Algorithmic Fairness Generalization under Covariate and Dependence Shifts Simultaneously

Chen Zhao*
Baylor University
Waco, Texas, USA
chen_zhao@baylor.edu

Kai Jiang*
The University of Texas at Dallas
Richardson, Texas, USA
kai.jiang@utdallas.edu

Xintao Wu
University of Arkansas
Fayetteville, Arkansas, USA
xintaowu@uark.edu

Haoliang Wang
The University of Texas at Dallas
Richardson, Texas, USA
haoliang.wang@utdallas.edu

Latifur Khan
The University of Texas at Dallas
Richardson, Texas, USA
lkhan@utdallas.edu

Christan Grant
University of Florida
Gainesville, Florida, USA
christan@ufl.edu

Feng Chen
The University of Texas at Dallas
Richardson, Texas, USA
feng.chen@utdallas.edu

ABSTRACT

The endeavor to preserve the generalization of a fair and invariant classifier across domains, especially in the presence of distribution shifts, becomes a significant and intricate challenge in machine learning. In response to this challenge, numerous effective algorithms have been developed with a focus on addressing the problem of fairness-aware domain generalization. These algorithms are designed to navigate various types of distribution shifts, with a particular emphasis on covariate and dependence shifts. In this context, covariate shift pertains to changes in the marginal distribution of input features, while dependence shift involves alterations in the joint distribution of the label variable and sensitive attributes. In this paper, we introduce a simple but effective approach that aims to learn a fair and invariant classifier by simultaneously addressing both covariate and dependence shifts across domains. We assert the existence of an underlying transformation model can transform data from one domain to another, while preserving the semantics related to non-sensitive attributes and classes. By augmenting various synthetic data domains through the model, we learn a fair and invariant classifier in source domains. This classifier can then be generalized to unknown target domains, maintaining both model prediction and fairness concerns. Extensive empirical studies on four benchmark datasets demonstrate that our approach surpasses state-of-the-art methods. The code repository is available at https://github.com/jk-kaijiang/FDDG.

*Both authors contributed equally to this research.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671909

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Applied computing** → *Law, social and behavioral sciences*; • **Social and professional topics** → User characteristics.

KEYWORDS

Fairness, Generalization, Distribution Shifts

ACM Reference Format:
Chen Zhao, Kai Jiang, Xintao Wu, Haoliang Wang, Latifur Khan, Christan Grant, and Feng Chen. 2024. Algorithmic Fairness Generalization under Covariate and Dependence Shifts Simultaneously. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671909

1 INTRODUCTION

While modern fairness-aware machine learning techniques have demonstrated significant success in various applications [31, 54, 56, 58–64], their primary objective is to facilitate equitable decision-making, ensuring algorithmic fairness across all demographic groups characterized by sensitive attributes, such as race and gender. Nevertheless, the generalization of a fair classifier learned in the source domain to a target domain during inference often demonstrates severe limitations in many state-of-the-art methods. The poor generalization can be attributed to the data distribution shifts from source to target domains, resulting in catastrophic failures.

There are two main lines of data distribution shifts [41]: general and fairness-specific shifts. The former focuses on shifts involving input features and labels. Specifically, covariate shift [45] and label shift [52] refer to variations due to different marginal distributions over feature and class variables, respectively. Concept shift [53] indicates "functional relation change" due to the change amongst the instance-conditional distributions [40]. Moreover, fairness-specific shifts consider additional sensitive attributes and hence place a greater emphasis on ensuring algorithmic fairness. Demographic

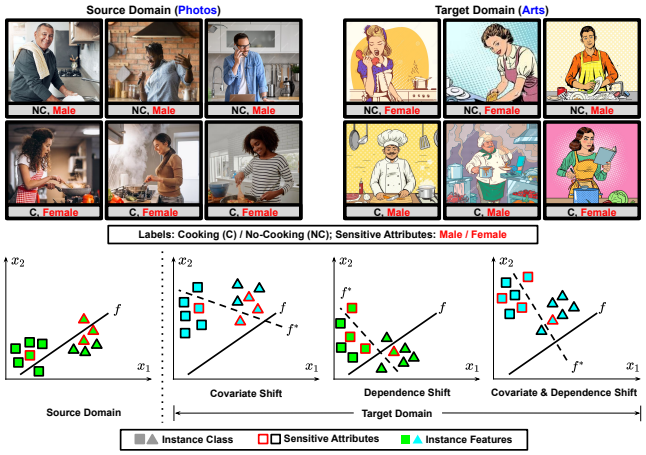

Figure 1: Illustration of the problem in generalizing fair classifiers across different data domains under covariate and dependence shifts simultaneously. (Upper) Images in source and target domains have different styles (Photos and Arts). Each data domain is linked to a distinct correlation between class labels (NC and C) and sensitive attributes (Male and Female). (Lower) We consider $\mathbf{x} = [x_1, x_2]^T$ a simple example of a two-dimensional feature vector. A fair classifier f learned using source data is applied to data sampled from various types of shifted target domains, resulting in misclassification and unfairness. f^* represents the true classifier in the target domain.

shift[1] [15] refers to certain sensitive population subgroups becoming more or less probable during inference. Dependence shift [41] captures the correlation change between the class variable and sensitive attributes. Within these distribution shifts, a trained fair classifier from source domains is directly influenced and may degrade when adapted to target domains.

To simplify, we narrow the scope of distribution shifts to two prominent ones: covariate shift, which has been extensively investigated in the context of out-of-distribution (OOD) generalization [40, 57], and dependence shift, a topic that has gained attention in recent research. In the illustrative example shown in Fig. 1, the source and target domains exhibit variations stemming from different image styles (Photos and Arts) and correlations between labels (No-cooking and Cooking) and sensitive attributes (Male and Female). Specifically, in the source domain, most males in the kitchen are not cooking, whereas in the target domain, a distinct correlation is observed with most males engaging in cooking. To learn a classifier that is both fair and accurate under such hybrid shifts, a variety of domain generalization approaches have been explored. Predominantly, these methods often exhibit two specific limitations: they (1) address either covariate shift [26, 40, 57] or dependence shift [8, 36], or (2) solely focus on covariate shift but not explicitly indicate the existence of dependence shift [37]. Therefore, there is a need for research that explores the problem of *fairness-aware domain generalization* (FDG), considering both covariate and dependence shifts simultaneously across source and target domains.

[1]Dependence shift is named as correlation shift in [15].

In this paper, we introduce a novel framework, namely *Fair disEntangled DOmain geneRAlization* (FEDORA). The key idea in our framework revolves around learning a fair and accurate classifier that can generalize from given source domains to target domains, which remain unknown and inaccessible during training. The variations in these domains result from the concurrent presence of covariate and dependence shifts. Notice that, unlike the settings in some works involving covariate shift [32, 38, 48], we assert each domain possesses a distinct data style (Photos and Arts), resulting in an alternation in feature spaces. Technically, we assert the existence of a transformation model that can disentangle input data to a semantic factor that remains invariant across domains, a style factor that characterizes covariate-related information, and a sensitive factor that captures attributes of a sensitive nature. To enhance the generalization of the training classifier and adapt it to unknown target domains, we augment the data by generating them through the transformation model. It utilizes semantic factors associated with various style and sensitive factors sampled from their respective prior distributions. Furthermore, we leverage this framework to systematically define the FDG problem as a semi-infinite constrained optimization problem. Theoretically, we apply this re-formulation to demonstrate that a tight approximation of the problem can be achieved by solving the empirical, parameterized dual for this problem. Moreover, we develop a novel interpretable bound focusing on fairness within a target domain, considering the domain generalization arising from both covariate and dependence shifts. Finally, extensive experimental results on the proposed new algorithm show that our algorithm significantly outperforms state-of-the-art baselines on several benchmarks. Our main contributions are summarized.

- We introduce a fairness-aware domain generalization problem within a framework that accommodates inter-domain variations arising from covariate and dependence shifts simultaneously. We also give a brief survey by comparing the setting of related works.
- We reformulate the problem to a novel constrained learning problem. We further establish duality gap bounds for the empirically parameterized dual of this problem and develop a novel upper bound that specifically addresses fairness within a target domain while accounting for the domain generalization stemming from both covariate and dependence shifts.
- We present a novel algorithm, FEDORA, that enforces invariance across unseen target domains by utilizing generative models derived from the observed source domains.
- Comprehensive experiments are conducted to verify the effectiveness of FEDORA. We empirically show that it significantly outperforms state-of-the-art baselines on four benchmarks.

2 RELATED WORKS

Domain generalization. Addressing the challenge of domain shift and the absence of OOD data has led to the introduction of several state-of-the-art methods in the domain generalization field [3, 40, 50, 57]. These methods are designed to enable deep learning models to possess intrinsic generalizability, allowing them to adapt effectively from one or multiple source domains to target domains characterized by unknown distributions [51]. They encompass various techniques, such as aligning source domain distributions to

Table 1: Different Types of Distribution Shifts.

Type of Shifts	Notations, $\forall s \in \mathcal{E}_s$
Covariate Shift (Cov.) [45]	$\mathbb{P}^s_X \neq \mathbb{P}^t_X$
Label Shift (Lab.) [52]	$\mathbb{P}^s_Y \neq \mathbb{P}^t_Y$
Concept Shift (Con.) [53]	$\mathbb{P}^s_{Y\|X} \neq \mathbb{P}^t_{Y\|X}$
Demographic Shift (Dem.) [15]	$\mathbb{P}^s_Z \neq \mathbb{P}^t_Z$
Dependence Shift (Dep.) [41]	$\mathbb{P}^s_{Y\|Z} \neq \mathbb{P}^t_{Y\|Z}$ and $\mathbb{P}^s_Z = \mathbb{P}^t_Z$; or, $\mathbb{P}^s_{Z\|Y} \neq \mathbb{P}^t_{Z\|Y}$ and $\mathbb{P}^s_Y = \mathbb{P}^t_Y$
Hybrid Shift	Any combination of the shifts above.

Table 2: An overview of different settings of existing approaches in mitigating unfairness under distribution shifts.

Refs.	Distribution Shifts					Spaces Change*, $\forall s \in \mathcal{E}^s$			$\|\mathcal{E}^s\|$	Access to Target
	Cov.	Lab.	Con.	Dem.	Dep.	$\mathcal{X}^s \neq \mathcal{X}^t$	$\mathcal{Y}^s \neq \mathcal{Y}^t$	$\mathcal{Z}^s \neq \mathcal{Z}^t$		
[32, 38, 48]	•								1	No
[9]	•								M	No
[10, 39]	•								1	Yes
[37]	•					•			M	No
[4]		•							1	Yes
[20, 21]			•						1	Yes
[15, 44]				•				•	1	Yes
[8]					•				M	No
[36]					•				1	No
[41]					•				1	Yes
[22]	•		•						1	Yes
[46]	•			•					1	Yes
[43]	•	•		•					1	Yes
[18]	•	•					•		1	No
[7]	•	•				•			1	Yes
FEDORA	•				•	•			M	No

* $\mathcal{Y}^s \neq \mathcal{Y}^t$ and $\mathcal{Z}^s \neq \mathcal{Z}^t$ indicate the introduction of new labels and new sensitive attributes. A change in $\mathcal{X}$ denotes a shift in feature variation, such as transitioning from photo images to arts.

facilitate domain-invariant representation learning [29], subjecting the model to domain shift during training through meta-learning [28], and augmenting data with domain analysis, among others [65], and so on. In the context of the number of source domains, a significant portion of research [5, 40, 57] has focused on the multi-source setting. This setting assumes the availability of multiple distinct but relevant domains for the generalization task. As mentioned in [5], the primary motivation for studying domain generalization is to harness data from multiple sources in order to unveil stable patterns. This entails learning representations invariant to the marginal distributions of data features, all while lacking access to the target data. Nevertheless, existing domain generalization methods tend to overlook the aspect of learning with fairness, where group fairness dependence patterns may not change domains.

Fairness learning for changing environments. Two primary research directions aim to tackle fairness-aware machine learning in dynamic or changing environments. The first approach involves equality-aware monitoring methods [1, 7, 15, 24, 37, 39, 46], which strive to identify and mitigate unfairness in a model's behavior by continuously monitoring its predictions. These methods adapt the model's parameters or structure when unfairness is detected. However, a significant limitation of such approaches is their assumption of invariant fairness levels across domains, which may not hold in real-world applications. The second approach [8, 36] focuses on assessing a model's fairness in a dynamic environment exclusively under dependence shifts. However, it does not consider other types of distribution shifts.

