally, to demonstrate the impact of different outlier source domains, we introduce posterior drift in the fourth and fifth source domain.

Similar to the experiments conducted by Liu et al. (2021); Zhou et al. (2023), we consider the mean squared error (MSE) of the estimated conditional mean $\mathbb{E}(Y \mid X)$ and the estimated conditional standard deviation $\mathrm{SD}(Y \mid X)$. We use a test data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$ of size $K = 2000$. For the proposed method, we first generate $J = 10000$ samples $\{\eta_j : j = 1, \ldots, J\}$ from the reference distribution P_η and calculate conditional samples $\left\{\hat{G}\left(\eta_j, \boldsymbol{x}_k\right), j = 1, \ldots, J, k = 1, \ldots, K\right\}$ The estimated conditional standard deviation is calculated as the sample standard deviation of the conditional samples. The MSE of the estimated conditional mean is $\mathrm{MSE(mean)} = (1/K)\sum_{k=1}^{K}\{\widehat{\mathbb{E}}\left(Y \mid X = \boldsymbol{x}_k\right) - \mathbb{E}\left(Y \mid X = \boldsymbol{x}_k\right)\}^2$. Similarly, the MSE of the estimated conditional standard deviation is $\mathrm{MSE(sd)} = (1/K)\sum_{k=1}^{K}\{\widehat{\mathrm{SD}}(Y \mid X = \boldsymbol{x}_k) - \mathrm{SD}(Y \mid X = \boldsymbol{x}_k)\}^2$.

Based on Figure 2, in all three data simulated models, the first source domain is considered a reliable source domain, while the others are identified as outlier source domains. The MSE(mean) and MSE(sd) are summarized in Table 1. Comparing with the models trained with only target domain, **STWGAN** has the smallest MSEs error in most cases.

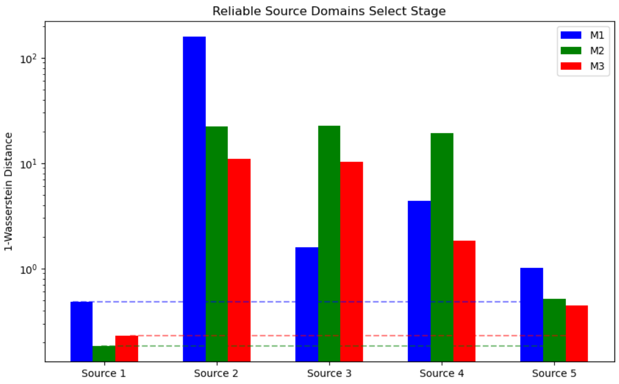


Figure 2: A simple visualization of STWGAN Stage 1

Table 1: Mean squared error (MSE) of the estimated conditional mean, the estimated standard deviation and the corresponding simulation standard errors (in parentheses). The experimental results are presented in three parts, arranged from top to bottom, corresponding to $n_t = 20000, 40000, 60000$ and $n_0 = 10000$. **Our complete experimental results are provided in the Appendix.**

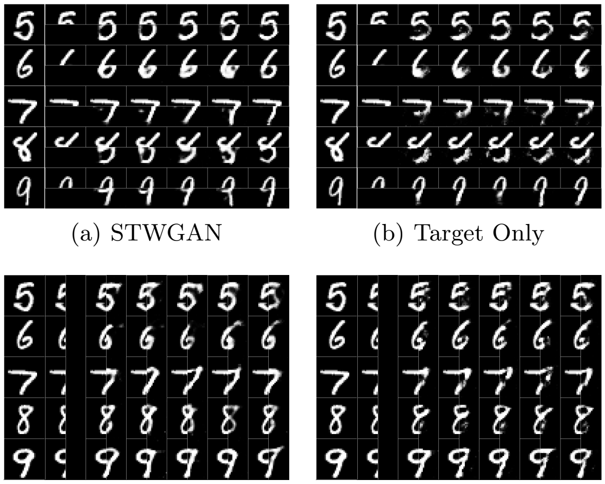
		STWGAN	Target-only	Pool
M1	Mean	**15.77**(1.29)	21.49(1.24)	16.90(1.63)
	SD	4.43(1.48)	8.21(2.84)	**1.89**(0.45)
M2	Mean	**4.40**(1.10)	9.51(3.63)	6.75(2.35)
	SD	1.95(0.30)	**1.39**(0.14)	1.84(0.18)
M3	Mean	**2.22**(0.99)	25.75(4.10)	3.07(1.42)
	SD	**0.47**(0.10)	10.14(5.20)	0.75(0.10)
M1	Mean	**10.64**(2.07)	17.06(1.91)	16.94(2.94)
	SD	6.69(4.47)	7.69(3.33)	**1.37**(0.22)
M2	Mean	**3.12**(1.14)	7.10(3.01)	5.15(1.77)
	SD	1.90(0.38)	**1.53**(0.23)	2.10(0.34)
M3	Mean	**2.09**(1.39)	26.89(7.13)	2.32(1.55)
	SD	**0.54**(0.13)	7.72(4.06)	0.61(0.51)
M1	Mean	**10.73**(1.16)	24.40(2.84)	17.56(1.69)
	SD	2.84(1.59)	9.84(2.56)	**1.61**(0.56)
M2	Mean	**2.22**(1.30)	7.73(4.01)	6.96(1.42)
	SD	2.37(1.16)	**1.51**(0.20)	2.05(0.13)
M3	Mean	**1.68**(1.34)	20.97(3.40)	2.32(1.55)
	SD	**0.56**(0.06)	5.67(2.67)	0.61(0.51)

