FEED: Fairness-Enhanced Meta-Learning for Domain Generalization

Kai Jiang[1], Chen Zhao[2], Haoliang Wang[1], Feng Chen[1]
[1]*Department of Computer Science, The University of Texas at Dallas, Richardson, Texas, USA*
[2]*Department of Computer Science, Baylor University, Waco, Texas, USA*
{kai.jiang, haoliang.wang, feng.chen}@utdallas.edu, chen_zhao@baylor.edu

***Abstract*—Generalizing to out-of-distribution data with being aware of model fairness is a significant and challenging problem in meta-learning. The goal of this problem is to find a set of fairness-aware invariant parameter of classifier that is trained using data drawn from a family of related training domains with distribution shift on non-sensitive features as well as different levels of dependence between model predictions and sensitive features so that the classifier can achieve good generalization performance on unknown but distinct test domains. To tackle this challenge, existing state-of-the-art methods either address the domain generalization problem but completely ignore learning with fairness, or solely specify shifted domains with various fairness levels. This paper introduces an approach to fairness-aware meta-learning that significantly enhances domain generalization capabilities. Our framework, Fairness-Enhanced Meta-Learning for Domain Generalization (FEED), disentangles latent data representations into content, style, and sensitive vectors. This disentanglement facilitates the robust generalization of machine learning models across diverse domains while adhering to fairness constraints. Unlike traditional methods that focus primarily on domain invariance or sensitivity to shifts, our model integrates a fairness-aware invariance criterion directly into the meta-learning process. This integration ensures that the learned parameters uphold fairness consistently, even when domain characteristics vary widely. We validate our approach through extensive experiments across multiple benchmarks, demonstrating not only superior performance in maintaining high accuracy and fairness but also significant improvements over existing state-of-the-art methods in domain generalization tasks.**

***Index Terms*—Fairness-aware Meta-Learning, Domain Generalization.**

I. Introduction

The widespread adoption of machine learning across various sectors has underscored the critical importance of developing algorithms that can perform well across diverse domains. This challenge, often termed domain generalization, is crucial in environments that differ from the training settings, a common scenario in real-world applications such as healthcare, finance, and social justice. In these applications, not only is high accuracy essential, but fairness cannot be overlooked, especially when sensitive attributes like gender or ethnicity are involved [6].

Recent advancements in domain generalization techniques [2], [12] have aimed at learning domain-invariant features. However, these methods often fail to address changes in distributions of sensitive attributes across domains, leading to potential fairness issues when deployed in varied real-world settings [16].

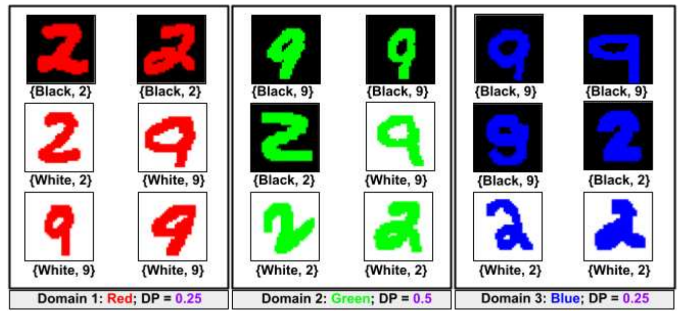

Fig. 1. Illustration of fairness-aware domain generalization problems using the ccMNIST digit dataset. The domains correspond to different digit colors (red/green/blue). Each image has a black or white background color as the sensitive label. For simplicity, digits 2 and 9 are used as toy examples to demonstrate the setting. Each domain is associated with various group fairness levels estimated using the demographic parity metric.

Recent approaches in fairness-aware meta-learning have shown promise in addressing this gap by not only adapting models to new tasks with minimal data but also by potentially incorporating fairness directly into the learning process. However, existing approaches such as [7], [8] focus predominantly on online learning scenarios or specific types of domain shifts, thus limiting their applicability in a broader range of domain generalization contexts.