In response to these limitations, this paper adopts a novel approach by attributing the distribution shift from source to target domains to both covariate shift and fairness dependence shift simultaneously. The objective is to train a fairness-aware invariant classifier capable of effective generalization across domains, ensuring robust performance in terms of both model accuracy and the preservation of fair dependence between predicted outcomes and sensitive attributes under both shifts.

3 PRELIMINARIES

Notations. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote a feature space, $\mathcal{Z} = \{-1, 1\}$ is a sensitive space, and $\mathcal{Y} = \{0, 1\}$ is a label space for classification. Let $\mathcal{C} \subseteq \mathbb{R}^c$, $\mathcal{A} \subseteq \mathbb{R}^a$, and $\mathcal{S} \subseteq \mathbb{R}^s$ be the semantic, sensitive and style latent spaces, respectively, induced from $\mathcal{X}$ and $\mathcal{A}$ by an underlying transformation model $T : \mathcal{X}\times\mathcal{Z}\times\mathcal{E} \to \mathcal{X}\times\mathcal{Z}$. We use X, Z, Y, C, A, S to denote random variables that take values in $\mathcal{X}, \mathcal{Z}, \mathcal{Y}, \mathcal{C}, \mathcal{A}, \mathcal{S}$ and $\mathbf{x}, z, y, \mathbf{c}, \mathbf{a}, \mathbf{s}$ the realizations. A domain $e \in \mathcal{E}$ is defined as a joint distribution $\mathbb{P}^e_{XZY} = \mathbb{P}(X^e, Z^e, Y^e) : \mathcal{X} \times \mathcal{Z} \times \mathcal{Y} \to [0, 1]$. A classifier f in a class space $\mathcal{F}$ denotes $f \in \mathcal{F} : \mathcal{X} \to \mathcal{Y}$. We denote $\mathcal{E}$ and $\mathcal{E}^s \subset \mathcal{E}$ as the set of domain labels for all domains and source domains, respectively. Superscripts in the samples denote their domain labels, while subscripts specify the indices of encoders. For example, $E_s(\mathbf{x}^s)$ denotes a sample $\mathbf{x}$ drawn from the s domain and encoded by a style encoder E_s.

Fairness notions. When learning a fair classifier $f \in \mathcal{F}$ that focuses on statistical parity across different sensitive subgroups, the fairness criteria require the independence between the sensitive random variables Z and the predicted model outcome $f(X)$ [11]. Addressing the issue of preventing group unfairness can be framed as the formulation of a constraint. This constraint mitigates bias by ensuring that $f(X)$ aligns with the ground truth Y, fostering equitable outcomes.

Definition 1 (Group Fairness Notion [35, 54]). *Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, z_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ sampled i.i.d. from $\mathbb{P}_{XZY}$, a classifier $f \in \mathcal{F} : \mathcal{X} \to \mathcal{Y}$ is fair when the prediction $\hat{Y} = f(X)$ is independent of the sensitive random variable Z. To get rid of the indicator function and relax the exact values, a linear approximated form of the difference between sensitive subgroups is defined as*

$$\rho(\hat{Y}, Z) = \left|\mathbb{E}_{\mathbb{P}_{XZY}} g(\hat{Y}, Z)\right|, \;\; g(\hat{Y}, Z) = \frac{1}{p_1(1-p_1)}\left(\frac{Z+1}{2} - p_1\right)\hat{Y} \quad (1)$$

p_1 and $1 - p_1$ are the proportion of samples in the subgroup $Z = 1$ and $Z = -1$, respectively.

Specifically, when $p_1 = \mathbb{P}(Z = 1)$ and $p_1 = \mathbb{P}(Z = 1, Y = 1)$, the fairness notion $\rho(\hat{Y}, Z)$ is defined as the difference of demographic parity and the difference of equalized opportunity, respectively [35]. In this paper, we will present the results under demographic parity (and then the expectation in Eq. (1) is over XZ), while the framework can be generalized to multi-class, multi-sensitive attributes and other fairness notions. Strictly speaking, a classifier f is fair over subgroups if it satisfies $\rho(\hat{Y}, Z) = 0$.

Problem setting. Given a dataset $\mathcal{D} = \{\mathcal{D}^e\}_{e=1}^{|\mathcal{E}|}$, where each $\mathcal{D}^e = \{(\mathbf{x}^e_i, z^e_i, y^e_i)\}_{i=1}^{|\mathcal{D}^e|}$ is *i.i.d.* sampled from a domain $\mathbb{P}^e_{XZY}$ and $e \in \mathcal{E}$, we consider multiple source domains $\{\mathbb{P}^s_{XZY}\}_{s=1}^{|\mathcal{E}^s|}$ and a distinct target domain $\mathbb{P}^t_{XZY}, t \neq s, \forall s \in \mathcal{E}^s \subset \mathcal{E}$ and $t \in \mathcal{E}\backslash\mathcal{E}^s$, which is unknown and inaccessible during training. Given samples

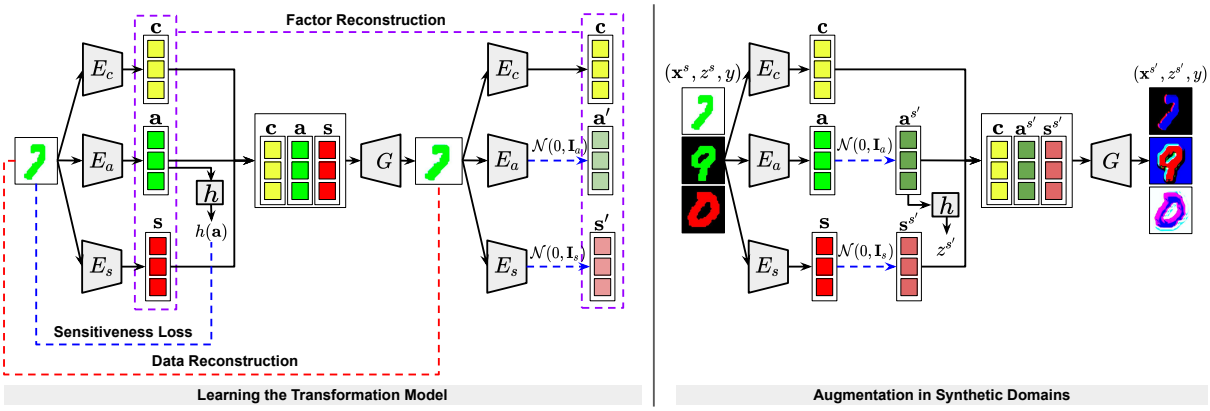

Figure 2: (Left) A transformation model T is trained using a bi-directional reconstruction loss (data reconstruction and factor reconstruction) and a sensitiveness loss. (Right) To enhance the generalization of the classifier f to unseen target domains, the transformation model T is used for augmentation in synthetic domains by generating data based on invariant semantic factors and randomly sampled sensitive and style factors that encode synthetic domains. We demonstrate the concept using the ccMNIST dataset, where the domains are distinguished by different digit colors and fair dependencies between class labels and sensitive attributes. Here, sensitive attributes are defined by image background colors.

$\{\mathcal{D}^s\}_{s=1}^{|\mathcal{E}^s|}$ from finite source domains, the goal of fairness-aware domain generalization problems is to learn a classifier $f \in \mathcal{F}$ that is generalizable across all possible domains.

Problem 1 (Fairness-aware Domain Generalization). *Let $\{\mathbb{P}^s_{XZY}\}_{s=1}^{|\mathcal{E}^s|}$ be a finite subset of source domains and assume that, for each $s \in \mathcal{E}^s$, we have access to its corresponding dataset $\mathcal{D}^s = \{(\mathbf{x}_i^s, z_i^s, y_i^s)\}_{i=1}^{|\mathcal{D}^s|}$ sampled i.i.d from $\mathbb{P}^s_{XZY}$. Given a classifier set $\mathcal{F}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the goal is to learn a fair classifier $f \in \mathcal{F}$ for any $\mathcal{D}^s$ that minimizes the worst-case risk over all domains in $\{\mathbb{P}^e_{XZY}\}_{e=1}^{|\mathcal{E}|}$ satisfying a group fairness constraint:*

$$\min_{f\in\mathcal{F}} \max_{e\in\mathcal{E}} \mathbb{E}_{\mathbb{P}^s_{XZY}} \ell(f(X^s), Y^s), \quad s.t.\ \rho(f(X^s), Z^s) = 0 \tag{2}$$

The goal of Prob. 1 is to seek a fair classifier f that generalizes from the given finite set of source domains to give a good generalization performance on all domains. Since we do not assume data from a target domain is accessible, it makes Prob. 1 challenging to solve.

Another challenge is how closely the data distributions in unknown target domains match those in the observed source domains. As discussed in Sec. 1 and Tab. 1, there are five different types of distribution shifts. In this paper, we narrow the scope and claim the shift between source and target domains is solely due to covariate and dependence shifts.

Definition 2 (Covariate Shift [40] and Dependence Shift[41]). *In Prob. 1, covariate shift occurs when domain variation is attributed to disparities in the marginal distributions over input features $\mathbb{P}^s_X \neq \mathbb{P}^t_X, \forall s$. On the other hand, Prob. 1 exhibits a dependence shift when domain variation arises from alterations in the joint distribution between Y and Z, denoted $\mathbb{P}^s_{YZ} \neq \mathbb{P}^t_{YZ}, \forall s$ where $\mathbb{P}^s_{Y|Z} \neq \mathbb{P}^t_{Y|Z}$ and $\mathbb{P}^s_Z = \mathbb{P}^t_Z$; or $\mathbb{P}^s_{Z|Y} \neq \mathbb{P}^t_{Z|Y}$ and $\mathbb{P}^s_Y = \mathbb{P}^t_Y$.*

Underlying transformation models. Inspired by existing domain generalization endeavors [19, 40, 57], distribution shifts can characterize generalization tasks across domains through an underlying transformation model T. The motivation behind using T lies in bolstering the robustness and adaptability of the classifier f across diverse domains. By learning a transformation model, the objective is twofold: (1) to enable the model to adapt domain-invariant data representations (factors) from the input data by disentangling domain-specific variations and (2) to generate augmented data in new domains by perturbing existing samples with various variations. This augmentation enhances the diversity of the source data and thereby improves the ability to generalize to unseen target domains.

4 METHODOLOGY

4.1 Learning the Transformation Model

One goal of the transformation model $T = \{E, G\}$ is to disentangle an input sample from source domains into three factors in latent spaces by learning a set of encoder $E = \{E_c, E_a, E_s\}$ and a decoder $G : \mathcal{C} \times \mathcal{A} \times \mathcal{S} \to \mathcal{X}$, where $E_c : \mathcal{X} \to \mathcal{C}$, $E_a : \mathcal{X} \to \mathcal{A}$, and $E_s : \mathcal{X} \to \mathcal{S}$ represent semantic, sensitive and style encoders, respectively.

Assumption 1 (Multiple Latent Factors). *Given dataset $\mathcal{D}^e = \{(\mathbf{x}_i^e, z_i^e, y_i^e)\}_{i=1}^{|\mathcal{D}^e|}$ sampled i.i.d. from $\mathbb{P}^e_{XZY}$ domain $e \in \mathcal{E}$, we assume that each instance $\mathbf{x}_i^e$ is generated from (1) a latent semantic factor $\mathbf{c} \in C$, where $C = \{\mathbf{c}_{y=0}, \mathbf{c}_{y=1}\}$; (2) a latent sensitive factor $\mathbf{a} \in \mathcal{A}$, where $\mathcal{A} = \{\mathbf{a}_{z=1}, \mathbf{a}_{z=-1}\}$; and (3) a latent style factor $\mathbf{s}^e$, where $\mathbf{s}^e$ is specific to the individual domain e. We assume that the semantic and sensitive factors in C and $\mathcal{A}$ do not change across domains. Each domain $\mathbb{P}^e_{XZY}$ is represented by a style factor $\mathbf{s}^e$ and the dependence score $\rho^e = \rho(Y^e, Z^e)$[2], denoted $e := (\mathbf{s}^e, \rho^e)$, where $\mathbf{s}^e$ and ρ^e are unique to the domain $\mathbb{P}^e_{XZY}$.*

Note that Assump. 1 is similarly related to the one made in [19, 34, 40, 57]. In our paper, with a focus on group fairness, we

[2]Here, ρ functions equivalently as it does in Eq. (1), by substituting $\hat{Y}$ to Y.

expand upon the assumptions of existing works by introducing three latent factors. Under Assump. 1, if two instances $(\mathbf{x}^{e_i}, z^{e_i}, y)$ and $(\mathbf{x}^{e_j}, z^{e_j}, y)$ where $e_i, e_j \in \mathcal{E}, i \neq j$ share the same class label, then the latter instance can be reconstructed by decoder G from the former using $\mathbf{c} = E_c(\mathbf{x}^{e_i})$, $\mathbf{s} = E_s(\mathbf{x}^{e_j})$, and $\mathbf{a} = E_a(\mathbf{x}^{e_j})$ through T, denoted $(\mathbf{x}^{e_j}, z^{e_j}) = T(\mathbf{x}^{e_i}, z^{e_i}, e_j)$.

