Fair Domain Generalization: An Information-Theoretic View

Tangzheng Lian[1], **Guanyu Hu**[2,3], **Dimitrios Kollias**[3], **Xinyu Yang**[2], **Oya Celiktutan**[1]
[1]King's College London
[2]Xi'an Jiaotong University
[3]Queen Mary University of London
lian.tangzheng@kcl.ac.uk, g.hu@qmul.ac.uk, d.kollias@qmul.ac.uk, yxyphd@mail.xjtu.edu.cn, oya.celiktutan@kcl.ac.uk

Abstract

Domain generalization (DG) and algorithmic fairness are two key challenges in machine learning. However, most DG methods focus solely on minimizing expected risk in the unseen target domain, without considering algorithmic fairness. Conversely, fairness methods typically do not account for domain shifts, so the fairness achieved during training may not generalize to unseen test domains. In this work, we bridge these gaps by studying the problem of Fair Domain Generalization (FairDG), which aims to minimize both expected risk and fairness violations in unseen target domains. We derive novel mutual information-based upper bounds for expected risk and fairness violations in multi-class classification tasks with multi-group sensitive attributes. These bounds provide key insights for algorithm design from an information-theoretic perspective. Guided by these insights, we propose a practical method that solves the FairDG problem through Pareto optimization. Experiments on real-world vision and language datasets show that our method achieves superior utility–fairness trade-offs compared to existing approaches.

Introduction

In real-world deployments, machine learning models often face domain shift, where test data comes from a domain that ***never seen*** during training (e.g., new demographics, lighting conditions, or image styles). Domain generalization (DG) tackles a more challenging setting, aiming to train models that perform well on unseen domains without access to their data or labels. Instead, it typically assumes the availability of multiple distinct but related source domains during training. Prior DG research has proposed various techniques, including domain-invariant representation learning (Ganin et al. 2016), data augmentation (Dunlap et al. 2023), and meta-learning (Li et al. 2018a). However, these methods focus solely on minimizing expected risk in the target domain and overlook algorithmic fairness. As a result, models that generalize well may still exhibit unfairness in unseen domains.

In parallel, the field of algorithmic fairness in machine learning focuses on mitigating biases in model predictions. Among fairness notions like individual and counterfactual fairness, we focus on the widely used ***group fairness*** (Caton and Haas 2024), which aims to prevent performance disparities across subgroups defined by a sensitive attribute. These disparities often arise from imbalances in the training data. For instance, as shown in Fig. 1, if white faces dominate the training data while black faces are under-represented, a model trained for facial expression recognition may achieve higher accuracy for white faces and lower accuracy for black faces simply because white faces are more frequent in the training set. Many methods have been proposed to enforce group fairness, typically categorized into pre-processing, in-processing, and post-processing techniques (Mehrabi et al. 2021). However, these methods generally do not account for domain shifts, so the fairness achieved during training may not generalize to unseen test domains. In this paper, we bridge these gaps by addressing the challenge of Fair Domain Generalization (FairDG), which aims to jointly minimize expected risk and fairness violations in unseen target domains. Our contributions are summarized as follows:

(1) We derive novel theoretical upper bounds based on

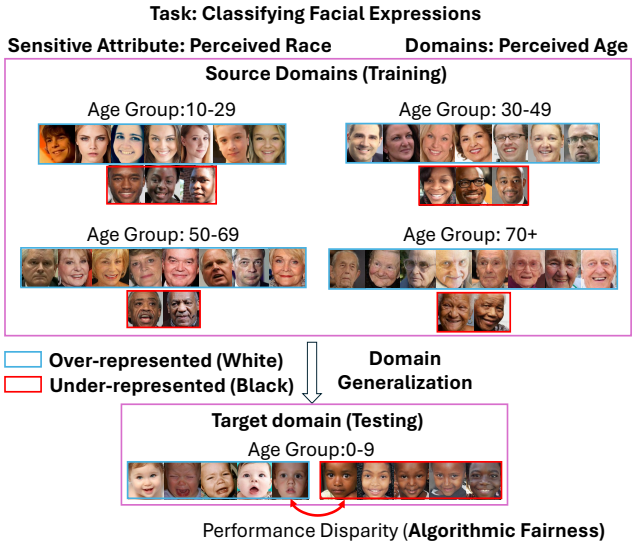

Figure 1: A real-world example of the FairDG problem. The goal is to train a model that generalizes to an unseen domain (age group: 0-9) while also ensuring fairness by minimizing performance disparities across perceived racial groups.

mutual information (MI) for both the expected risk and fairness violations in multi-class classification tasks with multi-group sensitive attributes, offering key insights from an information-theoretic perspective that inform algorithm design for solving the FairDG problem.

(2) We propose a method to solve FairDG while modeling the utility-fairness trade-off via Pareto optimization.

(3) Experimental results on real-world natural language and vision datasets show that our method outperforms existing approaches, achieving better utility-fairness trade-offs.

Related Works

The FairDG problem lies within the broader area of ensuring algorithmic fairness under distribution shifts. Please see (Shao et al. 2024; Barrainkua et al. 2025) for comprehensive surveys. However, most prior work focuses on fairness in the domain adaptation, which assumes access to unlabeled target data to adapt to the domain shifts (Chen et al. 2022; Wang et al. 2023; Rezaei et al. 2021; Singh et al. 2021). In contrast, only a few studies have addressed fairness in the more challenging DG setting as discussed below.

Lin et al. proposed two approaches to FairDG: one focusing on group fairness (Lin et al. 2024a) and the other on counterfactual fairness (Lin et al. 2024b). However, both methods were evaluated only on synthetic and tabular datasets, leaving their effectiveness on real-world, high-dimensional data such as text and images unclear. (Jiang et al. 2024) introduced a meta-learning method, while (Tian et al. 2024) proposed a plug-and-play fair identity attention module for medical image segmentation and classification. (Zhao et al. 2024) addressed FairDG using synthetic data augmentation with learned transformations, and (Palakkadavath et al. 2025) studied the setting with heterogeneous sensitive attributes across domains. Although these methods involve real-world data such as images, none provide theoretical guarantees to support their algorithmic designs. In contrast, our work offers upper bounds on both expected risk and fairness violations and presents a practical framework validated on real-world natural language and vision datasets.

(Pham, Zhang, and Zhang 2023) is the first work in FairDG that proposes a method with upper bounds for both fairness violations and expected risk. However, their bounds scale poorly as the number of classes, source domains, and attribute groups increases. In particular, their fairness bound applies only to binary classification with binary-group sensitive attributes and relies on matching distribution means, which is not sufficient to satisfy group fairness metrics. In contrast, we introduce novel bounds based on MI that scale to complex FairDG settings and align directly with the definitions of group fairness metrics. These bounds provide information-theoretic insights that perfectly support algorithm design for FairDG. A detailed comparison with prior theoretical bounds is provided in **Appendix D**.

Problem Formulation

Assumption 1. There exists a domain random variable $\mathbf{D} \sim \text{Categorical}(\{\pi_d\}_{d\in\mathcal{D}})$, where $\mathcal{D}$ contains source domains $\mathcal{D}_S$ with $|\mathcal{D}_S| \geq 2$ and an unseen target domain $d_T \notin \mathcal{D}_S$.