Different from existing settings for the problem of fairness-aware domain generalization, we define the same problem but in a more general way. We illustrate our setting using the ccMNIST image dataset. In this example, data domains are specified by various digit colors, but *we do not assume group fairness levels for domains are different or the same*. The goal of this problem is to learn an invariant classifier across observed training domains and achieve good generalization performance on testing domains with unknown non-sensitive variation and an unknown group fairness level. To address the challenges, our work introduces a novel fairness-aware meta-learning framework specifically tailored for domain generalization. This framework is designed to learn a robust set of initial parameters that are optimized for effective adaptation across a range of diverse domains, embedding fairness considerations directly at the meta-parameter level. This allows the model to rapidly adapt to new domains while adhering to stringent fairness constraints, thereby pushing the boundaries of traditional domain generalization approaches. Our main following contributions are:

- We introduce a meta-learning framework for fairness across domains, innovatively incorporating fairness at the meta-parameter level. It enables our model maintaining fairness while outperforming traditional domain generalization approaches.
- We formulate a fairness-aware invariance criterion for meta-learning settings. This criterion ensures that the learned initial parameters are consistent and fair across domain shifts, significantly enhancing the fairness level of machine-learning models across various unseen domains.
- We empirically validate our method on multiple domain generalization benchmarks, where it demonstrates the ability to maintain high accuracy and fairness. Our rigorous testing against state-of-the-art domain generalization and fairness methods highlights the critical role of initial parameter in meta-learning for achieving fairness across different domain shifts.

II. Related Work

Fairness-aware domain generalization. Fairness considerations in domain generalization have emerged as a concern due to challenges posed by domain shifts and the unavailability of out-of-distribution (OOD) data, which are traditionally tackled by several leading techniques [2], [9]–[11], [13], [14]. These methods strive to enhance the innate generalizability of machine learning models across source domains, each characterized by distinct but potentially overlapping distributions [22]. A prevalent approach involves aligning distributions across multiple sources to foster domain-invariant feature representations, crucial for stable pattern recognition across domains without target domain data access [23], [24]. Notably, some strategies incorporate meta-learning paradigms to acclimate the model to domain shifts during the training phase [12] or use domain analytic data augmentation techniques to broaden the model's exposure to potential shifts [24].

Despite these advancements, the integration of fairness into domain generalization remains scant. Most research in domain generalization [13], [14], [25], has predominantly focused on leveraging diverse source data to uncover invariant patterns. As [25] articulates, the principal goal is to derive representations that are robust to the marginal distributions of data features, thereby eschewing reliance on target data. However, this line of inquiry largely overlooks the nuances of ensuring that fairness across varying domains. Addressing this gap could enhance the robustness and ethical alignment of models deployed in real-world settings. A recent method for disentangling sensitive attributes, as proposed by Zhao et al. (2024) [15], focuses on learning domain-invariant parameters from training domains. These parameters are fixed and directly applied to new domains. However, the method lacks adaptability when applied to new domains with only a few examples. These methods optimize parameters for multiple domains but are limited in rapidly adapting to unseen tasks. In contrast, our method, based on meta-learning, learns initial parameters that quickly adapt to new domains while ensuring fairness.

Fairness-aware meta-learning. In the context of fairness-aware meta-learning, research efforts primarily focus on developing adaptable frameworks that can effectively handle shifts in domain characteristics while maintaining fairness standards. Strategies such as equality-aware monitoring [26] have been developed. These approaches continuously observe the outputs of a model to detect any deviations from fairness norms and adjust accordingly by modifying the model's parameters or its structure. However, these methods traditionally operate under the assumption that fairness metrics remain consistent across different domains, an assumption often contradicted by the complexities encountered in practical scenarios. Zeng et al. [31] introduced a Nash Bargaining solution to enhance fairness in meta-learning models. However, their approach sometimes struggled with the robustness of fairness across drastic domain shifts due to an overemphasis on bargaining outcomes in homogeneous domains. In contrast, our framework enhances domain generalization by disentangling latent representations into content, style, and sensitive factors, thereby maintaining fairness even when domain characteristics vary significantly. Furthermore, alternative approaches in the literature [16], [17] attempt to evaluate a model's fairness by recognizing changes in fairness benchmarks as indicative of domain shifts, yet they tend to overlook variations in the distribution of non-sensitive attributes, which can lead to inadequate generalization capabilities.

To address these challenges, our meta-learning framework innovatively partitions data attributes into sensitive and non-sensitive categories. Such a distinction is pivotal for the meta-learning algorithm, which is designed not merely to react to explicit domain labels but also to respond to more nuanced shifts in the distributions of data features. This approach enables our meta-learning algorithm to refine its strategy for learning initial parameters, ensuring domain generalization and fairness. By effectively distinguishing between these attribute categories, the algorithm can prioritize the learning of initial parameters that maintain high performance and fairness standards across a spectrum of environments.