To enhance the effectiveness of the transformation model T, our overall learning loss for these encoders and decoders consists of two main components: a bidirectional reconstruction loss and a sensitiveness loss.

Data reconstruction loss encourages learning reconstruction in the direction of data→latent→data. As for it, a data sample $\mathbf{x}^s$ from $\mathbb{P}^s_X, \forall s \in \mathcal{E}^s$ is required to be reconstructed by its encoded factors.

$$\mathcal{L}^{data}_{recon} = \mathbb{E}_{\mathbf{x}^s \sim \mathbb{P}^s_X}[\|G(E_c(\mathbf{x}^s), E_a(\mathbf{x}^s), E_s(\mathbf{x}^s)) - \mathbf{x}^s\|_1]$$

Factor reconstruction loss. Given latent factors $\mathbf{c}$, $\mathbf{a}$, and $\mathbf{s}^s$ encoded from a sample $\mathbf{x}^s$, they are encouraged to be reconstructed through some latent factors randomly sampled from the prior Gaussian distributions.

$$\begin{aligned}\mathcal{L}^{factor}_{recon} =& \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{a}\sim\mathbb{P}_A, \mathbf{s}^s\sim\mathbb{P}_S}[\|E_c(G(\mathbf{c}, \mathbf{a}, \mathbf{s}^s)) - \mathbf{c}\|_1] \\ &+ \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{a}\sim\mathcal{N}(0,\mathrm{I}_a), \mathbf{s}^s\sim\mathbb{P}_S}[\|E_a(G(\mathbf{c}, \mathbf{a}, \mathbf{s}^s)) - \mathbf{a}\|_1] \\ &+ \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{a}\sim\mathbb{P}_A, \mathbf{s}^s\sim\mathcal{N}(0,\mathrm{I}_s)}[\|E_s(G(\mathbf{c}, \mathbf{a}, \mathbf{s}^s)) - \mathbf{s}\|_1]\end{aligned}$$

where $\mathbb{P}_C$, $\mathbb{P}_A$, $\mathbb{P}_S$ are given by $E_c(\mathbf{x}^s)$, $E_a(\mathbf{x}^s)$, and $E_s(\mathbf{x}^s)$, respectively.

Sensitiveness loss. Since a sensitive factor is causally dependent on the sensitive attribute of data $(\mathbf{x}^s, z^s, y^s)$, a simple classifier $h : \mathcal{A} \to \mathcal{Z}$ is learned, and further it is used to label the sensitive attribute in augmented data when learning f.

$$\mathcal{L}_{sens} = CrossEntropy(z^s, h(E_a(\mathbf{x}^s)))$$

Total loss. We jointly train the encoders and the decoder to optimize the transformation model T with a weighted sum loss.

$$\min_{E_c, E_a, E_s, G} \beta_1 \mathcal{L}^{data}_{recon} + \beta_2 \mathcal{L}^{factor}_{recon} + \beta_3 \mathcal{L}_{sens} \tag{3}$$

where $\beta_1, \beta_2, \beta_3 > 0$ are hyperparameters that control the importance of each loss term.

4.2 Fair Disentangled Domain Generalization

Furthermore, with a trained transformation model T, to learn the fairness-aware invariant classifier f across domains, we make the following assumption.

Assumption 2 (Fairness-aware Domain Shift). *We assume that inter-domain variation is characterized by covariate and dependence shifts. As a consequence, we assume that the conditional distribution $\mathbb{P}^e_{Y|XZ}$ is stable across domains, $\forall e \in \mathcal{E}$. Given a transformation model T, it holds that $\mathbb{P}^{e_i}_{Y|XZ} = \mathbb{P}^{e_j}_{Y|XZ}$, $\forall e_i, e_j \in \mathcal{E}, i \neq j$, where $(X^{e_j}, Z^{e_j}) = T(\mathbf{X}^{e_i}, Z^{e_i}, e_j)$.*

In Assump. 2, the domain shift captured by T would characterize the mapping from the marginal distributions $\mathbb{P}^{e_i}_X$ and $\rho(Y^{e_i}, Z^{e_i})$ over $\mathcal{D}^{e_i}$ to the distribution $\mathbb{P}^{e_j}_X$ and $\rho(Y^{e_j}, Z^{e_j})$ over $\mathcal{D}^{e_j}$ sampled from a different data domain $\mathbb{P}^{e_j}_{XZY}$, respectively. With this in mind and under Assump. 2, we introduce a new definition of fairness-aware invariance with respect to the variation captured by T and satisfying the group fair constraint introduced in Defn. 1.

Definition 3 (Fairness-aware T-Invariance). *Given a transformation model T, a fairness-aware classifier $f \in \mathcal{F}$ is domain invariant if it holds for all $e_i, e_j \in \mathcal{E}$.*

$$f(\mathbf{x}^{e_i}) = f(\mathbf{x}^{e_j}), \ \text{and} \ \rho(f(X^{e_i}), Z^{e_i}) = \rho(f(X^{e_j}), Z^{e_j}) = 0 \tag{4}$$

almost surely when $(\mathbf{x}^{e_j}, z^{e_j}) = T(\mathbf{x}^{e_i}, z^{e_i}, e_j)$, $\mathbf{x}^{e_i} \sim \mathbb{P}^{e_i}_X$, $\mathbf{x}^{e_j} \sim \mathbb{P}^{e_j}_X$.

Defn. 3 is crafted to enforce invariance on the predictions generated by f directly. We expect a prediction to remain consistent across various data realizations T while considering group fairness.

Problem 2 (Fair Disentanglement Domain Generalization). *Under Defn. 3 and Assump. 2, if we restrict $\mathcal{F}$ of Prob. 1 to the set of invariant fairness-aware classifiers, the Prob. 1 is equivalent to the following problem*

$$\begin{aligned}P^\star \triangleq \min_{f\in\mathcal{F}} R(f) \triangleq \mathbb{E}_{\mathbb{P}^{s_i}_{XZY}} \ell(f(X^{s_i}), Y^{s_i}) \\ s.t.\ f(X^{s_i}) = f(X^{s_j}),\ \rho(f(X^{s_i}), Z^{s_i}) = \rho(f(X^{s_j}), Z^{s_j}) = 0\end{aligned} \tag{5}$$

where $(X^{s_j}, Z^{s_j}) = T(X^{s_i}, Z, s_j), \forall s_i, s_j \in \mathcal{E}^s, i \neq j$.

Similar to [40], Prob. 2 is not a composite optimization problem. Moreover, acquiring domain labels is often expensive or even unattainable, primarily due to privacy concerns. Consequently, under the assumptions of disentanglement-based invariance and domain shift, Prob. 1 can be approximated to Prob. 2 by removing the max operator over $\mathcal{E}$.

In addition, Prob. 2 offers a new and theoretically-principled perspective on Prob. 1, when data varies from domain to domain with respect to T. To optimize Prob. 2 is challenging because (1) The strict equality constraints in Prob. 2 are difficult to enforce in practice; (2) Enforcing constraints on deep networks is known to be a challenging problem due to non-convexity. Simply transforming them to regularization cannot guarantee satisfaction for constrained problems; and (3) As we have incomplete access to all domains, it limits the ability to enforce fairness-aware T-invariance and further makes it hard to estimate $R(f)$.

Due to such challenges, we develop a tractable method for approximately solving Prob. 2 with optimality guarantees. To address the first challenge, we relax constraints in Prob. 2

$$\begin{aligned}P^\star(\gamma_1, \gamma_2) \triangleq \min_{f\in\mathcal{F}} R(f) \\ \text{s.t. } \delta^{s_i, s_j}(f) \leq \gamma_1,\ \rho^{s_i}(f) \leq \frac{\gamma_2}{2},\ \text{and } \rho^{s_j}(f) \leq \frac{\gamma_2}{2}\end{aligned} \tag{6}$$

where

$$\delta^{s_i, s_j}(f) \triangleq \mathbb{E}_{\mathbb{P}^{s_i}_{XZ}} d\big[f(X^{s_i}), f(X^{s_j} = T(X^{s_i}, Z^{s_i}, s_j))\big], \tag{7}$$

$$\rho^{s_i}(f) \triangleq \rho(f(X^{s_i}), Z^{s_i}), \quad \rho^{s_j}(f) \triangleq \rho(f(X^{s_j}), Z^{s_j}) \tag{8}$$

and $\forall s_i, s_j \in \mathcal{E}^s, i \neq j$. Here, $\gamma_1, \gamma_2 > 0$ are constants controlling the extent of relaxation and $d[\cdot]$ is a distance metric, *e.g.,* KL-divergence. When $\gamma_1 = \gamma_2 = 0$, Eqs. (5) and (6) are equivalent.

Since it is unrealistic to have access to the full distribution and we only have access to source domains, given data sampled from $\mathcal{E}_s$, we consider the empirical dual problem.

$$\begin{aligned}D^\star_{\xi, N, \mathcal{E}_s}(\gamma_1, \gamma_2) \triangleq \max_{\lambda_1(s_i,s_j), \lambda_2(s_i,s_j)} \min_{\boldsymbol{\theta}\in\Theta} \hat{R}(\boldsymbol{\theta}) \\ + \frac{1}{|\mathcal{E}^s|} \sum_{s_i, s_j \in \mathcal{E}_s} \Big[\lambda_1(s_i, s_j)\big(\hat{\delta}^{s_i, s_j}(\boldsymbol{\theta}) - \gamma_1\big) \\ + \lambda_2(s_i, s_j)\big(\hat{\rho}^{s_i}(\boldsymbol{\theta}) + \hat{\rho}^{s_j}(\boldsymbol{\theta}) - \gamma_2\big)\Big]\end{aligned} \tag{9}$$

Algorithm 1 FEDORA: Fair Disentangled Domain Generalization.

Require: Encoders $E = \{E_c, E_a, E_s\}$, decoder G and sensitive classifier h.
Initialize: primal and dual learning rate η_p, η_d, empirical constant γ_1, γ_2.