6.2 Image reconstruction: MNIST dataset

We now illustrate the application of the proposed method to high-dimensional data problems. We use the MNIST handwritten digits dataset(Deng, 2012), which contains 60000 images for training and 10000 images for testing. The images are stored in 28 × 28 matrices with gray color intensity from 0 to 1. We use STWGAN to help reconstruct the missing part of an image. Specifically, we consider a scenario in which only the upper or left half of an image is observed, and the task is to reconstruct the missing part. In this setting, let $X \in \mathbb{R}^{28\times 14}$ represent the observed upper or left half of the image, while $Y \in \mathbb{R}^{28\times 14}$ denotes the missing part. We refer to the two experiments corresponding to different missing parts as ”upper2lower” and ”left2right”.

In practice, we construct the target domain and a source domain directly within the MNIST dataset. We select 5,000 images of digits 5-9 from the training set to serve as the target domain, and 50,000 images of digits 0-9 as the source domain. The experimental results are illustrated in Figure 3.

In Figure 3, the first column of each panel displays the true images, the second column shows the generator



(a) STWGAN (b) Target Only

(c) STWGAN (d) Target Only

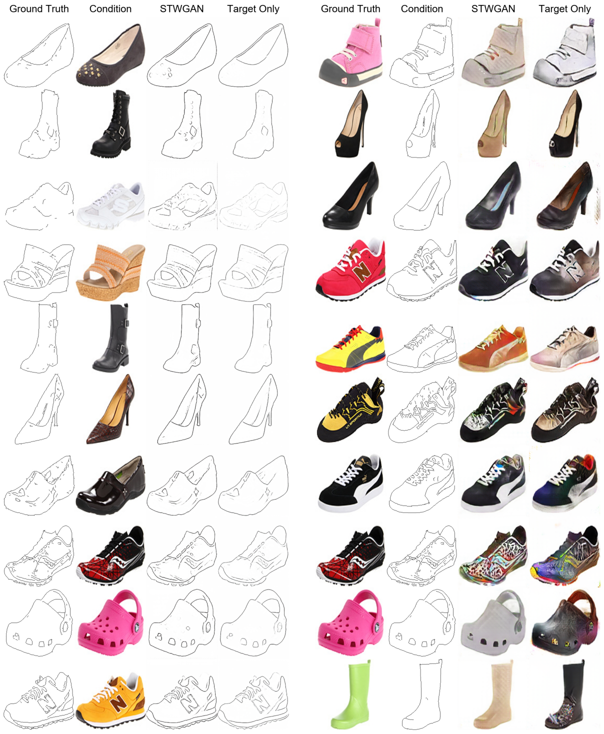
Figure 3: Comparison of STWGAN and Target Only: (a) and (b) show results of upper2lower, while figures (c) and (d) show results of left2right.

condition X, and the remaining columns present the generated images. Digits “6”, “7”, and “9” are reconstructed effectively with the aid of the source domain. However, when only the upper part of the digits is provided, digit “8” may be misinterpreted as “6” due to its similar structure, resulting in poorer reconstruction quality. Nonetheless, in terms of stroke consistency, the images generated after transfer exhibit greater realism.

6.3 Image-to-Image translation

We demonstrate the application of the proposed method to the task of image-to-image translation using the edges2shoes and edges2handbags datasets (Isola et al., 2017). The edges2shoes dataset includes over 40000 training images derived from the UT Zappos50K dataset (Yu and Grauman, 2017), while the edges2handbags dataset comprises more than 130000 training images from the iGAN project (Zhu et al., 2016). Each dataset consists of a real image of shoes or handbags paired with a corresponding edge map of the object, where the edges were generated using the HED edge detector (Xie and Tu, 2015). In both datasets, the edge maps and real images are stored as tensors with dimensions $1\times286\times286$ and $3\times286\times286$, respectively. Due to the smaller sample size in the edges2shoes dataset, we selected 40000 samples from it to serve as the target domain and selected 120000 samples from the edges2handbags dataset as the source domain.