III. Preliminaries

Notations. Consider the data space $\mathcal{P}=\mathcal{X}\times\mathcal{Z}\times\mathcal{Y}$, where $\mathcal{X}\subseteq\mathbb{R}^d$ denotes a feature space, $\mathcal{Z}\subseteq\{-1,1\}$ denotes binary sensitive attributes[1], and $\mathcal{Y}\subseteq\{0,1\}$ denotes the binary output space for binary classification. Define the parameterized latent spaces: $\mathcal{C}$ for content, $\mathcal{S}$ for style, and $\mathcal{A}$ for sensitivity factors.

The function $d(\cdot,\cdot)$ is a distance measure across the space $\mathcal{Y}\times\mathcal{Y}$. Variables and parameters in our framework are symbolically denoted as follows: vectors in boldface lowercase letters, and scalars in italic lowercase letters.

Problem setting. Given a dataset $\mathcal{D}$, we consider a set of data domains $\mathcal{E}=\{e_i\}_{i=1}^n$ where each domain corresponds to a distinct data subset $\mathcal{D}^{e_i}=(\mathbf{x}_j^{e_i},z_j^{e_i},y_j^{e_i})_{j=1}^{|\mathcal{D}^{e_i}|}$ over $\mathcal{P}$, and $\mathcal{D}=\bigcup_{i=1,\cdots,n}\mathcal{D}^{e_i}$. Data domains are partitioned

[1] In this study, we focus on a single binary sensitive attribute for clarity. Extensions to multiple sensitive attributes of various categories can be seamlessly integrated.

into multiple training domains $\mathcal{E}_{train} \subsetneq \mathcal{E}$ and testing domains $\mathcal{E}_{test} = \mathcal{E} \setminus \mathcal{E}_{train}$. The corresponding datasets are $\mathcal{D}_{train} = \bigcup_{i=1,\cdots,n_{tr}} \mathcal{D}^{e_i}$ where $e_i \in \mathcal{E}_{train}$ and $\mathcal{D}_{test} = \bigcup_{i=1,\cdots,n_{te}} \mathcal{D}^{e_i}$ where $e_i \in \mathcal{E}_{test}$. Given samples from finite training domains $\mathcal{E}_{train}$, the goal of fairness-aware domain generalization problems is to learn initial parameters $\boldsymbol{\theta} \in \Theta$ of classifier f that is generalizable across all possible domains.

Meta-learning. Define task $\mathcal{T} \sim p(\mathcal{T})$ where $p(\mathcal{T}) = \{(\mathcal{B}^{sup}, \mathcal{B}^{qry}) \mid \mathcal{B}^{sup} \cup \mathcal{B}^{qry} \subseteq \mathcal{D}_{train}, \mathcal{B}^{sup} \cap \mathcal{B}^{qry} = \emptyset\}$. The goal of meta-learning is to learn initial parameters on the training dataset, and it can be quickly adapted to the testing dataset (understream task).

Model-agnostic meta-learning (MAML) [30], as a state-of-the-art approach in the meta-learning landscape, exemplifies a robust framework designed for such tasks. In MAML, a unique set of model parameters $\boldsymbol{\theta}'$ is trained for each task $\mathcal{T}$ on its support set $\mathcal{B}^{sup}$. These parameters are specifically adapted from a shared set of meta-parameters $\boldsymbol{\theta}$, which are iteratively updated based on the aggregate loss observed across all query sets $\mathcal{B}^{qry}$.

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}^{sup}}(\boldsymbol{\theta}), \tag{1}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \sum_{\mathcal{T} \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{B}^{qry}}(\boldsymbol{\theta}') \tag{2}$$

where α is the task-specific learning rate and β denotes the meta-learning rate. $\mathcal{L}_{\mathcal{B}^{sup}}(\boldsymbol{\theta})$ represents the loss calculated on the support set using the initial meta-parameters $\boldsymbol{\theta}$, and $\mathcal{L}_{\mathcal{B}^{qry}}(\boldsymbol{\theta}')$ denotes the loss calculated on the query set using the parameters $\boldsymbol{\theta}'$.