1: **repeat**
2: **for** minibatch $\mathcal{B} = \{(\mathbf{x}_i, z_i, y_i)\}_{i=1}^m \subset \mathcal{D}_s$ **do**
3: $\mathcal{L}_{cls}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^m \ell(y_i, \hat{f}(\mathbf{x}_i, \boldsymbol{\theta}))$
4: Initialize $\mathcal{L}_{inv}(\boldsymbol{\theta}) = 0$ and $\mathcal{B}' = [\ ]$
5: **for** each $(\mathbf{x}_i, z_i, y_i)$ in the minibatch **do**
6: Generate $(\mathbf{x}_j, z_j, y_j) = T(\mathbf{x}_i, z_i, y_i)$ and add to $\mathcal{B}'$
7: $\mathcal{L}_{inv}(\boldsymbol{\theta}) += \frac{1}{m} d[\hat{f}(\mathbf{x}_i, \boldsymbol{\theta}), \hat{f}(\mathbf{x}_j, \boldsymbol{\theta})]$
8: **end for**
9: $\mathcal{L}_{fair}(\boldsymbol{\theta}) = \left|\frac{1}{m}\sum_{(\mathbf{x}_i, z_i)\in\mathcal{B}} g(\hat{f}(\mathbf{x}_i, \boldsymbol{\theta}), z_i)\right| + \left|\frac{1}{m}\sum_{(\mathbf{x}_j, z_j)\in\mathcal{B}'} g(\hat{f}(\mathbf{x}_j, \boldsymbol{\theta}), z_j)\right|$
10: $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{cls}(\boldsymbol{\theta}) + \lambda_1 \cdot \mathcal{L}_{inv}(\boldsymbol{\theta}) + \lambda_2 \cdot \mathcal{L}_{fair}(\boldsymbol{\theta})$
11: $\boldsymbol{\theta} \leftarrow \text{Adam}(\mathcal{L}(\boldsymbol{\theta}), \boldsymbol{\theta}, \eta_p)$
12: $\lambda_1 \leftarrow \max\{[\lambda_1 + \eta_d \cdot (\mathcal{L}_{inv}(\boldsymbol{\theta}) - \gamma_1)], 0\}$, $\lambda_2 \leftarrow \max\{[\lambda_2 + \eta_d \cdot (\mathcal{L}_{fair}(\boldsymbol{\theta}) - \gamma_2)], 0\}$
13: **end for**
14: **until** convergence
15: **procedure** $T(\mathbf{x}, z, y)$
16: $\mathbf{c}, \mathbf{a}, \mathbf{s} = E(\mathbf{x})$
17: Sample $\mathbf{a}' \sim \mathcal{N}(0, I_a)$, $\mathbf{s}' \sim \mathcal{N}(0, I_s)$
18: $\mathbf{x}' = G(\mathbf{c}, \mathbf{a}', \mathbf{s}')$, $z' = h(\mathbf{a}')$
19: **return** $(\mathbf{x}', z', y)$
20: **end procedure**

where $\xi = \mathbb{E}_{\mathbb{P}_X}||f(\mathbf{x}) - \hat{f}(\mathbf{x}, \boldsymbol{\theta})||_\infty > 0$ is a constant bounding the difference between f and its parameterized counterpart $\hat{f} : \mathcal{X} \times \Theta \to \mathbb{R}$ defined in the Defn. 5.1 of [40]. N is the number of samples drawn from $\mathbb{P}_{XZY}$ and it can be empirically replaced by $\sum_{s\in\mathcal{E}^s} |\mathcal{D}^s|$. $\lambda_1(s_i, s_j), \lambda_2(s_i, s_j) > 0$ are dual variables. $\hat{R}(\boldsymbol{\theta})$, $\hat{\delta}^{s_i,s_j}(\boldsymbol{\theta})$, $\hat{\rho}^{s_i}(\boldsymbol{\theta})$ and $\hat{\rho}^{s_j}(\boldsymbol{\theta})$ are the empirical counterparts of $R(\hat{f}(\cdot, \boldsymbol{\theta}))$, $\delta^{s_i,s_j}(\hat{f}(\cdot, \boldsymbol{\theta}))$, $\rho^{s_i}(\hat{f}(\cdot, \boldsymbol{\theta}))$ and $\rho^{s_j}(\hat{f}(\cdot, \boldsymbol{\theta}))$, respectively.

4.3 The FEDORA Algorithm

In practice, we propose a simple but effective algorithm, given in Algorithm 1, which is co-trained with the transformation model T. The detailed training process of T is provided in Algorithm 2 of the Appendix. In Algorithm 1, we harness the power of T to address the unconstrained dual optimization problem outlined in Eq. (9) through a series of primal-dual iterations.

Given a finite number of observed source domains, to enhance the generalization performance for unseen target domains, the invariant classifier $\hat{f}$ is trained by expanding the dataset with synthetic domains generated by T. These synthetic domains are created by introducing random sample style and random sensitive factors, hence a random sensitive attribute, resulting in an arbitrary fair dependence within such domains. As described in Fig. 2, the sensitive factor $\mathbf{a}^{s'}$ and the style factor $\mathbf{s}^{s'}$ are randomly sampled from their prior distributions $\mathcal{N}(0, \mathbf{I}_a)$ and $\mathcal{N}(0, \mathbf{I}_s)$, respectively. A sensitive attribute $z^{s'}$ is further predicted from $\mathbf{a}^{s'}$ through h. Along with the unchanged semantic factor $\mathbf{c}$ encoded by $(\mathbf{x}^s, z^s, y)$, they are further passed through G to generate $(\mathbf{x}^{s'}, z^{s'}, y)$ with the unchanged class labels in an augmented synthetic domain. Under Assump. 2 and Defn. 3, according to Eqs. (7) and (8), data augmented in synthetic domains are required to maintain invariance in terms of accuracy and fairness with the data in the corresponding original domains.

Specifically, in lines 15-20 of Algorithm 1, we describe the transformation procedure that takes an example $(\mathbf{x}, z, y)$ as INPUT and returns an augmented example $(\mathbf{x}', z', y)$ from a new synthetic domain as OUTPUT. The augmented example has the same semantic factor as the input example but has different sensitive and style factors sampled from their associated prior distributions that encode a new synthetic domain. Lines 1-14 show the main training loop for FEDORA. In line 6, for each example in the minibatch $\mathcal{B}$, we apply the procedure T to generate an augmented example from a new synthetic domain described above. In line 7, we consider KL-divergence as the distance metric for $d[\cdot]$. All the augmented examples are stored in the set $\mathcal{B}'$. The Lagrangian dual loss function is defined based on $\mathcal{B}$ and $\mathcal{B}'$ in line 10. The primal parameters $\boldsymbol{\theta}$ and the dual parameters λ_1 and λ_2 are updated in lines 11-12.

5 ANALYSIS

With the approximation on the dual problem in Eq. (9), the duality gap between $P^\star$ in Eq. (6) and $D^\star_{\xi,N,\mathcal{E}_s}(\gamma_1, \gamma_2)$ in Eq. (9) can be explicitly bounded.

Theorem 1 (Fairness-aware Data-dependent Duality Gap). *Given $\xi > 0$, assuming $\{\hat{f}(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{F}$ has finite VC-dimension, with M datapoints sampled from $\mathbb{P}_{XZY}$ we have*

$$|P^\star - D^\star_{\xi,N,\mathcal{E}_s}(\boldsymbol{\gamma})| \le L||\boldsymbol{\gamma}||_1 + \xi k(1 + ||\boldsymbol{\lambda}^\star_p||_1) + O(\sqrt{\log(M)/M})$$

where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2]^T$; L is the Lipschitz constant of $P^\star(\gamma_1, \gamma_2)$; k is a small universal constant defined in Proposition 3 of Appendix B; and $\boldsymbol{\lambda}^\star_p$ is the optimal dual variable for a perturbed version of Eq. (6).

The duality gap that arises when solving the empirical problem presented in Eq. (9) is minimal when the fairness-aware T-invariance in Defn. 3 margin $\boldsymbol{\gamma}$ is narrow, and the parametric space closely approximates $\mathcal{F}$.

Furthermore, we present the following theorem to establish an upper bound on fairness within an unseen target domain.

Theorem 2 (Fairness Upper Bound of the Unseen Target Domain). *In accordance with Defn. 1 and Eq.* (8)*, for any domain $e \in \mathcal{E}$, the fairness dependence under instance distribution $\mathbb{P}^e_{XZY}$ with respect to the classifier $f \in \mathcal{F}$ is defined as:*

$$\rho^e(f) = \left|\mathbb{E}_{\mathbb{P}^e_{XZ}} g(f(X^e), Z^e)\right|$$

With observed source domains $\mathcal{E}_s$, the dependence at any unseen target domain $t \in \mathcal{E}\backslash\mathcal{E}_s$ is upper bounded. $dist[\cdot]$ is the Jensen-Shannon distance metric [12].

$$\rho^t(f) \le \frac{1}{|\mathcal{E}_s|}\sum_{s_i\in\mathcal{E}_s} \rho^{s_i}(f) + \sqrt{2}\min_{s_i\in\mathcal{E}_s} dist\left[\mathbb{P}^t_{XZY}, \mathbb{P}^{s_i}_{XZY}\right] + \sqrt{2}\max_{s_i,s_j\in\mathcal{E}_s} dist\left[\mathbb{P}^{s_i}_{XZY}, \mathbb{P}^{s_j}_{XZY}\right]$$

where $dist[\mathbb{P}_1, \mathbb{P}_2] = \sqrt{\frac{1}{2}KL(\mathbb{P}_1||\frac{\mathbb{P}_1+\mathbb{P}_2}{2}) + \frac{1}{2}KL(\mathbb{P}_2||\frac{\mathbb{P}_1+\mathbb{P}_2}{2})}$ is JS divergence defined based on KL divergence.

Notice that the second term in Theorem 2 becomes uncontrollable during training as it relies on the unseen target domain. Therefore, to preserve fairness across target domains, we aim to learn semantic factors that map the transformation mode T, ensuring

Table 3: Statistics summary of all datasets.

Datasets	Domains	Sensitive Attr.	Labels	$(s, \rho^s), \forall s \in \mathcal{E}_s$
ccMNIST	digit color	background color	digit label	(R, 0.11), (G, 0.43), (B, 0.87)
FairFace	race	gender	age	(B, 0.91), (E, 0.87), (I, 0.58), (L, 0.48), (M, 0.87), (S, 0.39), (W, 0.49)
YFCC100M-FDG	year	location	in-,outdoor	(d_0, 0.73), (d_1, 0.84), (d_2, 0.72)
NYSF	city	race	stop record	(R, 0.93), (B, 0.85), (M, 0.81), (Q, 0.98), (S, 0.88)

that $\mathbb{P}^{s_i}_{C|XZY}, \forall s_i \in \mathcal{E}_s$ remains invariant across source domains. Simultaneously, we strive for the classifier f to achieve high fairness within the source domains. Proofs of Theorems 1 and 2 are provided in Appendices B and C.

6 EXPERIMENTS

Due to space limitations, we defer a detailed description of the experimental settings and comprehensive results to the arXiv version of this paper, which can be accessed at https://arxiv.org/pdf/2311.13816.

6.1 Settings

Datasets. We evaluate the performance of our FEDORA on four benchmarks. To highlight each source data and its fair dependence score ρ^s defined in Assump. 1, we summarize the statistics in Tab. 3.

(1) ccMNIST is a domain generalization benchmark created by colorizing digits and the backgrounds of the MNIST dataset [27]. ccMNIST consists of images of handwritten digits from 0 to 9. Similar to ColoredMNIST [3], for binary classification, digits are labeled with 0 and 1 for digits from 0-4 and 5-9, respectively. ccMNIST contains 70,000 images divided into three data domains, each characterized by a different digit color (*i.e.,* red, green, blue) and followed by a different correlation between the class label and sensitive attribute (digit background colors). **(2)** FairFace [23] is a dataset that contains a balanced representation of different racial groups. It includes 108,501 images from seven racial categories: Black (B), East Asian (E), Indian (I), Latino (L), Middle Eastern (M), Southeast Asian (S), and White (W). In our experiments, we set each racial group as a domain, gender as the sensitive attributes, and age ($\geq$ or $<$ 50) as the class label. **(3)** YFCC100M-FDG is an image dataset created by *Yahoo Labs* and released to the public in 2014. It is randomly selected from the YFCC100M [49] dataset with a total of 90,000 images. For domain variations, YFCC100M-FDG is divided into three domains. Each contains 30,000 images from different year ranges, before 1999 (d_0), 2000 to 2009 (d_1), and 2010 to 2014 (d_2). The outdoor or indoor tag is used as the binary class label for each image. Latitude and longitude coordinates, representing where images were taken, are translated into different continents. The North American or non-North American continent is the sensitive attribute (related to spatial disparity). **(4)** NYSF [25] is a real-world dataset on policing in New York City in 2011. It documents whether a pedestrian who was stopped on suspicion of weapon possession would, in fact, possess a weapon. NYSF consists of records collected in five different regions: Manhattan (M), Brooklyn (B), Queens (Q), Bronx (R), and Staten (S). We use regions as different domains. This data had a pronounced racial bias against African Americans, so we consider race (black or non-black) as the sensitive attribute.

Baselines. We compare the performance of FEDORA with 19 baseline methods that fall into two main categories: **(1)** 12 state-of-the-art domain generalizations methods, specifically designed to address covariate shifts: ColorJitter, ERM [50], IRM [3], GDRO [42], Mixup [55], MLDG [28], CORAL [47], MMD [29], DANN [14], CDANN [30], DDG [57], and MBDG [40], where ColorJitter is a naive function in *PyTorch* that randomly changes the brightness, contrast, saturation and hue of images; and **(2)** 7 state-of-the-art fairness-aware domain generalizations methods, specifically designed to address either covariate or dependence shifts: DDG-FC, MBDG-FC, EIIL [8], FarconVAE [36], FCR [2], FTCS [41], and FATDM [37], where DDG-FC and MBDG-FC are two baselines that built upon DDG [57] and MBDG [40], respectively by straightforwardly adding fairness constraints defined in Defn. 1 to the loss functions of the original models.