Symbol	Description
$\mathcal{X}$	Input space
$\mathcal{Y}$	Set of class labels
$\mathcal{D}_S$	Set of source domains available during training
$\mathcal{G}$	Set of group memberships (sensitive attribute S)
$\mathbf{X} \in \mathcal{X}$	Random input
$\mathbf{Y} \in \mathcal{Y}$	Class label corresponding to $\mathbf{X}$
$\mathbf{D}_S \in \mathcal{D}_S$	Source domain corresponding to $\mathbf{X}$
$\mathbf{G} \in \mathcal{G}$	Group membership corresponding to $\mathbf{X}$
d_T	Unknown target domain to generalize
x, y, d_S, g	Realizations of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{D}_S$, $\mathbf{G}$

Table 1: Notation table.

Assumption 2. We focus on the same sensitive attribute S and its groups $\mathcal{G}$ when moving from any $d_S \in \mathcal{D}_S$ to d_T.

Domain Generalization: Let $\hat{f}_\theta : \mathcal{X} \to \mathcal{Y}$ be a model parameterized by $\theta \in \Theta$ and let $\mathcal{L}(\cdot)$ denote a loss function. The objective of domain generalization is to find the set of optimal parameters θ_{DG} that satisfies:

$$\theta_{\text{DG}} = \arg\min_{\theta\in\Theta} \mathcal{R}_{d_T}(\hat{f}_\theta), \tag{1}$$

where $\mathcal{R}_{d_T}(\hat{f}_\theta) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim d_T}[\mathcal{L}(\hat{f}_\theta(\mathbf{X}), \mathbf{Y})]$ is the expected risk on an unseen target domain d_T.

Algorithmic Fairness: We consider two group fairness metrics that are conditioned on the true label $\mathbf{Y}$: *Equalized Odds (EOD)* and *Equal Opportunity (EO)*[1]. In this paper, we focus on deriving EOD, since EO is a special case of EOD that only considers the true positive rate, and ***all the theoretical results for EOD can be naturally extended to EO.***

Let $\hat{\mathbf{Y}} = \hat{f}_\theta(\mathbf{X})$ be the model prediction for a random input, and let $\hat{y} \in \mathcal{Y}$ denote its realization. EOD requires that, for any true label $y \in \mathcal{Y}$ and any pair of distinct groups $g, g' \in \mathcal{G}$, the conditional distributions of the predictions—both true and false positive rates—are identical:

$$P(\hat{\mathbf{Y}} \mid \mathbf{Y} = y, \mathbf{G} = g) = P(\hat{\mathbf{Y}} \mid \mathbf{Y} = y, \mathbf{G} = g'),$$

which we denote as $P_{\hat{\mathbf{Y}}|y,\,g} = P_{\hat{\mathbf{Y}}|y,\,g'}$. Equivalently, in probability mass functions, this condition is expressed as: $p(\hat{y} \mid y, g) = p(\hat{y} \mid y, g')\ \forall \hat{y}$. Violations of EOD are measured using the Total-Variation (TV) distance, defined as:

$$\delta_{\text{TV}}(P_{\hat{\mathbf{Y}}|y,\,g}, P_{\hat{\mathbf{Y}}|y,\,g'}) = \frac{1}{2}\sum_{\hat{y}\in\mathcal{Y}} |\,p(\hat{y} \mid y, g) - p(\hat{y} \mid y, g')|.$$

The overall EOD violation averaging across all classes and group pairs for a model $\hat{f}_\theta$ is then given by:

$$\Delta^{\text{EOD}}(\hat{f}_\theta) = C^{\text{D}} \sum_{y\in\mathcal{Y}} \sum_{\{g,g'\}\subset\mathcal{G}} \delta_{\text{TV}}(P_{\hat{f}_\theta(\mathbf{X})|y,\,g}, P_{\hat{f}_\theta(\mathbf{X})|y,\,g'}),$$

where the normalization constant $C^{\text{D}} = \frac{2}{|\mathcal{Y}||\mathcal{G}|(|\mathcal{G}|-1)}$, and the summation is over all unordered unique group pairs $\{g, g'\}$. The objective is to find the optimal parameters θ_{Fair} that minimize the EOD violation on d_T:

$$\theta_{\text{Fair}} = \arg\min_{\theta\in\Theta} \Delta^{\text{EOD}}_{d_T}(\hat{f}_\theta). \tag{2}$$

[1]Other metrics like demographic parity and disparate impact do not consider the correctness of model predictions. Conditioning on the true label allows fairness evaluation via the confusion matrix, better supporting decision-making (Hardt, Price, and Srebro 2016).

Theoretical Bounds

Minimizing Eq. (1) and (2) is challenging as the target domain d_T is unknown. Therefore, we derive their upper bounds and minimize the bounds instead. (See proofs of the theorems and supporting lemmas from **Appendix A** to C.)

Theorem 1 (upper bound for the expected risk on the target domain). Let $\mathcal{L}(\cdot)$ be any non-negative loss function upper bounded[2] by a constant C. Then the expected risk on the target domain d_T satisfies the following upper bound:

$$\mathcal{R}_{d_T}(\hat{f}_\theta) \leq \underbrace{\mathcal{R}_{\mathcal{D}_S}(\hat{f}_\theta)}_{\textbf{Term (1)}} + \underbrace{C \cdot \delta_{TV}(P^{\mathbf{X},\mathbf{Y}}_{d_T}, P^{\mathbf{X},\mathbf{Y}})}_{\textbf{Term (2)}} + \underbrace{\frac{\sqrt{2}C}{2}\sqrt{I\big(\mathbf{D}_S;\, \hat{f}_\theta(\mathbf{X}), \mathbf{Y}\big)}}_{\textbf{Term (3)}}$$

Term (1) is the expected risk of the source domains $\mathcal{D}_S$ available during training. **Term (2)** is the discrepancy between the joint distribution of inputs and labels in the target domain and the mixture distribution, measured by the TV distance. The mixture distribution is computed from the training data as: $P^{\mathbf{X},\mathbf{Y}} = \sum_{d_S \in \mathcal{D}_S} p(d_S) P^{\mathbf{X},\mathbf{Y}}_{d_S}$. However, in DG, the target domain distribution is unknown, making **Term (2)** uncontrollable. **Term (3)** is the MI between the source domain variable $\mathbf{D}_S$ and the joint variable of the model prediction and label $(\hat{f}_\theta(\mathbf{X}), \mathbf{Y})$, which can then be factorized by the chain rule as: $I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X}), \mathbf{Y}) = I(\mathbf{D}_S; \mathbf{Y}) + I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$, where $I(\mathbf{D}_S; \mathbf{Y})$ is constant and can be estimated from the training data.

Takeaways: To minimize $\mathcal{R}_{d_T}(\hat{f}_\theta)$, one should focus on minimizing the ***controllable*** and ***parameterized*** components of the upper bound: $\mathcal{R}_{\mathcal{D}_S}(\hat{f}_\theta)$ and $I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$.