A. Assumptions

Assumption 1 (Latent Spaces)**.** *Given a batch $\mathcal{B} = \{(\mathbf{x}_j^{e_i}, z_j^{e_i}, y_j^{e_i})\}_{j=1}^{|\mathcal{B}|}$ sampled from a specific domain $e_i \in \mathcal{E}$, as illustrated in Fig. 2, we postulate that each data point $\mathbf{x}_j^{e_i}$ within the task originates from:*

- *a latent content factor $\mathbf{c} \in \mathcal{C}$, where $\mathcal{C}$ denotes a content space that is invariant across all domains $\mathcal{E}$;*
- *a latent style factor $\mathbf{s} \in \mathcal{S}$ that is unique to the specific domain e_i;*
- *a latent sensitive factor $\mathbf{a} \in \mathcal{A}$.*

where $\mathcal{C} \cap \mathcal{S} \cap \mathcal{A} = \emptyset$. Each domain e_i is uniquely characterized by its style factors, denoted as $e_i := \mathbf{s}$.

Assumps. 1 echoes the assumptions made in prior works such as [13], [14], [18], [27]. Specifically, UNIT [18] hypothesizes a fully shared latent space across all factors, whereas MUNIT [27] suggests a hybrid latent space model where some components are shared across domains and others are domain-specific. In our framework, considering group fairness, we extend these concepts to include three distinct latent spaces: a content space $\mathcal{C}$, a style space $\mathcal{S}$, and a sensitive space $\mathcal{A}$. Additionally, we posit that domain labels are typically unattainable in both training and testing phases due to practical limitations or excessive costs, as supported by [28].

It is essential for fairness that the labels remain independent of variations across domains. This requirement translates to a scenario where instance conditional distributions $\{\mathbb{P}(Y^{e_i}|X^{e_i}, Z^{e_i})\}_{e_i \in \mathcal{E}}$ differ by domain, reflective of inherent domain-specific characteristics. Within the context of this research, we posit that differences across domains, termed as domain shifts, are governed exclusively by a transformation model $T : \mathcal{X} \times \mathcal{Z} \times \mathcal{E} \to \mathcal{X} \times \mathcal{Z}$. Specifically, if two samples $(\mathbf{x}^{e_i}, z^{e_i})$ and $(\mathbf{x}^{e_j}, z^{e_j})$ from different domains $e_i, e_j \in \mathcal{E}$, where $i \neq j$, exhibit identical content factors, then the sample from domain e_j can be reconstructed from the sample of domain e_i using the transformation T. This process involves T extracting the invariant content from $(\mathbf{x}^{e_i}, z^{e_i})$ and subsequently applying domain-specific style and sensitivity information encoded in e_j to regenerate $(\mathbf{x}^{e_j}, z^{e_j})$.

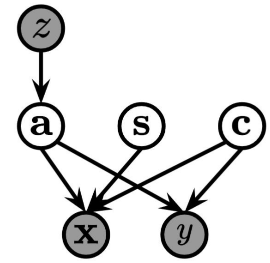

Fig. 2. Causal interpretation of fairness-aware domain generalization tasks. We assume that the raw features ($\mathbf{x}$) and class label (y) of each example are generated by the latent content factor ($\mathbf{c}$), style factor ($\mathbf{s}$), and sensitive factor ($\mathbf{a}$). The sensitive factor ($\mathbf{a}$) is dependent on the sensitive attribute (z) of this example and may or may not be dependent on the domain. The style factor $\mathbf{s}$ depends on the domain, but the content factor $\mathbf{c}$ is independent of the domain e. Each domain label is unobserved.

Assumption 2 (Fairness-aware Domain Invariance)**.** *We hypothesize that the variations observed between domains are primarily driven by changes in the marginal distributions $\mathbb{P}(X^e)$ and $\mathbb{P}(Z^e)$ for each domain $e \in \mathcal{E}$. Consequently, we posit that the conditional distribution $\mathbb{P}(Y^e|X^e, Z^e)$ remains consistent across different domains. With a domain transformation function T, we assert that for any feature vector $\mathbf{x} \in \mathcal{X}$, sensitive attribute $z \in \mathcal{Z}$, and class label $y \in \mathcal{Y}$:*

$$\begin{aligned}&\mathbb{P}(Y^{e_i} = y|X^{e_i} = \mathbf{x}^{e_i}, Z^{e_i} = z^{e_i}) = \mathbb{P}(Y^{e_j} = y|(X^{e_j}, Z^{e_j})\\&= T(\mathbf{x}^{e_i}, z^{e_i}, e_j)) \quad \forall e_i, e_j \in \mathcal{E}, i \neq j\end{aligned}$$

In relation to existing literature, Robey et al. [14] describe a version of T that incorporates content and style factors, but overlooks the sensitive factors which are crucial for ensuring fairness in domain generalization. The domain shift driven by T effectively represents how the distinct distributions $\mathbb{P}(X^{e_i})$ and $\mathbb{P}(Z^{e_i})$ map to the corresponding distributions $\mathbb{P}(X^{e_j})$ and $\mathbb{P}(Z^{e_j})$ in domains. Moreover, it is fundamental in our framework that class labels $y \sim Y$ should remain invariant to changes in fairness-sensitive attributes across domains. In this context, inter-domain variation is exclusively defined by the transformations dictated by T.