Evaluation metrics. Three metrics are used for evaluation. Two of them are for fairness quantification, Demographic Parity (*DP*) [11] and the Area Under the ROC Curve (AUC_{fair}) between predictions of sensitive subgroups [33].

- *Demographic Parity* (*DP*) [11] is formalized as

$$\text{DP} = k, \text{ if DP} \leq 1; \text{DP} = 1/k, \text{ otherwise}$$

  where $k = \mathbb{P}(\hat{Y} = 1|Z = -1)/\mathbb{P}(\hat{Y} = 1|Z = 1)$ This is also known as a lack of disparate impact [13]. A value closer to 1 indicates fairness.
- *The Area Under the ROC Curve* (AUC_{fair}) [6] varies from zero to one, and it is symmetric around 0.5, which represents random predictability or zero bias effect on predictions.

$$AUC_{fair} = \frac{\sum_{(\mathbf{x}_i, z=-1, y_i) \in \mathcal{D}_{-1}} \sum_{(\mathbf{x}_j, z=1, y_j) \in \mathcal{D}_1} I\big(\mathbb{P}(\hat{y}_i = 1) > \mathbb{P}(\hat{y}_j = 1)\big)}{|\mathcal{D}_{-1}| \times |\mathcal{D}_1|}$$

  where $|\mathcal{D}_{-1}|$ and $|\mathcal{D}_1|$ represent sample size of subgroups $z = -1$ and $z = 1$, respectively. $I(\cdot)$ is the indicator function that returns 1 when its argument is true and 0 otherwise.

Notice that the AUC_{fair} is not the same as the one commonly used in classification based on TPR and FPR. The intuition behind this AUC_{fair} is based on the nonparametric *Mann-Whitney U* test, in which a fair condition is defined as the classifier's prediction probability of a randomly selected sample $\mathbf{x}_{-1}$ from one sensitive subgroup being greater than a randomly selected sample $\mathbf{x}_1$ from the other sensitive subgroup is equal to the probability of $\mathbf{x}_1$ being greater than $\mathbf{x}_{-1}$ [6, 58]. A value of *DP* closer to 1 indicates fairness, and 0.5 of AUC_{fair} represents zero bias effect on predictions.

Model selection. The model selection in domain generalization is intrinsically a learning problem, followed by [40], we use leave-one-domain-out validation criteria, which is one of the three selection methods stated in [17]. Specifically, we evaluate FEDORA on the held-out source domain and average the performance of $|\mathcal{E}_s| - 1$ domains over the held-out one.

Hyperparameter Search We follow the same set of the MUNIT [19] for the hyperparameters. More specifically, the learning rate is 0.0001, the number of iterations is 600000, and the batch size is 1. The loss weights in learning T are chosen from $\{1, 5, 10\}$. The selected best ones are $\beta_1 = 10, \beta_2 = 1, \beta_3 = 1, \beta_4 = 1$. We monitor the loss of the validation set and choose the β with the lowest validation loss. For the hyperparameters in learning the classifier f, the learning rate is chosen from $\{0.000005, 0.00001, 0.00005, 0.0001, 0.0005\}$.

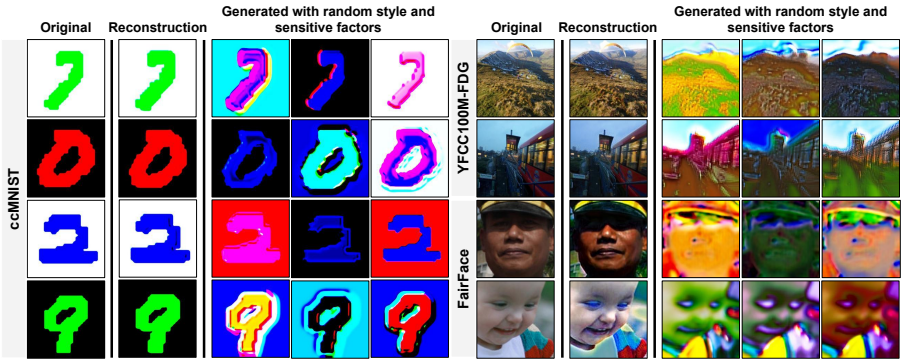

Figure 3: Visualizations for images under reconstruction and the transformation model T with random style and sensitive factors.

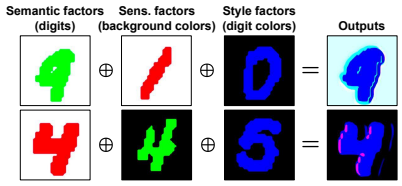

Figure 4: Example results of generating images using latent factors encoded from three images.

η is chosen from $\{0.01, 0.05, 0.1\}$. γ is chosen from $\{0.01, 0.025, 0.05\}$. λ is chosen from $\{0.1, 1, 10, 20\}$. The batch size is chosen from $\{22, 64, 80, 128, 512, 1024, 2048\}$. The numbers of iterations is chosen from $\{500, 1000, ..., 8000\}$ on the ccMNIST and NYSF datasets. The number of iterations are chosen from $\{300, 600, ..., 7800, 8000\}$ on the FairFace and YFCC100M-FDG datasets. The selected best ones are: the learning rate is 0.00005, $\eta_1 = \eta_2 = 0.05$, $\gamma_1 = \gamma_2 = 0.025$, $\lambda_1 = \lambda_2 = 1$. The batch size on the ccMNIST and YFCC100M-FDG datasets is 64, and it is 22 on the FairFace dataset and 1024 on the NYSF dataset. The number of iterations on the ccMNIST dataset is 3000, 500, 7000 for domains R, G, B, respectively. The number of iterations on the FairFace dataset is 7200, 7200, 7800, 8000, 6600, 7200, 6900 for domains B, E, I, L, M, S, W, respectively. The number of iterations on the YFCC100M-FDG dataset is 7200, 6000, 6900 for d_0, d_1, d_2, respectively. The number of iterations on the NYSF dataset is 500, 3500, 4000, 1500, 8000 for domains R, B, M, Q, and S, respectively. We monitor the accuracy and the value of fairness metrics from the validation set and select the best ones. The grid space of the grid search on all the baselines is the same as that of our method.

6.2 Results

Data augmentation in synthetic domains via T. We visualize the augmented samples with random variations in Fig. 3. The first column (Original) shows the images sampled from the datasets. In the second column (Reconstruction), we display images generated from latent factors encoded from the images in the first column. The images in the second column closely resemble those in the first column. Images in the last three columns are generated using the semantic factors encoded from images in the first column, associated with style and sensitive factors randomly sampled from their respective Gaussian distributions. The images in the last three columns preserve the fundamental semantic information of the corresponding samples in the first column. However, their style and sensitive attributes undergo significant changes at random. The generated images within synthetic domains enhance the classifier's generalization (f) to unseen source domains. This demonstrates that the transformation model T effectively extracts latent factors and produces diverse transformations of the provided data domains.

Effectiveness of T. To further validate the effectiveness of T, drawing inspiration from [19], we train a separate transformation model for each domain. Subsequently, we generate an output image by utilizing distinct latent factors from each domain. Using ccMNIST as an example, we individually train three transformation models $\{T^i\}_{i=1}^3$ within each domain. Each T^i includes unique encoders E_c^i, E_a^i, and E_s^i. As shown in Fig. 4, an output image is generated through G using a semantic factor (digit class, $E_c^1(\mathbf{x}^1)$), a sensitive factor (background color, $E_a^2(\mathbf{x}^2)$), and a style factor (digit color, $E_s^3(\mathbf{x}^3)$) from images in different domains. As a result, the output image is constructed from the digit of $\mathbf{x}^1$, the background color of $\mathbf{x}^2$, and the digit color of $\mathbf{x}^3$, with given variations. This suggests that the augmented data with random variations in Fig. 3 for the synthetic domain are not merely altering colors; instead, they are precisely generated with unchanged semantics and random sensitive and style factors.

The effectiveness of FEDORA across domains in terms of predicted fairness and accuracy. Comprehensive experiments showcase that FEDORA consistently outperforms baselines by a considerable margin. For all tables in the main paper and Appendix, results shown in each column represent performance on the target domain, using the rest as source domains. Due to space limit, selected results for three domains of FairFace are shown in Tab. 4, but the average results are based on all domains. As shown in Tab. 4, for the FairFace dataset, our method has the best accuracy and fairness level for the average DG performance over all the domains. More specifically, our method has better fairness metrics (3% for DP, 2% for AUC_{fair}) and comparable accuracy (0.19% better) than the best of the baselines for individual metrics. As shown in Tab. 5, for YFCC100M-FDG, our method excels in fairness metrics (8% for DP, 4% for AUC_{fair}) and comparable accuracy (0.35% better) compared to the best baselines.

Ablation studies. We conduct three ablation studies to study the robustness of FEDORA on FairFace. In-depth descriptions and the pseudocodes for these studies and more results can be found in the arXiv version of our paper at https://arxiv.org/pdf/2311.13816. (1) In FEDORA w/o E_a, we modify the encoder within T by restricting its output to only latent semantic and style factors. (2) FEDORA w/o T skips data augmentation in synthetic domains via T and results are conducted only based f constrained by fair notions outlined in Defn. 1. (3) In FEDORA w/o $\mathcal{L}_{fair}$, the fair constraint on f is not included, and we eliminate the $\mathcal{L}_{fair}$ in line 9 of Algorithm 1. We include the performance of such ablation studies in

Table 4: Performance on FairFace. (bold is the best; underline is the second best).