Theorem 2 (upper bound for the EOD violation on the target domain). The EOD violation for multi-class classification with a multi-group sensitive attribute on the target domain d_T satisfies the following upper bound:

$$\Delta^{\text{EOD}}_{d_T}(\hat{f}_\theta) \leq \underbrace{\frac{\sqrt{2I\big(\mathbf{G};\, \hat{f}_\theta(\mathbf{X}) \mid \mathbf{Y}, \mathbf{D}_S\big)}}{|\mathcal{Y}||\mathcal{G}| \min_{y,g} p(y,g)}}_{\textbf{Term (1)}} + \underbrace{\frac{2}{|\mathcal{Y}||\mathcal{G}|}\sum_{y\in\mathcal{Y}}\sum_{g\in\mathcal{G}} \delta_{\text{TV}}\big(P^{\mathbf{X}|y,g}_{d_T},\, P^{\mathbf{X}|y,g}\big)}_{\textbf{Term (2)}} + \underbrace{\frac{\sqrt{2I\big(\mathbf{D}_S;\, \hat{f}_\theta(\mathbf{X}) \mid \mathbf{Y}, \mathbf{G}\big)}}{|\mathcal{Y}||\mathcal{G}| \min_{y,g} p(y,g)}}_{\textbf{Term (3)}}$$

Term (1) is the MI between the group variable $\mathbf{G}$ and the model prediction $\hat{f}_\theta(\mathbf{X})$, conditioned on the joint variable of the label and source domain $(\mathbf{Y}, \mathbf{D}_S)$. The denominator includes $p(y,g)$, the joint probability of observing label y and group g, which is constant and can be estimated from the training data. Similar to **Theorem 1**, in **Term (2)**, $P^{\mathbf{X}|y,g}$ is computed from the training data by $P^{\mathbf{X}|y,g} = \sum_{d_S \in \mathcal{D}_S} p(d_S) P^{\mathbf{X}|y,g}_{d_S}$. However, the target domain distribution is unknown in DG, making **Term (2)** uncontrollable. **Term (3)** measures the MI between the source domain variable $\mathbf{D}_S$ and the model prediction $\hat{f}_\theta(\mathbf{X})$, conditioned on the joint variable of the label and group $(\mathbf{Y}, \mathbf{G})$.

Takeaways: To reduce $\Delta^{\text{EOD}}_{d_T}(\hat{f}_\theta)$, one should focus on minimizing the the two ***parameterized*** conditional MI terms $I\big(\mathbf{G};\, \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S\big)$ and $I\big(\mathbf{D}_S;\, \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{G}\big)$.

An Information-Theoretic View

Combining the two takeaways from **Theorems 1&2**, the FairDG objectives can be summarized as finding the optimal parameter set θ^* that minimizes the following four terms:

$$\begin{aligned}\theta^* = \arg\min_{\theta\in\Theta} \{&\mathcal{R}_{\mathcal{D}_S}(\hat{f}_\theta),\ I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}),\\ &I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S), I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{G})\}.\end{aligned} \tag{3}$$

Theorem 3 (Risk minimization $\Longleftrightarrow$ MI maximization). When $\mathcal{L}(\cdot)$ is the cross-entropy loss, the optimal set θ^* that attain the minimum of $\mathcal{R}_{\mathcal{D}_S}(\hat{f}_\theta)$ (Bayes-optimal) coincides with the set that attains the maximum of $I\big(\hat{f}_\theta(\mathbf{X}); \mathbf{Y}|\mathbf{D}_S\big)$ (Please refer to detailed proof in the **Appendix B**):

$$\arg\min_{\theta\in\Theta} \mathcal{R}_{\mathcal{D}_S}(\hat{f}_\theta) = \arg\max_{\theta\in\Theta} I\big(\hat{f}_\theta(\mathbf{X})\,;\, \mathbf{Y} \mid \mathbf{D}_S\big).$$

Therefore, Eq. (3) can be interpreted entirely in MI terms:

$$\begin{aligned}\theta^* = \arg\max_{\theta\in\Theta} &I(\hat{f}_\theta(\mathbf{X});\, \mathbf{Y} \mid \mathbf{D}_S),\\ \arg\min_{\theta\in\Theta}\{&I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}), I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S),\\ &I(\mathbf{D}_S;\, \hat{f}_\theta(\mathbf{X}) \mid \mathbf{Y}, \mathbf{G})\}.\end{aligned} \tag{4}$$

Theorem 4 (Chain-rule bounds). For the random variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{D}_S$, $\mathbf{G}$, and for any parameter set θ, the MI terms in Eq. (4) satisfy the following inequalities based on chain rules (Please refer to detailed proof in the **Appendix B**):

$$\begin{aligned}I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{G}) \leq\ &I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})\\ &+ I(\mathbf{G};\, \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S),\end{aligned} \tag{5}$$

$$I\big(\hat{f}_\theta(\mathbf{X}); \mathbf{Y}\big) \geq I\big(\hat{f}_\theta(\mathbf{X}); \mathbf{Y}|\mathbf{D}_S\big) - I\big(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}\big), \tag{6}$$

$$\begin{aligned}I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}) \leq\ &I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})\\ &+ I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S).\end{aligned} \tag{7}$$

Eq. (5) shows that for any parameter set θ, the last MI term in Eq. (4) is upper-bounded by the sum of the second and third MI terms. Hence, reducing the second and third MI terms tightens this upper bound and can already help to decrease the last MI term. We can then simplify Eq. (4) to:

$$\begin{aligned}\theta^* = \arg\max_{\theta\in\Theta} &I(\hat{f}_\theta(\mathbf{X});\, \mathbf{Y} \mid \mathbf{D}_S),\\ \arg\min_{\theta\in\Theta}\{&I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}), I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S)\}\end{aligned} \tag{8}$$

[2]For example, the CE loss can be bounded by C by modifying the softmax output from $(p_1, p_2, \cdots, p_{|\mathcal{Y}|})$ to $(\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{|\mathcal{Y}|})$, where $\hat{p}_i = p_i(1 - \exp(-C)|\mathcal{Y}|) + \exp(-C)$, $\forall i \in |\mathcal{Y}|$.

We further show that Eq. (8) indeed provides a solution to the FairDG problem from an information-theoretic view. For DG, Eq. (6) indicates that maximizing $I(\hat{f}_\theta(\mathbf{X}); \mathbf{Y}|\mathbf{D}_S)$ and minimizing $I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$ in the Eq. (8) raises the lower bound of $I(\hat{f}_\theta(\mathbf{X}); \mathbf{Y})$, thereby helping to increase it. This aligns with the goal of DG: by increasing $I(\hat{f}_\theta(\mathbf{X}); \mathbf{Y})$, the model learns to make predictions more informative about the true labels regardless of source domains. As a result, the model becomes source domain-invariant and is better positioned to generalize to unseen target domains. Similarly, for algorithmic fairness, Eq. (7) indicates that minimizing both $I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$ and $I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S)$ in the Eq. (8) reduces the upper bound of $I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$, thereby help to decrease it. This aligns with our goal of algorithmic fairness based on EOD: $I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$ characterizes the MI form of EOD violations regardless of source domains. Minimizing this term encourages the model to produce EOD-consistent predictions that are source domain-invariant, thereby enhancing its ability to generalize the EOD-based algorithmic fairness to unseen target domains.