Our approach delves into the domain generalization problem, where inter-domain variability is specifically attributed to domain shifts driven by T, representing environmental discrepancies across a collection of marginal distributions $\{\mathbb{P}(X^{e_i}), \mathbb{P}(Z^{e_i})\}_{e_i \in \mathcal{E}}$. Following the assumptions set in Assumps. 2, the generation of data within each domain $e_i \in \mathcal{E}$ is conceptualized through a transformation model T.

To address the challenges of domain-specific variation, our methodology introduces a rigorous definition of invariance, predicated on maintaining fairness across domains as defined by the transformation model T.

Definition 1 (Fairness-aware T-Invariance)**.** *Let T denote the domain transformation model under which a set of classifier parameters $\boldsymbol{\theta} \in \Theta$ is evaluated. A classifier is deemed fairness-aware and domain invariant if:*

$$f(\mathbf{x}^{e_i}, \boldsymbol{\theta}) = f(\mathbf{x}^{e_j}, \boldsymbol{\theta}), \text{ and}$$
$$\mathbb{E}_{\mathbb{P}(X^{e_i}, Z^{e_i}), \mathbb{P}(X^{e_j}, Z^{e_j})}\big[g(X^{e_i}, Z^{e_i}) + g(X^{e_j}, Z^{e_j})\big] = 0$$

is satisfied almost surely, where $(\mathbf{x}^{e_j}, z^{e_j}) = T(\mathbf{x}^{e_i}, z^{e_i}, e_j)$, $\mathbf{x}^{e_i} \sim \mathbb{P}(X^{e_i})$, $\mathbf{x}^{e_j} \sim \mathbb{P}(X^{e_j})$, and $e_i, e_j \in \mathcal{E}$.

Definition 1 establishes the groundwork for ensuring that predictions by f remain consistent across transformations induced by T, affirming the model's adherence to group fairness principles. The intent is that f should uniformly return equivalent predictions for any data instances transformed under T, thereby ensuring the fairness in domain generalization.

IV. Methodology

A. Disentanglement for Fairness-aware Domain Generalization

In our approach to enhance fairness in domain generalization, we leverage a disentanglement strategy. This strategy decomposes the samples into three distinct components: content, style, and sensitive vectors. These content vectors capture domain-invariant features essential for prediction performance, while the style vector encapsulates domain-specific variations that are irrelevant to the labels. The sensitive vector captures the sensitive attributes that could potentially lead to bias. Each sample is decomposed into these three latent vectors, enabling the generation of new samples in a synthetic domain by replacing the style and sensitive vectors with sampled ones, independent of the original domain characteristics. It allows the exploration of a more extensive and varied synthetic domain space, potentially uncovering and mitigating unfair biases that were not explicit in the original data distribution.

We use a transformation model to transfer a sample to a new sample in a synthetic domain by utilizing encoders E^m, E^c and decoders G^i, G^o, which are parameterized by $\boldsymbol{\theta}_m, \boldsymbol{\theta}_c \in \Theta$ and $\boldsymbol{\phi}_i, \boldsymbol{\phi}_o \in \Phi$ respectively. Specifically, when transferring a datapoint to a new datapoint in a synthetic domain, the datapoint is first encoded to a semantic factor $\mathbf{m} \in \mathcal{M}$ through the semantic encoder $E^m : \mathcal{X} \times \Theta \to \mathcal{M}$. The semantic factor $\mathbf{m}$ is further encoded to a content factor $\mathbf{c} \in \mathcal{C}$ through $E^c : \mathcal{M} \times \Theta \to \mathcal{C}$. After sampling a sensitive factor $\mathbf{a} \in \mathcal{A}$ and a style factor $\mathbf{s} \in \mathcal{S}$, two decoders $G^i : \mathcal{C} \times \mathcal{A} \times \Phi \to \mathcal{M}$ and $G^o : \mathcal{M} \times \mathcal{S} \times \Phi \to \mathcal{X}$ are used for generating a new sample in a synthetic domain. Details of learning the transformation model are introduced in Sec. VII-A.