Methods	$DP\uparrow / AUC_{fair}\downarrow /$ Accuracy $\uparrow$			
	(B, 0.91)	(W, 0.49)	(L, 0.48)	Avg
ColorJitter	0.64±0.26 / 0.64±0.15 / 93.47±1.56	0.34±0.09 / 0.64±0.02 / 92.07±0.55	0.39±0.10 / 0.70±0.02 / 91.77±0.61	0.42 / 0.66 / 92.94
ERM	0.67±0.17 / 0.58±0.02 / 91.89±1.10	0.39±0.09 / 0.61±0.01 / 92.82±0.38	0.57±0.15 / 0.62±0.01 / 91.96±0.51	0.51 / 0.61 / 93.08
IRM	0.63±0.12 / 0.58±0.01 / 93.39±1.03	0.32±0.19 / 0.66±0.01 / 90.54±1.56	0.41±.021 / 0.63±0.05 / 92.06±1.89	0.43 / 0.62 / 92.48
GDRO	0.71±0.16 / 0.57±0.02 / 89.81±1.10	0.48±0.09 / 0.60±0.01 / 92.50±0.38	0.54±0.15 / 0.62±0.01 / 91.59±0.51	0.55 / 0.60 / 92.55
Mixup	0.58±0.19 / 0.59±0.02 / 92.46±0.69	0.43±0.19 / 0.61±0.01 / 92.98±0.03	0.55±0.22 / 0.61±0.02 / 93.43±2.02	0.51 / 0.60 / 93.19
MLDG	0.63±0.25 / 0.58±0.02 / 92.71±2.36	0.47±0.20 / 0.59±0.01 / 92.82±1.65	0.53±0.18 / 0.62±0.03 / 92.99±0.86	0.51 / 0.60 / 93.39
CORAL	0.69±0.19 / 0.58±0.01 / 92.09±2.03	0.50±0.14 / 0.60±0.02 / 92.47±2.04	0.56±0.23 / **0.59**±0.03 / 92.62±1.11	0.54 / 0.60 / 93.21
MMD	0.69±0.25 / 0.56±0.01 / 93.87±0.14	0.39±0.20 / 0.68±0.02 / 91.75±1.37	0.55±0.16 / 0.61±0.02 / 92.53±1.41	0.50 / 0.60 / 92.34
DANN	0.46±0.07 / 0.61±0.02 / 91.80±0.64	0.11±0.09 / 0.66±0.01 / 86.80±1.18	0.39±0.21 / 0.67±0.01 / 90.82±2.44	0.47 / 0.70 / 90.10
CDANN	0.62±0.24 / 0.59±0.03 / 91.22±0.33	0.35±0.17 / 0.67±0.02 / 90.19±0.60	0.42±0.23 / 0.61±0.03 / 92.42±2.19	0.43 / 0.66 / 91.48
DDG	0.60±0.20 / 0.59±0.02 / 91.76±1.03	0.51±0.07 / 0.60±0.01 / 91.34±0.80	0.44±0.17 / 0.62±0.02 / 93.46±0.32	0.49 / 0.61 / 92.74
MBDG	0.60±0.15 / 0.58±0.01 / 91.29±1.41	0.30±0.04 / 0.62±0.01 / 91.05±0.53	0.56±0.09 / 0.61±0.01 / 93.49±0.97	0.50 / 0.60 / 92.71
DDG-FC	0.61±0.06 / 0.58±0.03 / 92.27±1.65	0.48±0.15 / 0.62±0.02 / 92.45±1.55	0.50±0.25 / 0.62±0.03 / 92.42±0.30	0.52 / 0.61 / 93.23
MBDG-FC	0.70±0.15 / 0.56±0.03 / 92.12±0.43	0.32±0.07 / 0.60±0.03 / 91.50±0.57	0.57±0.23 / 0.62±0.02 / 91.89±0.81	0.53 / 0.60 / 92.48
EIIL	0.88±0.07 / 0.59±0.05 / 84.75±2.16	0.46±0.05 / 0.65±0.03 / 86.53±1.02	0.49±0.07 / **0.59**±0.01 / 88.39±1.25	0.64 / 0.61 / 87.78
FarconVAE	0.93±0.03 / **0.54**±0.01 / 89.61±0.64	0.51±0.07 / 0.60±0.01 / 86.40±0.42	**0.58**±0.05 / 0.60±0.05 / 88.70±0.71	0.66 / **0.58** / 88.46
FCR	0.81±0.05 / 0.59±0.02 / 79.66±0.25	0.39±0.06 / 0.63±0.02 / 82.33±0.89	0.38±0.12 / 0.66±0.02 / 85.22±2.33	0.54 / 0.63 / 83.68
FTCS	0.75±0.10 / 0.60±0.02 / 80.00±0.20	0.40±0.06 / 0.60±0.02 / 79.66±1.05	0.42±0.23 / 0.65±0.03 / 79.64±1.00	0.57 / 0.64 / 80.91
FATDM	0.93±0.03 / 0.57±0.02 / 92.20±0.36	0.46±0.05 / 0.63±0.01 / 92.56±0.31	0.51±0.16 / 0.63±0.02 / 93.33±0.20	0.67 / 0.61 / 92.54
FEDORA	**0.94**±0.05 / 0.55±0.02 / **93.91**±0.33	**0.52**±0.17 / **0.58**±0.03 / **93.02**±0.50	**0.58**±0.15 / **0.59**±0.01 / **93.73**±0.26	**0.70** / **0.58** / **93.42**

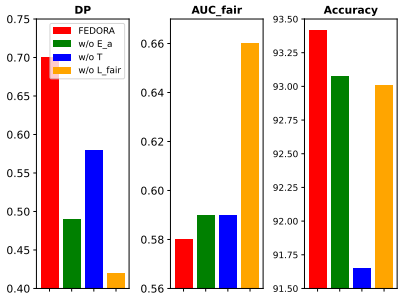

Figure 5: Ablation study on FairFace. Averaged results are plotted across all domains.

Table 5: Performance on YFCC100M-FDG. (Bold is the best; underline is the second best.)

Methods	$DP\uparrow / AUC_{fair}\downarrow /$ Accuracy $\uparrow$			
	$(d_0, 0.73)$	$(d_1, 0.84)$	$(d_2, 0.72)$	Avg
ColorJitter	0.67±0.06 / 0.57±0.02 / 57.47±1.20	0.67±0.34 / 0.61±0.01 / 82.43±1.25	0.65±0.21 / 0.64±0.02 / 87.88±0.35	0.66 / 0.61 / 75.93
ERM	0.81±0.09 / 0.58±0.01 / 40.51±0.23	0.71±0.18 / 0.66±0.03 / 83.91±0.33	0.89±0.08 / 0.59±0.01 / 82.06±0.33	0.80 / 0.61 / 68.83
IRM	0.76±0.10 / 0.58±0.02 / 50.51±2.44	0.87±0.08 / 0.60±0.02 / 73.26±0.03	0.70±0.24 / 0.57±0.02 / 82.78±2.19	0.78 / 0.58 / 68.85
GDRO	0.80±0.05 / 0.59±0.01 / 53.43±2.29	0.73±0.22 / 0.60±0.01 / 87.56±2.20	0.79±0.13 / 0.65±0.02 / 83.10±0.64	0.78 / 0.62 / 74.70
Mixup	0.82±0.07 / 0.57±0.03 / 61.15±0.28	0.79±0.14 / 0.63±0.03 / 78.63±0.97	0.89±0.05 / 0.60±0.01 / 85.18±0.80	0.84 / 0.60 / 74.99
MLDG	0.75±0.13 / 0.67±0.01 / 49.56±0.69	0.71±0.19 / 0.57±0.02 / 89.45±0.44	0.71±0.14 / 0.57±0.03 / 87.51±0.18	0.72 / 0.60 / 75.51
CORAL	0.80±0.11 / 0.58±0.02 / 58.96±2.34	0.72±0.11 / 0.64±0.03 / 91.66±0.85	0.70±0.07 / 0.64±0.03 / 89.28±1.77	0.74 / 0.62 / 79.97
MMD	0.79±0.11 / 0.59±0.02 / 61.51±1.79	0.71±0.15 / 0.64±0.03 / 91.15±2.33	0.79±0.17 / 0.60±0.01 / 86.69±0.19	0.76 / 0.61 / 79.87
DANN	0.70±0.13 / 0.78±0.02 / 47.71±1.56	0.79±0.12 / 0.53±0.01 / 84.80±1.14	0.77±0.17 / 0.59±0.02 / 58.50±1.74	0.75 / 0.64 / 63.67
CDANN	0.74±0.13 / 0.58±0.02 / 55.87±2.09	0.70±0.22 / 0.65±0.02 / 87.06±2.43	0.72±0.13 / 0.63±0.02 / 85.76±2.43	0.72 / 0.62 / 76.23
DDG	0.81±0.14 / 0.57±0.03 / 60.08±1.08	0.74±0.12 / 0.66±0.03 / 92.53±0.91	0.71±0.21 / 0.59±0.03 / **95.02**±1.92	0.75 / 0.61 / 82.54
MBDG	0.79±0.15 / 0.58±0.01 / 60.46±1.90	0.73±0.07 / 0.67±0.01 / 94.36±0.23	0.71±0.11 / 0.59±0.03 / 93.48±0.65	0.74 / 0.61 / 82.77
DDG-FC	0.76±0.06 / 0.58±0.03 / 59.96±2.36	0.83±0.06 / 0.58±0.01 / **96.80**±1.28	0.82±0.09 / 0.59±0.01 / 86.38±2.45	0.80 / 0.58 / 81.04
MBDG-FC	0.80±0.13 / 0.58±0.01 / 62.31±0.13	0.72±0.09 / 0.63±0.01 / 94.73±2.09	0.80±0.07 / **0.53**±0.01 / 87.78±2.11	0.77 / 0.58 / 81.61
EIIL	**0.87**±0.11 / 0.55±0.02 / 56.74±0.60	0.76±0.05 / 0.54±0.03 / 68.99±0.91	0.87±0.06 / 0.78±0.03 / 72.19±0.75	0.83 / 0.62 / 65.98
FarconVAE	0.67±0.06 / 0.61±0.03 / 51.21±0.61	0.90±0.06 / 0.59±0.01 / 72.40±2.13	0.85±0.12 / 0.55±0.01 / 74.20±2.46	0.81 / 0.58 / 65.93
FCR	0.62±0.21 / 0.70±0.01 / 55.32±0.04	0.63±0.14 / 0.66±0.10 / 70.89±0.22	0.66±0.30 / 0.78±0.02 / 70.58±0.23	0.64 / 0.71 / 65.60
FTCS	0.72±0.03 / 0.60±0.01 / 60.21±0.10	0.79±0.02 / 0.59±0.01 / 79.96±0.05	0.69±0.10 / 0.60±0.06 / 72.99±0.50	0.73 / 0.60 / 71.05
FATDM	0.80±0.10 / 0.55±0.01 / 61.56±0.89	0.88±0.08 / 0.56±0.01 / 90.00±0.66	0.86±0.10 / 0.60±0.02 / 89.12±1.30	0.84 / 0.57 / 80.22
FEDORA	**0.87**±0.09 / **0.53**±0.01 / **62.56**±2.25	**0.94**±0.05 / **0.52**±0.01 / 93.36±1.70	**0.93**±0.03 / **0.53**±0.02 / 93.43±0.73	**0.92** / **0.53** / **83.12**

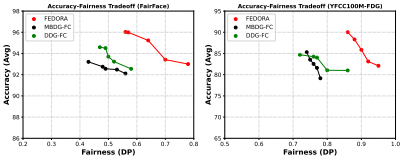

Figure 6: Results of accuracy-fairness tradeoff on Fairface (left) and YFCC100M-FDG (right) sweeping over a range of λ_2.

Fig. 5. The results illustrate that when data is disentangled into three factors, and the model is designed accordingly, it can enhance generalization performance due to covariate and dependence shifts. Generating data in synthetic domains with random fairness dependence patterns proves to be an effective approach for ensuring fairness invariance across domains.

Fairness-accuracy tradeoff. In our Algorithm 1, because λ_2 (lines 10 and 12) is the parameter that regularizes the fair loss, we conduct additional experiments to show the change of tradeoffs between accuracy and fairness sweeping over a range of $\lambda_2 \in [0.01, 0.05, 0.1, 1, 10]$. Our results show that the larger (small) λ_2, the better(worse) model fairness for each domain as well as in average, but it gives worse (better) model accuracy. Evaluation on FairFace and YFCC100M-FDG is given in Fig. 6. Results in the top-right of the figure indicate good performance. This result is plotted on the average performance over all target domains.

7 CONCLUSION

In this paper, we introduce a novel approach designed to tackle the challenges of domain generalization when confronted with covariate shift and dependence shift simultaneously. In our pursuit of learning a fairness-aware invariant classifier, we assert the existence of an underlying transformation model that can transform instances across domains. This model plays a crucial role in achieving fairness-aware domain generalization by generating samples in synthetic domains characterized by novel data styles and fair dependence patterns. We present a tractable algorithm and showcase its effectiveness through comprehensive analyses and exhaustive empirical studies.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant numbers 2147375, IIS-2107449, and IIS-1954376, and in part by the National Center for Transportation Cybersecurity and Resiliency (TraCR), a U.S. Department of Transportation National University Transportation Center headquartered at Clemson University, Clemson, South Carolina, USA. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of TraCR, and the U.S. Government assumes no liability for the contents or use thereof.

REFERENCES

[1] Sergio Alonso, Rosana Montes, Daniel Molina, Iván Palomares, Eugenio Martínez-Cámara, Manuel Chiachio, Juan Chiachio, Francisco J Melero, Pablo García-Moral, Bárbara Fernández, et al. 2021. Ordering artificial intelligence based recommendations to tackle the sdgs with a decision-making model based on surveys. *Sustainability* 13, 11 (2021), 6038.

[2] Bang An, Zora Che, Mucong Ding, and Furong Huang. 2022. Transferring fairness under distribution shifts via fair consistency regularization. *Advances in Neural Information Processing Systems* 35 (2022), 32582–32597.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[4] Arpita Biswas and Suvam Mukherjee. 2021. Ensuring fairness under prior probability shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 414–424.

[5] Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems* 24 (2011).

[6] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling Attribute Effect in Linear Regression. *ICDM* (2013).

[7] Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. 2022. Fairness transferability subject to bounded distribution shift. *Advances in Neural Information Processing Systems* 35 (2022), 11266–11278.

[8] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*. PMLR, 2189–2200.

[9] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. 2020. Causal modeling for fairness in dynamical systems. In *International conference on machine learning*. PMLR, 2185–2195.

[10] Wei Du and Xintao Wu. 2021. Fair and robust classification under sample selection bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2999–3003.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. *CoRR* (2011).

[12] Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory* 49, 7 (2003), 1858–1860.

[13] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. *KDD* (2015).

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[15] Stephen Giguere, Blossom Metevier, Yuriy Brun, Bruno Castro da Silva, Philip S Thomas, and Scott Niekum. 2022. Fairness guarantees under demographic shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[17] Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434* (2020).