Proposed Method

Although Eq. (8) offers a theoretical formulation of the FairDG problem from an information-theoretic perspective, the direct computation of the MI terms is impractical as the underlying probability distributions of the involved random variables are unknown. To address this, as shown in Fig. 2, we introduce a practical method designed to approximate and optimize Eq. (8) with finite training data.

First, directly optimizing the predicted label $\hat{f}_\theta(\mathbf{X})$ is infeasible due to non-differentiable discrete operations (e.g., argmax over logits). A common strategy in fair or domain-invariant representation learning is decomposing the $\hat{f}_\theta$ into a feature encoder $\hat{f}_{\theta_E}$ and a classifier $\hat{f}_{\theta_C}$ to enable optimization at the representation level, such that by the data processing inequality the classifier applied after can be fair or domain-invariant (Ganin et al. 2016; Quadrianto, Sharmanska, and Thomas 2019; Dehdashtian, Sadeghi, and Boddeti 2024). As shown in Fig. 2, $\hat{f}_{\theta_E}$ maps inputs to representations $\mathbf{Z}_E = \hat{f}_{\theta_E}(\mathbf{X})$, and the objectives of minimizing $I(\mathbf{D}_S; \hat{f}_\theta(\mathbf{X})|\mathbf{Y})$ and $I(\mathbf{G}; \hat{f}_\theta(\mathbf{X})|\mathbf{Y}, \mathbf{D}_S)$ can be reformulated as minimizing $I(\mathbf{D}_S; \mathbf{Z}_E|\mathbf{Y})$ and $I(\mathbf{G}; \mathbf{Z}_E|\mathbf{Y}, \mathbf{D}_S)$.

However, calculating MI for high-dimensional representations is difficult and often requires approximations like the Mutual Information Neural Estimator (MINE) (Belghazi et al. 2018) or bounds-based methods (Poole et al. 2019). These methods can introduce approximation errors and may rely on unrealistic assumptions (e.g., assuming $\mathbf{Z}_E$ follows a Gaussian distribution). A more practical alternative is to use a differentiable dependence metric that captures both linear and non-linear dependencies, and works well with high-dimensional random vectors. This dependence metric is then used to enforce conditional independence relations $\mathbf{D}_S \perp \mathbf{Z}_E|\mathbf{Y}$ and $\mathbf{G} \perp \mathbf{Z}_E|\mathbf{Y}, \mathbf{D}_S$. Common choices include the Hilbert-Schmidt Independence Criterion (HSIC) (Quadrianto, Sharmanska, and Thomas 2019; Bahng et al. 2020) and Distance Correlation (dCor) (Liu et al. 2022; Zhen et al. 2022). Our experimental results show that dCor consistently outperforms both MINE and HSIC, making it the preferred choice for implementing our method (see Section **Experiments** for detailed discussion). Therefore, the optimization goal becomes minimizing two conditional dCor terms: $\mathrm{dCor}(\mathbf{D}_S, \mathbf{Z}_E|\mathbf{Y})$ and $\mathrm{dCor}(\mathbf{G}, \mathbf{Z}_E|\mathbf{Y}, \mathbf{D}_S)$.

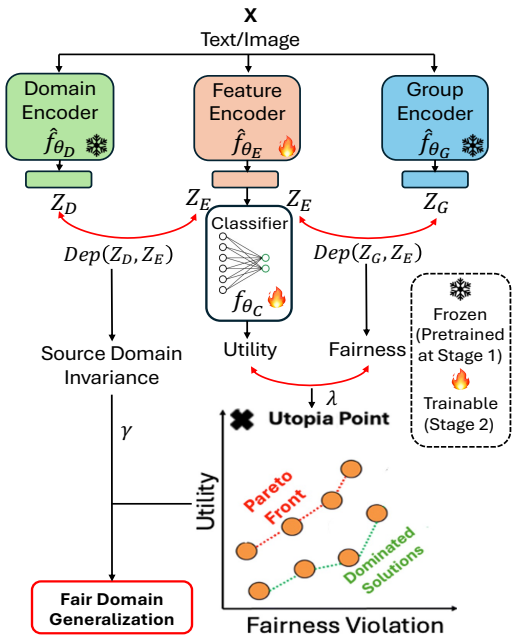

Figure 2: Our proposed method. $\hat{f}_{\theta_D}$ and $\hat{f}_{\theta_G}$ are trained in stage 1 and frozen to produce Z_D and Z_G, which guide $\hat{f}_{\theta_E}$ to learn fair and domain-invariant Z_D by minimizing two dependence terms, where λ controls the utility–fairness trade-off and γ adjusts the source domain invariance.

In real-world settings, $\mathbf{D}_S$ and $\mathbf{G}$ are usually discrete, while $\mathbf{Z}_E$ is continuous. Unlike previous studies that compute dCor directly between discrete and continuous variables (Guo et al. 2022; Zhang et al. 2019). We argue that it is more effective to represent $\mathbf{D}_S$ and $\mathbf{G}$ as continuous vectors because categorical labels fail to capture intra-group similarities in the way latent representations do (Zhen et al. 2022; Bahng et al. 2020). Accordingly, as shown in Fig. 2, we introduce two additional encoders: a domain encoder $\hat{f}_{\theta_D}$ and a group encoder $\hat{f}_{\theta_G}$ to extract domain and group representations $\mathbf{Z}_D = \hat{f}_{\theta_D}(\mathbf{X})$ and $\mathbf{Z}_G = \hat{f}_{\theta_G}(\mathbf{X})$ (we have empirically validated this design in the **Experiments** section). Therefore, our objectives become minimizing $\mathrm{dCor}(\mathbf{Z}_D, \mathbf{Z}_E|\mathbf{Y})$ and $\mathrm{dCor}(\mathbf{Z}_G, \mathbf{Z}_E|\mathbf{Y}, \mathbf{D}_S)$. Given a training set with n samples $\mathcal{D}_{train} = \{(x_i, y_i, d_S^i, g_i)\}_{i=1}^n$, the empirical version of these objectives are:

$$\mathrm{dCor}_n(\hat{f}_{\theta_D}(x_i), \hat{f}_{\theta_E}(x_i)|y) \tag{9}$$

and

$$\mathrm{dCor}_n(\hat{f}_{\theta_G}(x_i), \hat{f}_{\theta_E}(x_i)|y, d_S). \tag{10}$$

Here, dCor_n ranges from 0 to 1, with $\mathrm{dCor}_n = 0$ indi-

cating no observable dependence among the samples. Similar to the empirical risk minimization (ERM) in Eq. (11), dCor_n almost surely converges to the population value as $n \to \infty$ (see Theorem 2 in (Székely, Rizzo, and Bakirov 2007)). Full derivations of Eq. (9) and Eq. (10) are provided in the **Appendix E**. In parallel, as implied by **Theorem 3**, maximizing $I(\hat{f}_\theta(\mathbf{X}); \mathbf{Y}|\mathbf{D}_S)$ can be achieved by minimizing the expected risk over the source domains $\mathcal{R}_{\mathcal{D}_S}(\hat{f}_\theta)$, which reduces to ERM under the training data $\mathcal{D}_{train}$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\hat{f}_{\theta_C}(\hat{f}_{\theta_E}(x_i)), y_i) \tag{11}$$