This disentanglement allows the exploration of a more extensive and varied synthetic domain space, potentially uncovering and mitigating unfair biases that were not explicit in the original data distribution. Through this disentanglement, we aim to enhance the parameters' ability to generalize across domains by learning from a richer and more diverse synthetic domain data. It ensures that the learned parameters exhibit robustness to domain shifts and maintain fairness by not carrying over or amplifying biases inherent in the original data.

B. Fairness-aware Meta-Learning

Problem 1 (Meta-Learning for Fairness-aware Domain Generalization)**.** *Given the definitions and assumptions under Definition 1 and Assumps. 2 and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we define the meta-learning problem as follows:*

$$\begin{aligned} \boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} & \sum_{e_i \in \mathcal{E}_{train}} \mathbb{E}_{\mathbb{P}(X^{e_i}, Z^{e_i}, Y^{e_i})} \ell(f(X^{e_i}, \boldsymbol{\theta}^{e_i}), Y^{e_i}) \quad (3) \\ \text{subject to } & f(X^{e_i}, \boldsymbol{\theta}) = f(T(X^{e_i}, Z^{e_i}, e_j), \boldsymbol{\theta}), \\ & \mathbb{E}_{\mathbb{P}(X^{e_i}, Z^{e_i}), \mathbb{P}(X^{e_j}, Z^{e_j})}\big[g(X^{e_i}, Z^{e_i}) \\ & + g(X^{e_j}, Z^{e_j})\big] = 0 \end{aligned}$$

where the inner loop problem is defined as:

$$\begin{aligned} \boldsymbol{\theta}^{e_i} = \operatorname*{argmin}_{\boldsymbol{\theta}'} & \mathbb{E}_{\mathbb{P}(X^{e_i}, Z^{e_i}, Y^{e_i})} \ell(f(X^{e_i}, \boldsymbol{\theta}'), Y^{e_i}) \quad (4) \\ \text{subject to } & f(X^{e_i}, \boldsymbol{\theta}') = f(T(X^{e_i}, Z^{e_i}, e_j), \boldsymbol{\theta}'), \\ & \mathbb{E}_{\mathbb{P}(X^{e_i}, Z^{e_i}), \mathbb{P}(X^{e_j}, Z^{e_j})}\big[g(X^{e_i}, Z^{e_i}) \\ & + g(X^{e_j}, Z^{e_j})\big] = 0 \end{aligned}$$

where $\mathbf{x}^{e_i} \sim \mathbb{P}(X^{e_i})$, $\mathbf{x}^{e_j} \sim \mathbb{P}(X^{e_j})$, $\mathbf{z}^{e_i} \sim \mathbb{P}(Z^{e_i})$, $\forall e_i, e_j \in \mathcal{E}_{train}$, $i \neq j$. $\boldsymbol{\theta}$ is the initialization of $\boldsymbol{\theta}'$ in $\boldsymbol{\theta}^{e_i}$.

The downstream problem is defined as follows:

$$\begin{aligned} \min_{\tilde{\boldsymbol{\theta}}} & \sum_{e_k \in \mathcal{E}_{test}} \mathbb{E}_{\mathbb{P}(X^{e_k}, Z^{e_k}, Y^{e_k})} \ell(f(X^{e_k}, \tilde{\boldsymbol{\theta}}), Y^{e_k}) \quad (5) \\ \text{subject to } & f(X^{e_k}, \tilde{\boldsymbol{\theta}}) = f(T(X^{e_k}, Z^{e_k}, e_l), \tilde{\boldsymbol{\theta}}), \\ & \mathbb{E}_{\mathbb{P}(X^{e_k}, Z^{e_k}), \mathbb{P}(X^{e_l}, Z^{e_l})}\big[g(X^{e_k}, Z^{e_k}) \\ & + g(X^{e_l}, Z^{e_l})\big] = 0 \end{aligned}$$

where $\mathbf{x}^{e_k} \sim \mathbb{P}(X^{e_k})$, $\mathbf{x}^{e_l} \sim \mathbb{P}(X^{e_l})$, $\mathbf{z}^{e_k} \sim \mathbb{P}(Z^{e_k})$, $\forall e_k, e_l \in \mathcal{E}_{test}$, $i \neq j$. $\boldsymbol{\theta}^*$ is the initialization of $\tilde{\boldsymbol{\theta}}$.