[18] Xiao Han, Lu Zhang, Yongkai Wu, and Shuhan Yuan. 2023. Achieving Counterfactual Fairness for Anomaly Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 55–66.

[19] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 172–189.

[20] Vasileios Iosifidis and Eirini Ntoutsi. 2020. FABBOO-Online Fairness-Aware Learning Under Class Imbalance. In *International Conference on Discovery Science*. Springer, 159–174.

[21] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. 2019. Fairness-enhancing interventions in stream classification. In *Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30*. Springer, 261–276.

[22] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.

[23] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1548–1558.

[24] Gill Kirton. 2019. Unions and equality: 50 years on from the fight for fair pay at Dagenham. *Employee Relations: The International Journal* 41, 2 (2019), 344–356.

[25] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *ICML*.

[26] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*. PMLR, 5815–5826.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[29] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5400–5409.

[30] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 624–639.

[31] Yujie Lin, Dong Li, Chen Zhao, Xintao Wu, Qin Tian, and Minglai Shao. 2024. Supervised Algorithmic Fairness in Distribution Shifts: A Survey. *arXiv preprint arXiv:2402.01327* (2024).

[32] Yujie Lin, Chen Zhao, Minglai Shao, Baoluo Meng, Xujiang Zhao, and Haifeng Chen. 2023. Pursuing Counterfactual Fairness via Sequential Autoencoder Across Domains. *ArXiv:2309.13005* (2023).

[33] Charles X Ling, Jin Huang, Harry Zhang, et al. 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, Vol. 3. 519–524.

[34] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems* 30 (2017).

[35] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. 2020. Too Relaxed to Be Fair. In *ICML*.

[36] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. 2022. Learning Fair Representation via Distributional Contrastive Disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1295–1305.

[37] Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. 2023. Fairness and accuracy under domain generalization. *Proceedings of the International Conference on Learning Representations* (2023).

[38] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. 2020. Fairness for robust log loss classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5511–5518.

[39] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.

[40] Alexander Robey, George J Pappas, and Hamed Hassani. 2021. Model-based domain generalization. *Advances in Neural Information Processing Systems* 34 (2021), 20210–20229.

[41] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2023. Improving fair training under correlation shifts. *ICML* (2023).

[42] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. *International Conference on Learning Representations* (2020).

[43] Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. 2022. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems* 35 (2022), 19304–19318.

[44] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. 2019. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688* (2019).

[45] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.

[46] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[47] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 443–450.

[48] Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. 2020. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530* (2020).

[49] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[50] Vladimir Vapnik. 1999. *The nature of statistical learning theory*. Springer science & business media.

[51] Riccardo Volpi, Diane Larlus, and Grégory Rogez. 2021. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4443–4453.

[52] Ke Wang, Senqiang Zhou, Chee Ada Fu, and Jeffrey Xu Yu. 2003. Mining changes of classification by correspondence tracing. In *Proceedings of the 2003 SIAM International Conference on Data Mining*. SIAM, 95–106.

[53] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning* 23 (1996), 69–101.

[54] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On Convexity and Bounds of Fairness-aware Classification. *WWW*.

[55] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. 2020. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677* (2020).

[56] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. *ICML* (2013).

[57] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. 2022. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8024–8034.

[58] Chen Zhao and Feng Chen. 2019. Rank-Based Multi-task Learning For Fair Regression. *IEEE International Conference on Data Mining (ICDM)* (2019).

[59] Chen Zhao and Feng Chen. 2020. Unfairness Discovery and Prevention For Few-Shot Regression. *ICKG* (2020).

[60] Chen Zhao, Feng Chen, and Bhavani Thuraisingham. 2021. Fairness-Aware Online Meta-learning. *ACM SIGKDD* (2021).

[61] Chen Zhao, Changbin Li, Jincheng Li, and Feng Chen. 2020. Fair Meta-Learning For Few-Shot Classification. *ICKG* (2020).

[62] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, and Feng Chen. 2022. Adaptive Fairness-Aware Online Meta-Learning for Changing Environments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2565–2575. https://doi.org/10.1145/3534678.3539420

[63] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, and Feng Chen. 2024. Dynamic Environment Responsive Online Meta-Learning with Fairness Awareness. *ACM Transactions on Knowledge Discovery from Data* 18, 6, Article 153 (2024), 23 pages. https://doi.org/10.1145/3648684

[64] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, Christan Grant, and Feng Chen. 2023. Towards Fair Disentangled Online Learning for Changing Environments. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3480–3491.

[65] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 561–578.

A DETAILS OF LEARNING THE TRANSFORMATION MODEL

For simplicity, we denote the transformation model T consisting of three encoders E_c, E_a, E_s, and a decoder G. However, in practice, we consider a bi-level auto-encoder (see Fig. 7), wherein an additional content encoder $E_m : \mathcal{X} \to \mathcal{M}$ takes data as input and outputs a content factor. Furthermore, the decoder G used in the main paper is renamed G_o. Specifically, the inner level decoder is denoted as $G_i : \mathcal{C} \times \mathcal{A} \to \mathcal{M}$. As a consequence, the transformation model T consists of encoders $E = \{E_m, E_s, E_c, E_a\}$ and decoders $G = \{G_i, G_o\}$.

Specifically, in the outer level, an instance is first encoded to a content factor $\mathbf{m} \in \mathcal{M}$ and a style factor $\mathbf{s} \in \mathcal{S}$ through the corresponding encoders E_m and E_s, respectively. In the inner level, the content factor $\mathbf{m}$ is further encoded to a content factor $\mathbf{c} \in C$ and a sensitive factor $\mathbf{a} \in \mathcal{A}$, through encoders E_c and E_a. Therefore, the bidirectional reconstruction loss and the sensitiveness loss stated in Sec. 4 are reformulated.

$$\begin{aligned}\mathcal{L}_{recon}^{data} =& \mathbb{E}_{\mathbf{x}^s\sim\mathbb{P}_X^s}\left[\left\|G_o\left(\hat{\mathbf{m}}, E_s(\mathbf{x}^s)\right) - \mathbf{x}^s\right\|_1\right] \\ &+ \mathbb{E}_{\mathbf{m}\sim\mathbb{P}_M}\left[\left\|G_i\left(E_c(\mathbf{m}), E_a(\mathbf{m})\right) - \mathbf{m}\right\|_1\right]\end{aligned}$$

Algorithm 2 Learning the Transformation Model T.

Require: learning rate $\alpha_1, \alpha_2, \alpha_3$, initial coefficients $\beta_1, \beta_2, \beta_3, \beta_4$.
Initialize: Parameter of encoders $\{\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_c, \boldsymbol{\theta}_a\}$, decoders $\{\boldsymbol{\phi}_i, \boldsymbol{\phi}_o\}$, sensitive classifier $\boldsymbol{\theta}_z$, and discriminators $\{\boldsymbol{\psi}_i, \boldsymbol{\psi}_o\}$.

repeat
  for minibatch $\{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^q \in \mathcal{D}_s$ **do**
    Compute $\mathcal{L}_{total}$ for Stage 1 using Eq. (10).
    $\boldsymbol{\psi}_o, \boldsymbol{\psi}_i \leftarrow \text{Adam}(\beta_4 \mathcal{L}_{adv}, \boldsymbol{\psi}_o, \boldsymbol{\psi}_i, \alpha_1)$
    $\boldsymbol{\theta}_m, \boldsymbol{\theta}_c, \boldsymbol{\theta}_s, \boldsymbol{\theta}_a, \boldsymbol{\phi}_o, \boldsymbol{\phi}_i \leftarrow \text{Adam}(\beta_1 \mathcal{L}_{recon}^{data} + \beta_2 \mathcal{L}_{recon}^{factor}, \boldsymbol{\theta}_m, \boldsymbol{\theta}_c, \boldsymbol{\theta}_s, \boldsymbol{\theta}_a, \boldsymbol{\phi}_o, \boldsymbol{\phi}_i, \alpha_2)$
    $\boldsymbol{\theta}_z \leftarrow \text{Adam}(\beta_3 \mathcal{L}_{sens}, \boldsymbol{\theta}_z, \alpha_3)$
  end for
until convergence
Return $\{\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_c, \boldsymbol{\theta}_a, \boldsymbol{\theta}_z, \boldsymbol{\phi}_i, \boldsymbol{\phi}_o\}$

where $\hat{\mathbf{m}} = G_i(\mathbf{c}, \mathbf{a}) = G_i\big(E_c(E_m(\mathbf{x}^s)), E_a(E_m(\mathbf{x}^s))\big)$; $\mathbb{P}_M$ is given by $\mathbf{m} = E_m(\mathbf{x}^s)$.

$$\begin{aligned}\mathcal{L}_{recon}^{factor} =& \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{a}\sim\mathcal{N}(0,\mathbf{I}_a)}\left[\left\|E_c\big(G_i(\mathbf{c},\mathbf{a})\big) - \mathbf{c}\right\|_1\right] \\ &+ \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{a}\sim\mathcal{N}(0,\mathbf{I}_a)}\left[\left\|E_a\big(G_i(\mathbf{c},\mathbf{a})\big) - \mathbf{a}\right\|_1\right] \\ &+ \mathbb{E}_{\mathbf{m}\sim\mathbb{P}_M, \mathbf{s}^s\sim\mathcal{N}(0,\mathbf{I}_s)}\left[\left\|E_s\big(G_o(\mathbf{m},\mathbf{s})\big) - \mathbf{s}\right\|_1\right] \\ &+ \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{s}^s\sim\mathcal{N}(0,\mathbf{I}_s), \mathbf{a}\sim\mathcal{N}(0,\mathbf{I}_a)}\left[\left\|E_s\big(G_o(G_i(\mathbf{c},\mathbf{a}),\mathbf{s})\big) - \mathbf{s}\right\|_1\right] \\ &+ \mathbb{E}_{\mathbf{m}\sim\mathbb{P}_M, \mathbf{s}^s\sim\mathcal{N}(0,\mathbf{I}_s)}\left[\left\|E_m\big(G_o(\mathbf{m},\mathbf{s})\big) - \mathbf{m}\right\|_1\right]\end{aligned}$$

where $\mathbb{P}_C$ and $\mathbb{P}_M$ are given by $\mathbf{c} = E_c(E_m(\mathbf{x}^s))$ and $\mathbf{m} = E_m(\mathbf{x}^s)$. $\mathbf{a} = E_a(E_m(\mathbf{x}^s))$, and $\mathbf{s} = E_s(\mathbf{x}^s)$.

$$\mathcal{L}_{sens} = CrossEntropy(z^s, h(E_a(E_m(\mathbf{x}^s))))$$

Additionally, motivated by the observation that GANs [16] can improve data quality for evaluating the disentanglement effect in the latent spaces, we use GANs to match the distribution of reconstructed data to the same distribution. Followed by [19], data and semantic factors generated through encoders and decoders should be indistinguishable from the given ones in the same domain.

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{s}^s\sim\mathcal{N}(0,\mathbf{I}_s), \mathbf{a}\sim\mathcal{N}(0,\mathbf{I}_a)}\left[\log\left(1 - D_o(G_o(\hat{\mathbf{m}}, \mathbf{s}^s))\right)\right] \\ &+ \mathbb{E}_{\mathbf{x}^s\sim\mathbb{P}_X^s}\left[\log D_o(\mathbf{x}^s)\right] \\ &+ \mathbb{E}_{\mathbf{c}\sim\mathbb{P}_C, \mathbf{a}\sim\mathcal{N}(0,\mathbf{I}_a)}\left[\log\left(1 - D_i(G_i(\mathbf{c},\mathbf{a}))\right)\right] \\ &+ \mathbb{E}_{\mathbf{m}\sim\mathbb{P}_M}\left[\log D_i(\mathbf{m})\right]\end{aligned}$$

where $D_o : \mathcal{X} \to \mathbb{R}$ and $D_i : \mathcal{M} \to \mathbb{R}$ are the discriminators for the outer and inner levels, respectively.

Total Loss. We jointly train the encoders, decoders, and discriminators to optimize the final objective, a weighted sum of the three loss terms.

$$\min_{E_m, E_s, E_c, E_a, G_i, G_o} \max_{D_i, D_o} \beta_1 \mathcal{L}_{recon}^{data} + \beta_2 \mathcal{L}_{recon}^{factor} + \beta_3 \mathcal{L}_{sens} + \beta_4 \mathcal{L}_{adv} \tag{10}$$

where $\beta_1, \beta_2, \beta_3, \beta_4 > 0$ are hyperparameters that control the importance of each loss term. To optimize, the learning algorithm is given in Algorithm 2.