As prior fairness research shows (Taufiq, Ton, and Liu 2024; Sadeghi, Dehdashtian, and Boddeti 2022; Dehdashtian, Sadeghi, and Boddeti 2024), there is often a trade-off between utility and fairness. Thus, the objectives in Eq. (10) and Eq. (11) may conflict. This means no single optimal set θ^* can simultaneously satisfy θ_{DG} and θ_{Fair}. Instead, the problem should be framed as a multi-objective optimization (MOO) and optimized to yield Pareto-optimal solutions. By combining Eq. (9), Eq. (10), and Eq. (11) with linear scalarization[3], we formulate the empirical objective as:

$$\begin{aligned}\theta^{\mathcal{P}^*} = \underset{\substack{\theta_E\in\Theta_E\\ \theta_C\in\Theta_C\\ \theta_D\in\Theta_D\\ \theta_G\in\Theta_G}}{\arg\min}\ & \underbrace{\frac{1-\lambda}{n}\sum_{i=1}^{n}\mathcal{L}(\hat{f}_{\theta_C}(\hat{f}_{\theta_E}(x_i)), y_i)}_{\textbf{Utility (ERM)}} \\ & + \lambda \underbrace{\mathrm{dCor}_n(\hat{f}_{\theta_G}(x_i), \hat{f}_{\theta_E}(x_i)|y, d_S)}_{\textbf{Fairness (EOD)}} \\ & + \gamma \underbrace{\mathrm{dCor}_n(\hat{f}_{\theta_D}(x_i), \hat{f}_{\theta_E}(x_i)|y)}_{\textbf{Source Domain Invariance}}.\end{aligned} \tag{12}$$

Here, $\lambda \in [0, 1)$ balances the utility-fairness trade-off and γ controls the strength of the regularization for source domain invariance. We set the upper bound $C = 1$ for the loss function $\mathcal{L}(\cdot)$ (CE loss) as described in footnote 2.

Training & Evaluation

Training: A key challenge in optimizing Eq. (11) is training stability, as the framework includes four network components. To address this, as shown in Fig. 2, we adopt a two-stage training procedure. Since we have both domain and group labels in the training set, in the first stage, we train $\hat{f}_{\theta_D}$ and $\hat{f}_{\theta_G}$ by attaching classification heads to predict the source domains and group memberships of training samples. As $\hat{f}_{\theta_D}$ and $\hat{f}_{\theta_G}$ are only used to train $\hat{f}_{\theta_E}$ in a way that it learns to encode $\mathbf{Z}_E$ to be conditionally independent of $\mathbf{Z}_D$ and $\mathbf{Z}_G$, $\hat{f}_{\theta_D}$ and $\hat{f}_{\theta_G}$ are ***discarded at inference time.*** Therefore, obtaining $\hat{f}_{\theta_D}$ and $\hat{f}_{\theta_G}$ are simple in this case as we just need to train them to ***overfit*** the training set so that $\mathbf{Z}_D$ and $\mathbf{Z}_G$ are nearly the optimal representation of $\mathbf{D}_S$ and $\mathbf{G}$ for training samples. In the second stage, we freeze $\hat{f}_{\theta_D}$ and $\hat{f}_{\theta_G}$ and then train $\hat{f}_{\theta_E}$ and $\hat{f}_{\theta_C}$ for the main task.

Another challenge stems from the MOO setting: achieving different utility-fairness trade-offs requires training a new model from scratch for each λ, which is computationally expensive. To address this, we adopt the loss-conditional training strategy proposed in (Dosovitskiy and Djolonga 2019). Instead of training separate models for each λ, we train a single model that conditions on λ during training. Specifically, we sample a range of λ values and train the model on $(\mathbf{X}, \lambda)$ input pairs. This enables the network to adapt its behavior depending on the desired trade-off. So during inference, we just pass different λ to obtain a model tuned for that particular balance between utility and fairness.

Evaluation: We evaluate models with different λ values during testing using fairness metric V (either EO or EOD violations) and utility metric U (accuracy). Let the solution set be $\mathcal{S}_{(V,U)} = \{(V_i, U_i) \mid i = 1, 2, \ldots, N\}$, where each solution corresponds to a model with a different λ. We define the set of solutions that dominate a solution (V_i, U_i) as:

$$\mathcal{D}_{(i)} = \left\{(V_j, U_j) \in \mathcal{S}_{(V,U)} \;\middle|\; (V_j \le V_i) \wedge (U_j \ge U_i)\right\}.$$

The Pareto front $\mathcal{P}$, containing all non-dominated (Pareto optimal) solutions $(V_i, U_i) \in \mathcal{S}_{(V,U)}$, is defined as:

$$\mathcal{P} = \left\{(V_i, U_i) \in \mathcal{S}_{(V,U)} \;\middle|\; \mathcal{D}_{(i)} = \emptyset\right\}.$$

We evaluate both the Pareto front and selected single solution. We use the Hypervolume Indicator (HVI) (Zitzler, Brockhoff, and Thiele 2007) as the evaluation metric to measure both the convergence and diversity of the Pareto front $\mathcal{P}$. HVI measures the area in the solution space dominated by $\mathcal{P}$, bounded by a reference point $R = (V_{\mathrm{ref}}, U_{\mathrm{ref}})$, where $V_{\mathrm{ref}} > V_{\max}$ and $U_{\mathrm{ref}} < U_{\min}$. As the utility and fairness may vary in scale ($\Delta V = V_{\max} - V_{\min}$, $\Delta U = U_{\max} - U_{\min}$), we follow the standard practice (Miettinen 1999; Branke 2008) and normalize both metrics to $[0, 1]$:

$$\mathcal{P}_{\mathrm{norm}} = \left\{\left(\frac{V_i - V_{\min}}{\Delta V}, \frac{U_i - U_{\min}}{\Delta U}\right) \;\middle|\; (V_i, U_i) \in \mathcal{P}\right\}.$$

In $\mathcal{P}_{\mathrm{norm}}$, no solution can have both lower V and higher U. Thus, the solutions can be sorted as $\mathcal{P}_{\mathrm{norm}} = \{(V_1, U_1), \ldots, (V_n, U_n)\}$ with $V_1 < V_2 < \cdots < V_n$ and $U_1 < U_2 < \cdots < U_n$. The HVI is calculated as the non-overlapping rectangular area under $\mathcal{P}_{\mathrm{norm}}$ bounded by R:

$$\mathrm{HVI}(\mathcal{P}_{\mathrm{norm}}) = \sum_{i=2}^{n+1}(V_i - V_{i-1}) \times (U_{i-1} - U_{\mathrm{ref}}), \tag{13}$$

where $V_{n+1} = V_{\mathrm{ref}}$ with higher HVI for a better Pareto front.