We then conduct two sets of experiments. In the first experiment, we use the edge map as $Y \in \mathbb{R}^{81,796}$ and the real image as $X \in \mathbb{R}^{254,388}$, with the results shown in the figure (4.a). In the second experiment, we reverse the X and Y with the results also presented in the figure (4.b). It can be observed that the STWGAN method effectively transfers knowledge of complex patterns from the edges2handbags dataset, resulting in more accurate edge representations on shoes and enhancing the richness of patterns in the generated shoe images. For example, in the case of sneakers with intricate edge details, our method often produces brighter, more vibrant, and realistic images, while other approaches tend to generate duller, grayish color patterns, lacking in vibrancy and detail.



(a) shoes2edges (b) edges2shoes

Figure 4: Comparison of STWGAN and Target Only.

7 CONCLUSIONS

We proposed *Selected Transfer-WGAN* (STWGAN), a robust transfer approach designed to address the challenges of multi-source conditional generation models. This is achieved through a two-stage training process that maintains the training stability of WGAN. Our algorithm does not rely on pre-trained models from large datasets and provides both non-asymptotic error bounds and asymptotic guarantees. Future work will discuss how neural networks can learn complex dimensionality reduction structures and retain useful information.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984.

Cai, T. T. and Pu, H. (2024). Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*.

Chen, M., Liao, W., Zha, H., and Zhao, T. (2020). Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30.

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. (2016). Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al. (2021). Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Harun, M. Y., Lee, K., Gallardo, J., Krishnan, G., and Kanan, C. (2024). What variables affect out-of-distribution generalization in pretrained models? *arXiv preprint arXiv:2405.15018*.

He, B., Liu, H., Zhang, X., and Huang, J. (2024). Representation transfer learning for semiparametric regression. *arXiv preprint arXiv:2406.13197*.

Hoffman, J., Rodner, E., Donahue, J., Kulis, B., and Saenko, K. (2014). Asymmetric and category invariant feature transformations for domain adaptation. *International journal of computer vision*, 109:28–41.

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., and Yang, Y. (2022). An error analysis of generative adversarial networks for learning distributions. *Journal of machine learning research*, 23(116):1–43.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Kallenberg, O. (1997). *Foundations of modern probability*, volume 2. Springer.

Kandemir, M. (2015). Asymmetric transfer learning with deep gaussian processes. In *International Conference on Machine Learning*, pages 730–738. PMLR.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC.

Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173.

Li, Y., Zhang, R., Lu, J., and Shechtman, E. (2020). Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*.

Liang, T. (2021). How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41.

Lin, H. and Reimherr, M. (2022). Transfer learning for functional linear regression with structural interpretability. *arXiv preprint arXiv:2206.04277*.

Liu, S., Zhou, X., Jiao, Y., and Huang, J. (2021). Wasserstein generative learning of conditional distribution. *arXiv preprint arXiv:2112.10039*.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. (2014). Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1417.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Ortego, D., Arazo, E., Albert, P., O'Connor, N. E., and McGuinness, K. (2021). Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2017). Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Song, S., Wang, T., Shen, G., Lin, Y., and Huang, J. (2023). Wasserstein generative regression. *arXiv preprint arXiv:2306.15163*.

Suh, N. and Cheng, G. (2024). A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *arXiv preprint arXiv:2401.07187*.

Tan, Z., Zhou, L., and Lin, H. (2024). Generative adversarial learning with optimal input dimension and its adaptive generator architecture. *arXiv preprint arXiv:2405.03723*.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697.

Tian, Y., Gu, Y., and Feng, Y. (2023). Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, T., Huang, J., and Ma, S. (2024). Penalized generative variable selection. *arXiv preprint arXiv:2402.16661*.

Wang, Y., Gonzalez-Garcia, A., Berga, D., Herranz, L., Khan, F. S., and Weijer, J. v. d. (2020). Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341.

Wang, Y., Wu, C., Herranz, L., Van de Weijer, J., Gonzalez-Garcia, A., and Raducanu, B. (2018). Transferring gans: generating images from limited data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 218–234.

Xie, S. and Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403.

Yu, A. and Grauman, K. (2017). Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579.

Zhao, Y., Ding, H., Huang, H., and Cheung, N.-M. (2022). A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9140–9150.

Zhou, D., Liu, M., Li, M., and Cai, T. (2024). Doubly robust augmented model accuracy transfer inference with high dimensional features. *Journal of the American Statistical Association*, (just-accepted):1–26.

Zhou, X., Jiao, Y., Liu, J., and Huang, J. (2023). A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543):1837–1848.

Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   See Appendix for full description.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   We have included the complete assumptions and theoretical results in the Appendix.
   (b) Complete proofs of all theoretical results. [Yes]
   We provide proofs for all the proposed theoretical results in the Appendix.
   (c) Clear explanations of any assumptions. [Yes]
   For Assumptions 1-6, we provide clear explanations and analyses in the Appendix.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   We will open source the complete code after acceptance.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   Most training details are included in the Appendix, while the values of some commonly used hyperparameters that are not mentioned can be found in the code.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
   Our description on this aspect is included in Section 1.1 of the Appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]