B SKETCH PROOF OF THEOREM 1

Before we prove Theorem 1, we first make the following propositions and assumptions.

Proposition 1. *Let d' be a distance metric between probability measures for which it holds that $d'[\mathbb{P}, \mathbb{T}] = 0$ for two distributions $\mathbb{P}$ and $\mathbb{T}$ if and only if $\mathbb{P} = \mathbb{T}$ almost surely. Then $P^\star(0,0) = P^\star$*

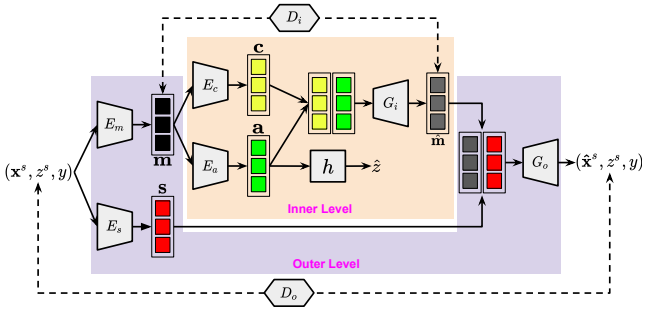

Figure 7: A two-level approach for learning the transformation model T.

Proposition 2. *Assuming the perturbation function $P^\star(\gamma_1, \gamma_2)$ is L-lipschitz continuous in γ_1, γ_2. Then given Proposition 1, it follows that $|P^\star - P^\star(\gamma_1, \gamma_2)| \leq L||\boldsymbol{\gamma}||_1$, where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2]^T$.*

Definition 4. *Let $\Theta \subseteq \mathbb{R}^p$ be a finite-dimensional parameter space. For $\xi > 0$, a function $\hat{f} : \mathcal{X} \times \Theta \to \mathcal{Y}$ is said to be an ξ-parameterization of $\mathcal{F}$ if it holds that for each $f \in \mathcal{F}$, there exists a parameter $\boldsymbol{\theta} \in \Theta$ such that $\mathbb{E}_{\mathbb{P}_X}\|\hat{f}(\mathbf{x}, \boldsymbol{\theta}) - f(\mathbf{x})\|_\infty \leq \xi$. Given an ξ-parameterization $\hat{f}$ of $\mathcal{F}$, consider the following saddle-point problem:*

$$
\begin{aligned}
&D^\star_\xi(\gamma_1, \gamma_2) \\
&\triangleq \max_{\lambda_1(s_i,s_j), \lambda_2(s_i,s_j)} \min_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}) + \int_{s_i,s_j\in\mathcal{E}_s} [\delta^{s_i,s_j}(\boldsymbol{\theta}) - \gamma_1] \mathrm{d}\lambda_1(s_i, s_j) \\
&+ \int_{s_i,s_j\in\mathcal{E}_s} [\rho^{s_i}(\boldsymbol{\theta}) + \rho^{s_j}(\boldsymbol{\theta}) - \gamma_2] \mathrm{d}\lambda_2(s_i, s_j)
\end{aligned}
$$

where $R(\boldsymbol{\theta}) = R(\hat{f}(\cdot, \boldsymbol{\theta}))$ and $\mathcal{L}^{s_i,s_j}(\boldsymbol{\theta}) = \mathcal{L}^{s_i,s_j}(\hat{f}(\cdot, \boldsymbol{\theta}))$.

Assumption 3. *The loss function ℓ is non-negative, convex, and L_ℓ-Lipschitz continuous in its first argument,*

$$|\ell(f_1(\mathbf{x}), y) - \ell(f_2(\mathbf{x}), y)| \leq \|f_1(\mathbf{x}) - f_2(\mathbf{x})\|_\infty$$

Assumption 4. *The distance metric d is non-negative, convex, and satisfies the following uniform Lipschitz-like inequality for some constant $L_d > 0$:*

$$
\begin{aligned}
&|d[f_1(\mathbf{x}), f_1(\mathbf{x}' = T(\mathbf{x}, z, s))] - d[f_2(\mathbf{x}), f_2(\mathbf{x}' = T(\mathbf{x}, z, s))]| \\
&\leq L_d\|f_1(\mathbf{x}) - f_2(\mathbf{x})\|_\infty, \quad \forall s \in \mathcal{E}_s
\end{aligned}
$$

Assumption 5. *The fairness metric g is non-negative, convex, and satisfies the following uniform Lipschitz-like inequality for some constant $L_g > 0$:*

$$|(g \circ f_1)(\mathbf{x}, z) - (g \circ f_2)(\mathbf{x}, z)| \leq L_g\|f_1(\mathbf{x}) - f_2(\mathbf{x})\|_\infty$$

Assumption 6. *There exists a parameter $\boldsymbol{\theta} \in \Theta$ such that $\delta^{s_i,s_j}(\boldsymbol{\theta}) < \gamma_1 - \xi\cdot\max\{L_\ell, L_d\}$ and $\rho^{s_i}(\boldsymbol{\theta}) + \rho^{s_j}(\boldsymbol{\theta}) < \gamma_2 - \xi\cdot\max\{L_\ell, L_g\}, \forall s_i, s_j \in \mathcal{E}_s$*

Proposition 3. *Let $\gamma_1, \gamma_2 > 0$ be given. With the assumptions above, it holds that*

$$P^\star(\gamma_1, \gamma_2) \leq D^\star_\xi(\gamma_1, \gamma_2) \leq P^\star(\gamma_1, \gamma_2) + \xi(1 + \|\lambda^\star_p\|_1) \cdot k$$

where $\lambda^\star_p$ is the optimal dual variable for a perturbed version of Eq. (6) in which the constraints are tightened to hold with margin $\gamma - \xi \cdot k$, $k = \max\{L_\ell, L_d, L_g\}$. In particular, this result implies that

$$|P^\star(\gamma_1, \gamma_2) - D^\star_\xi(\gamma_1, \gamma_2)| \leq \xi k(1 + \|\lambda^\star_p\|_{L_1})$$

Proposition 4 (Empirical gap). *Assume ℓ and d are non-negative and bounded in $[-B, B]$ and let d_{VC} denote the VC-dimension of the hypothesis class $\mathcal{A}_\xi = \{\hat{f}(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{F}$. Then it holds with probability $1 - \omega$ over the N samples from each domain that*

$$|D^\star_\xi(\gamma_1, \gamma_2) - D^\star_{\xi,N,\mathcal{E}_s}(\gamma_1, \gamma_2)| \leq 2B\sqrt{\frac{1}{N}[1 + \log(\frac{4(2N)^{d_{\mathrm{VC}}}}{\omega})]}$$

Let $\xi > 0$ be given, and let $\hat{f}$ be an ξ-parameterization of $\mathcal{F}$. Let the assumptions hold, and further assume that ℓ, d, and g are $[0, B]$-bounded and that $d[\mathbb{P}, \mathbb{T}] = 0$ if and only if $\mathbb{P} = \mathbb{T}$ almost surely, and that $P^\star(\gamma_1, \gamma_2)$ is L-Lipschitz. Then assuming that $\mathcal{A}_\xi = \{\hat{f}(\cdot, \theta) : \theta \in \Theta\} \subseteq \mathcal{F}$ has finite VC-dimension, it holds with probability $1 - \omega$ over the N samples that

$$|P^\star - D^\star_{\xi,N,\mathcal{E}_s}(\boldsymbol{\gamma})| \leq L||\boldsymbol{\gamma}||_1 + \xi k(1 + ||\lambda^\star_p||_1) + O(\sqrt{\log(M)/M})$$

Now we prove Theorem 1.

Proof. The proof of this theorem is a simple consequence of the triangle inequality. Indeed, by combining Propositions 2 to 4, we find that

$$
\begin{aligned}
&|P^\star - D^\star_{\xi,N,\mathcal{E}_s}(\gamma_1, \gamma_2)| \\
=&|P^\star + P^\star(\gamma_1, \gamma_2) - P^\star(\gamma_1, \gamma_2) + D^\star_\xi(\gamma_1, \gamma_2) - D^\star_\xi(\gamma_1, \gamma_2) - D^\star_{\xi,N,\mathcal{E}_s}(\gamma_1, \gamma_2)| \\
\leq&|P^\star - P^\star(\gamma_1, \gamma_2)| + |P^\star(\gamma_1, \gamma_2) - D^\star_\xi(\gamma_1, \gamma_2)| + |D^\star_\xi(\gamma_1, \gamma_2) - D^\star_{\xi,N,\mathcal{E}_s}(\gamma_1, \gamma_2)| \\
\leq&L\|\gamma\|_1 + \xi k(1 + \|\lambda^\star_p\|_1) + 2B\sqrt{\frac{1}{N}[1 + \log(\frac{4(2N)^{d_{\mathrm{VC}}}}{\omega})]}
\end{aligned}
$$

□

C SKETCH PROOF OF THEOREM 2

Lemma 1. *Given two domains $e_i, e_j \in \mathcal{E}$, $\mathbb{E}_{\mathbb{P}^{e_j}_{XZ}} g(f(X^{e_j}), Z^{e_j})$ can be bounded by $\mathbb{E}_{\mathbb{P}^{e_i}_{XZ}} g(f(X^{e_i}), Z^{e_i})$ as follows:*

$$\mathbb{E}_{\mathbb{P}^{e_j}_{XZ}} g(f(X^{e_j}), Z^{e_j}) \leq \mathbb{E}_{\mathbb{P}^{e_i}_{XZ}} g(f(X^{e_i}), Z^{e_i}) + \sqrt{2}dist[\mathbb{P}^{e_j}_{XZY}, \mathbb{P}^{e_i}_{XZY}]$$

Lemma 2. *Given two domains $e_i, e_j \in \mathcal{E}$, under Lemma 1, $\rho^{e_j}(f)$ can be bounded by $\rho^{e_i}(f)$ as follows:*

$$\rho^{e_j}(f) \leq \rho^{e_i}(f) + \sqrt{2}dist[\mathbb{P}^{e_j}_{XZY}, \mathbb{P}^{e_i}_{XZY}]$$

Under Lemmas 1 and 2, we now prove Theorem 2

Proof. Let $s_\star \in \mathcal{E}_s$ be the source domain nearest to the target domain $t \in \mathcal{E}\backslash\mathcal{E}_s$. Under Lemma 2, we have

$$\rho^t(f) \leq \rho^{s_i}(f) + \sqrt{2}dist[\mathbb{P}^t_{XZY}, \mathbb{P}^{e_i}_{XZY}]$$

where $s_i \in \mathcal{E}_s$. Taking the average of upper bounds based on all source domains, we have:

$$
\begin{aligned}
\rho^t(f) \leq& \frac{1}{|\mathcal{E}_s|}\sum_{s_i\in\mathcal{E}_s}\rho^{s_i}(f) + \frac{\sqrt{2}}{|\mathcal{E}_s|}\sum_{s_i\in\mathcal{E}_s} dist[\mathbb{P}^t_{XZY}, \mathbb{P}^{e_i}_{XZY}] \\
\leq& \frac{1}{|\mathcal{E}_s|}\sum_{s_i\in\mathcal{E}_s}\rho^{s_i}(f) + \frac{\sqrt{2}}{|\mathcal{E}_s|}|\mathcal{E}_s| dist[\mathbb{P}^t_{XZY}, \mathbb{P}^{s_\star}_{XZY}] \\
&+ \frac{\sqrt{2}}{|\mathcal{E}_s|}\sum_{s_i\in\mathcal{E}_s} dist[\mathbb{P}^{s_\star}_{XZY}, \mathbb{P}^{s_i}_{XZY}] \\
\leq& \frac{1}{|\mathcal{E}_s|}\sum_{s_i\in\mathcal{E}_s}\rho^{s_i}(f) + \sqrt{2}\min_{s_i\in\mathcal{E}_s} dist[\mathbb{P}^t_{XZY}, \mathbb{P}^{e_i}_{XZY}] \\
&+ \sqrt{2}\max_{s_i,s_j\in\mathcal{E}_s} dist[\mathbb{P}^{e_i}_{XZY}, \mathbb{P}^{e_j}_{XZY}]
\end{aligned}
$$

□