For the single solution, we argue that preferences should be set by the human decision makers; thus, we do not assume a preference between objectives by default. A well-known approach for selecting the optimal solution when preference information is not provided is the global criterion method (Zeleny 2012; Hwang and Masud 2012). Let the utopia point $U = (V^*, U^*)$ denote the ideal but typically unreachable solution. In the normalized space $\mathcal{P}_{\mathrm{norm}}$, the optimal solution is the one closest to the utopia point in L_2 distance:

$$(V_{\mathrm{opt}}, U_{\mathrm{opt}}) = \underset{(V_i,U_i)\in\mathcal{P}_{\mathrm{norm}}}{\arg\min}\sqrt{(V_i - V^*)^2 + (U^* - U_i)^2}. \tag{14}$$

[3]Linear scalarization ensures Pareto-optimal solutions, but as both the encoders and the classifier are non-convex, it may not fully characterize the Pareto front (Martinez, Bertran, and Sapiro 2020).

Dataset	CelebA			AffectNet			Jigsaw		
Methods	Acc ↑	EOD ↓	EO ↓	Acc ↑	EOD ↓	EO ↓	Acc ↑	EOD ↓	EO ↓
ERM (Ours)	82.8	14.2	10.5	62.8	15.0	10.1	82.4	17.8	11.1
† ERM+SDI (Ours)	**89.6**	13.5	11.5	**69.8**	15.2	11.1	**87.4**	18.8	11.6
† DANN	88.6	15.5	14.4	66.8	16.5	12.4	84.2	17.7	12.1
† CORAL	87.6	14.6	14.5	67.3	15.6	14.5	83.1	18.6	13.2
† MMD-AAE	88.4	16.5	13.2	68.4	17.5	15.2	84.4	16.8	10.7
† DDG	86.9	17.2	14.2	69.4	18.2	14.2	84.7	18.5	11.9
∗ ERM+Fair (Ours)	79.4	13.8	10.4	59.4	14.8	8.4	81.4	15.5	10.9
∗ LNL	72.4	13.2	10.4	61.4	13.6	9.6	77.4	17.5	7.1
∗ MaxEnt-ARL	71.5	13.6	9.0	61.5	13.8	10.0	78.7	16.5	7.7
∗ FairHSIC	82.4	12.5	9.3	62.4	12.6	9.8	79.4	15.6	9.8
∗ U-FaTE	69.5	14.1	8.2	59.5	14.6	6.4	79.7	16.1	8.7
†∗ FEDORA	85.7	10.6	8.1	65.8	10.2	5.9	84.4	12.6	6.1
†∗ FATDM-StarGAN	85.4	10.9	6.7	65.6	9.8	6.3	85.7	12.3	5.7
†∗ Ours-S	84.3	11.8	5.8	64.5	10.8	6.3	84.6	12.4	4.9
†∗ **Ours**	88.7	**8.2**	**3.1**	68.9	**8.4**	**4.3**	86.3	**9.7**	**3.7**

Table 2: Comparison of existing methods for $(V_{\text{opt}}, U_{\text{opt}})$ evaluation. DG methods are marked with †, fairness-only methods with ∗, and FairDG methods with †∗. Higher Acc (%) reflects better utility, while lower EOD (%) and EO (%) violations indicate better fairness. The best-performing method is shown in bold and underlined; the second-best is underlined. We report only the mean values here; please refer to the **Appendix G** for the full tables including variances.

Experiments

Datasets: Prior works on the FairDG problem use datasets that are either synthetic or tabular (Lin et al. 2024b,a), or restricted to binary classification tasks (Pham, Zhang, and Zhang 2023; Tian et al. 2024). In contrast, real-world FairDG problems may involve high-dimensional data (e.g., text or images), multi-class classification, and multi-group sensitive attributes. We use three datasets that reflect these complexities: **CelebA** (Liu et al. 2015), **AffectNet** (Mollahosseini, Hasani, and Mahoor 2017), and **Jigsaw** (Kivlichan et al. 2020). See details on the datasets in the **Appendix F**.

Implementation Details: For CelebA, we used ResNet18 (He et al. 2016) for all encoders; for AffectNet, we used the Swin Transformer (Base) (Liu et al. 2021); and for Jigsaw, we employed Sentence-BERT (Reimers and Gurevych 2019) for all encoders. All classifiers were implemented as two-layer MLPs. Training was performed using SGD, and the trade-off parameter λ was varied in the range [0, 1) with a step size of 0.01 ($N = 100$). The hyperparameter γ was tuned on the validation set via grid search over $\{0.1, 0.2, 0.4, 0.7, 1\}$. For Pareto front evaluation, we used (1.1, –0.1) as the reference point[4], and (0, 1) as the utopia point. We report the mean and variance of all experimental results over three independent runs. All experiments were done using PyTorch and run on two NVIDIA A100 GPUs. Please refer to **Appendix F** for more implementation details.

Comparisons with Existing Methods: We compared our proposed method against several baselines, including four DG methods: DANN (Ganin et al. 2016), CORAL (Sun and Saenko 2016), MMD-AAE (Li et al. 2018b), and DDG (Zhang et al. 2022). We also evaluated four optimization-based fairness methods capable of producing utility-fairness trade-offs: LNL (Kim et al. 2019), MaxEnt-ARL (Roy and Boddeti 2019), FairHSIC (Quadrianto, Sharmanska, and Thomas 2019), and U-FaTE (Dehdashtian, Sadeghi, and Boddeti 2024). In addition, we included two recent works targeting the FairDG problem: FEDORA (Zhao et al. 2024) and FATDM-StarGAN (Pham, Zhang, and Zhang 2023). Experiments were conducted on the CelebA, AffectNet, and Jigsaw datasets, evaluating both the Pareto front by HVI (%) and the selected single solution (Eq. (13)). ***As DG methods do not consider fairness, they do not yield a Pareto front and are therefore only compared with the single solution.***

As shown in Table 2, DG methods achieve high accuracy on the unseen domain but show large fairness violations. Fairness-only methods improve fairness by sacrificing accuracy, but due to poor robustness to domain shifts, their fairness violations remain high compared to FairDG methods. FairDG approaches outperform fairness-only methods by yielding both higher accuracy and lower fairness violations. As shown in Fig. 3, they also produce higher HVI scores than fairness-only methods. Among FairDG methods, our method consistently achieves the best Pareto front with the highest HVI (%) across datasets and fairness metrics. It also provides a single solution that dominates existing FairDG baselines, delivering large fairness gains with minimal accuracy loss. To evaluate the benefit of our representation-level encoder design, we implemented a variant Ours-S, which computes conditional dCor using $\mathbf{D}_S$ and $\mathbf{G}$ instead of $\mathbf{Z}_D$ and $\mathbf{Z}_G$. Our method consistently outperforms Ours-S, confirming the advantage of our proposed design.

Ablation Studies: As shown in Table 2, we ablate different components of Eq. (11): ERM alone, ERM with the fairness constraint (ERM+Fair), and ERM with the source domain invariance constraint (ERM+SDI). Compared to ERM, ERM+SDI significantly improves target domain accuracy on both datasets, confirming the benefit of enforcing domain invariance. ERM+Fair lowers EO and EOD violations compared to ERM, demonstrating the effectiveness of the fairness constraint, though at the cost of accuracy. The full optimization of Eq. (11) achieves the best trade-off between utility and fairness in the target domain.

Different Dependence Metrics: As shown in Table 3, dCor consistently outperforms HSIC and MINE across all settings. This is because MINE (Belghazi et al. 2018) only approximates the lower bound of MI through the Donsker-Varadhan (DV) representation. Minimizing the lower bound, however, does not guarantee that MI is minimized. Moreover, MINE introduces an inherent approximation error in addition to stochastic error that could be reduced as the sample size increases, whereas dCor and HSIC are only subject to stochastic error. However, HSIC is sensitive to kernel selection and parameter tuning. In contrast, dCor is parameter-free and operates directly on pairwise distances in the original data space, making it robust against kernel distortions.

Impact of the number of source domains: Like other DG methods, our method relies on a set of source domains to generalize to an unseen target domain. To discuss the effect of the number of source domains during training, we split the source domains (perceived age groups) in the AffectNet

[4]The 2D solution space bounded by R is 1.1 × 1.1 = 1.21. Each HVI (Eq. (12)) is normalized to a percentage: $\text{HVI}(\%) = \frac{\text{HVI} \times 10^2}{1.21}$.

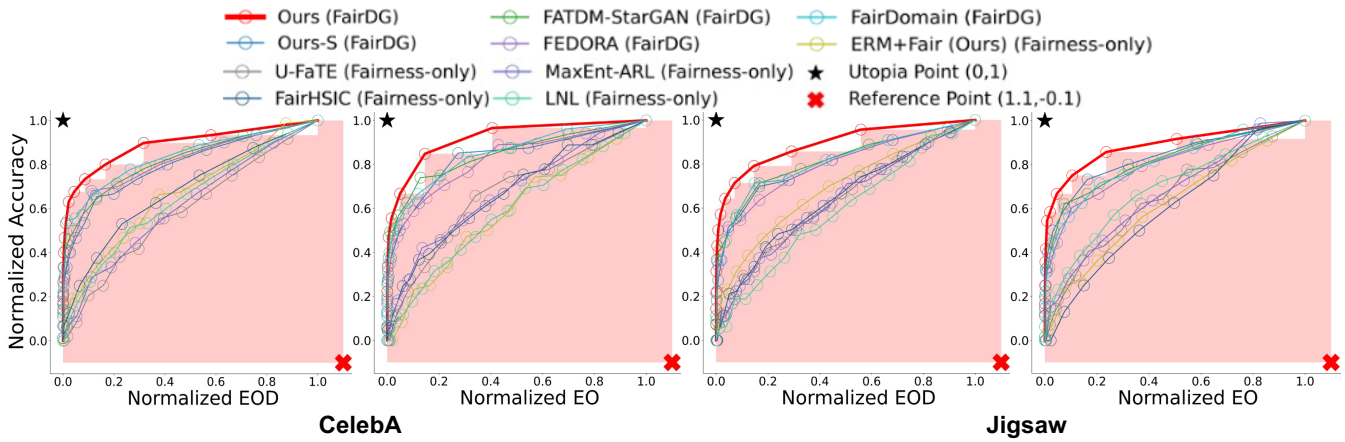

Figure 3: Visualization of Pareto fronts ($\mathcal{P}_{\text{norm}}$) for the fairness-only and FairDG methods. The method with the highest HVI is visualized (the shaded area). Please refer to the **Appendix G** for the complete figure for all datasets and exact HVI values.

Dataset	CelebA		AffectNet		Jigsaw		CelebA			AffectNet			Jigsaw		
Methods	HVI (EOD) ↑	HVI (EO) ↑	HVI (EOD) ↑	HVI (EO) ↑	HVI (EOD) ↑	HVI (EO) ↑	Acc ↑	EOD ↓	EO ↓	Acc ↑	EOD ↓	EO ↓	Acc ↑	EOD ↓	EO ↓
MINE	56.4 ±0.9	59.5 ±1.2	56.2 ±0.7	56.9 ±1.1	57.3 ±0.6	56.5 ±1.0	78.9 ±0.8	13.8 ±0.5	9.2 ±1.3	60.9 ±1.1	14.4 ±0.7	9.3 ±1.4	80.4 ±1.2	15.7 ±0.6	9.3 ±1.0
HSIC	74.2 ±0.6	77.4 ±1.3	74.2 ±0.5	71.4 ±1.2	72.3 ±0.7	73.5 ±0.9	86.8 ±1.0	8.4 ±0.6	3.9 ±1.1	66.8 ±0.8	8.6 ±1.4	6.3 ±0.9	85.3 ±1.2	10.2 ±0.5	4.7 ±1.3
dCor	**75.4** ±1.1	**78.3** ±0.7	**76.4** ±0.8	**74.9** ±1.4	**75.8** ±0.9	**75.7** ±1.2	**88.7** ±1.3	**8.2** ±0.6	**3.1** ±0.7	**68.9** ±0.5	**8.4** ±0.8	**4.3** ±1.1	86.3 ±0.9	**9.7** ±1.4	**3.7** ±0.6

Table 3: Comparison of our method with different dependence metrics for both $\mathcal{P}_{\text{norm}}$ and $(V_{\text{opt}}, U_{\text{opt}})$ evaluations.

Methods	HVI (EOD) ↑	HVI (EO) ↑	Acc ↑	EOD ↓	EO ↓
Ours (2 domains)	75.3 ±0.6	73.5 ±1.0	67.8 ±1.1	8.8 ±0.5	5.2 ±1.3
Ours (3 domains)	76.4 ±0.7	74.9 ±0.9	68.9 ±0.8	8.4 ±1.1	4.3 ±0.6
Ours (5 domains)	**77.3** ±0.8	**75.5** ±1.2	**69.7** ±0.9	**8.1** ±0.7	**3.9** ±1.0

Table 4: Comparisons of our proposed method with different numbers of source domains using the AffectNet dataset.

Methods	HVI (EOD) ↑	HVI (EO) ↑	Acc ↑	EOD ↓	EO ↓
†∗ FEDORA	76.6 ±0.2	73.1 ±0.4	66.7 ±0.6	8.5 ±1.0	4.5 ±0.6
†∗ FATDM-StarGAN	77.9 ±0.8	76.7 ±0.8	67.9 ±0.6	7.5 ±1.1	4.9 ±0.4
†∗ **Ours**	**79.3** ±0.4	**77.1** ±0.1	**68.1** ±0.3	**6.0** ±0.6	**4.1** ±0.5

Table 5: Comparisons of the existing FairDG methods with a different domain split using the AffectNet dataset.

dataset into two and five groups: 30–59, 60–70+ and 30–39, 40–49, 50–59, 60–69, 70+, respectively, in addition to the original setup with three source domains 30–49, 50–69, 70+. The validation and test domains remain unchanged. As shown in Table 4, we observe that increasing the number of source domain splits improves the trade-off achieved by our method. However, splitting domains at a finer granularity may be difficult and require extra human effort.

Robustness of domains splits: In addition to evaluating our method across datasets of different modalities, we further assess its robustness to varying domain splits within the same dataset. We run an experiment on the AffectNet dataset, where the unseen test domain is instead set to the 70+ age group, and the remaining age groups are used for training and validation. As shown in Table 5, our method still consistently outperforms existing FairDG approaches in both the Pareto front and single-solution under this setting.

Conclusion

In this paper, we study the FairDG problem, which aims to minimize both expected risk and fairness violations in unseen target domains. We derive novel upper bounds based on MIs for both the expected risk and fairness violations in multi-class classification tasks with multi-group sensitive attributes, offering key insights from an information-theoretic perspective that inform algorithm design. Guided by these insights, we introduce a practical framework that models the utility-fairness trade-off through Pareto optimization. Experimental results on real-world natural language and vision datasets show that our method outperforms existing approaches, achieving better utility-fairness trade-offs.

Ethical Statement. The sensitive attributes used in this work are either perceived or purely linguistic in meaning. They do not represent any individual's self-identification.

Acknowledgments

Tangzheng Lian is fully funded by the faculty of Natural, Mathematics & Engineering Sciences (NMES) PhD studentship at King's College London.

References

Bahng, H.; Chun, S.; Yun, S.; Choo, J.; and Oh, S. J. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 528–539. PMLR.

Barrainkua, A.; Gordaliza, P.; Lozano, J. A.; and Quadrianto, N. 2025. Preserving the Fairness Guarantees of Classifiers in Changing Environments: A Survey. *ACM Comput. Surv.*, 57(6).

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual Information Neural Estimation. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 531–540. PMLR.

Branke, J. 2008. *Multiobjective optimization: Interactive and evolutionary approaches*, volume 5252. Springer Science & Business Media.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7).

Chen, Y.; Raab, R.; Wang, J.; and Liu, Y. 2022. Fairness transferability subject to bounded distribution shift. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Dehdashtian, S.; Sadeghi, B.; and Boddeti, V. N. 2024. Utility-Fairness Trade-Offs and How to Find Them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12037–12046.

Dosovitskiy, A.; and Djolonga, J. 2019. You only train once: Loss-conditional training of deep networks. In *International conference on learning representations*.

Dunlap, L.; Umino, A.; Zhang, H.; Yang, J.; Gonzalez, J. E.; and Darrell, T. 2023. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, 36: 79024–79034.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Guo, D.; Wang, C.; Wang, B.; and Zha, H. 2022. Learning fair representations via distance correlation minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2): 2139–2152.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hwang, C.-L.; and Masud, A. S. M. 2012. *Multiple objective decision making—methods and applications: a state-of-the-art survey*, volume 164. Springer Science & Business Media.

Jiang, K.; Zhao, C.; Wang, H.; and Chen, F. 2024. Feed: Fairness-enhanced meta-learning for domain generalization. In *2024 IEEE International Conference on Big Data (BigData)*, 949–958. IEEE.

Kim, B.; Kim, H.; Kim, K.; Kim, S.; and Kim, J. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9012–9020.

Kivlichan, I.; Sorensen, J.; Elliott, J.; Vasserman, L.; Görner, M.; and Culliton, P. 2020. Jigsaw Multilingual Toxic Comment Classification. https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification. Kaggle.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Lin, Y.; Li, D.; Zhao, C.; and Shao, M. 2024a. FADE: Towards Fairness-aware Augmentation for Domain Generalization via Classifier-Guided Score-based Diffusion Models. *arXiv preprint arXiv:2406.09495*.

Lin, Y.; Zhao, C.; Shao, M.; Meng, B.; Zhao, X.; and Chen, H. 2024b. Towards counterfactual fairness-aware domain generalization in changing environments. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24. ISBN 978-1-956792-04-1.

Liu, J.; Li, Z.; Yao, Y.; Xu, F.; Ma, X.; Xu, M.; and Tong, H. 2022. Fair Representation Learning: An Alternative to Mutual Information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 1088–1097. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, 6755–6764. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Miettinen, K. 1999. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Palakkadavath, R.; Le, H.; Nguyen-Tang, T.; Gupta, S.; and Venkatesh, S. 2025. Fair Domain Generalization with Heterogeneous Sensitive Attributes Across Domains. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7389–7398. IEEE.

Pham, T.-H.; Zhang, X.; and Zhang, P. 2023. Fairness and Accuracy under Domain Generalization. In *International Conference on Learning Representations.*

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, 5171–5180. PMLR.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8227–8236.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rezaei, A.; Liu, A.; Memarrast, O.; and Ziebart, B. D. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9419–9427.

Roy, P. C.; and Boddeti, V. N. 2019. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2586–2594.

Sadeghi, B.; Dehdashtian, S.; and Boddeti, V. 2022. On Characterizing the Trade-off in Invariant Representation Learning. In *Transactions on Machine Learning Research*.

Shao, M.; Li, D.; Zhao, C.; Wu, X.; Lin, Y.; and Tian, Q. 2024. Supervised algorithmic fairness in distribution shifts: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24. ISBN 978-1-956792-04-1.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3–13.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*, 443–450. Springer.

Székely, G. J.; Rizzo, M. L.; and Bakirov, N. K. 2007. Measuring and testing dependence by correlation of distances.

Taufiq, M. F.; Ton, J.-F.; and Liu, Y. 2024. Achievable Fairness on Your Data With Utility Guarantees. *arXiv preprint arXiv:2402.17106*.

Tian, Y.; Wen, C.; Shi, M.; Afzal, M. M.; Huang, H.; Khan, M. O.; Luo, Y.; Fang, Y.; and Wang, M. 2024. Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision*, 251–271. Springer.

Wang, H.; Hong, J.; Zhou, J.; and Wang, Z. 2023. How robust is your fairness? evaluating and sustaining fairness under unseen distribution shifts. *Transactions on machine learning research*, 2023: https–openreview.

Zeleny, M. 2012. *Multiple criteria decision making Kyoto 1975*, volume 123. Springer Science & Business Media.

Zhang, H.; Zhang, Y.-F.; Liu, W.; Weller, A.; Schölkopf, B.; and Xing, E. P. 2022. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8024–8034.

Zhang, S.; Dang, X.; Nguyen, D.; Wilkins, D.; and Chen, Y. 2019. Estimating feature-label dependence using Gini distance statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 1947–1963.

Zhao, C.; Jiang, K.; Wu, X.; Wang, H.; Khan, L.; Grant, C.; and Chen, F. 2024. Algorithmic fairness generalization under covariate and dependence shifts simultaneously. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4419–4430.

Zhen, X.; Meng, Z.; Chakraborty, R.; and Singh, V. 2022. On the versatile uses of partial distance correlation in deep learning. In *European Conference on Computer Vision*, 327–346. Springer.

Zitzler, E.; Brockhoff, D.; and Thiele, L. 2007. The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration. In *Evolutionary Multi-Criterion Optimization: 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007. Proceedings 4*, 862–876. Springer.