The challenges presented in problem 1 arise from the need for meta-learning models. Specifically, the framework conducts meta-training across all the training domains $\mathcal{E}_{train}$, utilizing the breadth of these domains to learn a robust set of initial parameters. However, the true test of generalization and fairness occurs during the subsequent phase, where the meta parameters serve as initial parameters for downstream tasks on a limited subset of samples from the testing domains $\mathcal{E}_{test}$. This problem underscores a significant challenge: ensuring that the model not only adapts to new, unseen domains with very few examples but also maintains consistent and fair performance across the comprehensive domain set $\mathcal{E}$. The sparse availability of samples in $\mathcal{E}_{test}$ compounds this difficulty, demanding that the initial parameters derived from meta-training possess an intrinsic capability to generalize effectively and equitably, even under constrained conditions. A key aspect of tackling this problem involves addressing how closely the data feature distributions in testing domains resemble those in the observed training domains $\mathcal{E}_{train}$. The existing methods on domain generalization [13], [14] incorporate this consideration and introduce solutions primarily focused on decomposing

Algorithm 1 Fairness-Enhanced Meta-Learning for Domain Generalization.

Require: domain transformation model T.
Require: $\{\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_c, \boldsymbol{\theta}_a, \boldsymbol{\theta}_z, \boldsymbol{\phi}_i, \boldsymbol{\phi}_o\}$
Initialize: primal and dual learning rate η_p, η_d, empirical constant γ_1, γ_2.

1: **while** not done **do**
2:     Sample batch of tasks $\mathcal{T} = \{\mathcal{B}^{sup}, \mathcal{B}^{qry}\} \sim p(\mathcal{T})$
3:     **for** each $\mathcal{T}$ **do**
4:         $\boldsymbol{\theta}^{'} = \boldsymbol{\theta}$
5:         $\mathcal{B}^{sup}_{aug} \leftarrow \{T(\mathbf{x}_i, z_i, y_i) \mid (\mathbf{x}_i, z_i, y_i) \in \mathcal{B}^{sup}\}$
6:         $\boldsymbol{\theta}^{'} \leftarrow \text{Adam}(\mathcal{L}(\boldsymbol{\theta}^{'}, \mathcal{B}^{sup}, \mathcal{B}^{sup}_{aug}), \boldsymbol{\theta}^{'}, \eta_p)$
7:         update $\lambda_1^{'}, \lambda_2^{'}$ on $\boldsymbol{\theta}', \mathcal{B}^{sup}, \mathcal{B}^{sup}_{aug}$
8:         $\mathcal{B}^{qry}_{aug} \leftarrow \{T(\mathbf{x}_j, z_j, y_j) \mid (\mathbf{x}_j, z_j, y_j) \in \mathcal{B}^{qry}\}$
9:         calculate $\mathcal{L}(\boldsymbol{\theta}^{'}, \mathcal{B}^{qry}, \mathcal{B}^{sup}_{aug})$
10:     **end for**
11:     $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta_p \cdot \nabla_{\boldsymbol{\theta}} \sum_{\mathcal{T} \sim p(\mathcal{T})} \mathcal{L}(\boldsymbol{\theta}^{'}, \mathcal{B}^{qry}, \mathcal{B}^{qry}_{aug})$
12:     update λ_1, λ_2 on $\boldsymbol{\theta}, \mathcal{B}^{qry}, \mathcal{B}^{qry}_{aug}$
13: **end while**
14: **procedure** $T(\mathbf{x}, z, y)$
15:     $\mathbf{c} = E^c(E^m(\mathbf{x}, \boldsymbol{\theta}_m), \boldsymbol{\theta}_c)$
16:     Sample $\mathbf{a}' \sim \mathcal{N}(0, I_a)$
17:     Sample $\mathbf{s}' \sim \mathcal{N}(0, I_s)$
18:     $\mathbf{x}' = G^o(G^i(\mathbf{c}, \mathbf{a}', \boldsymbol{\phi}_i), \mathbf{s}', \boldsymbol{\phi}_o)$
19:     $z' = h(\mathbf{a}', \boldsymbol{\theta}_z)$
20:     **return** $(\mathbf{x}', z', y)$
21: **end procedure**

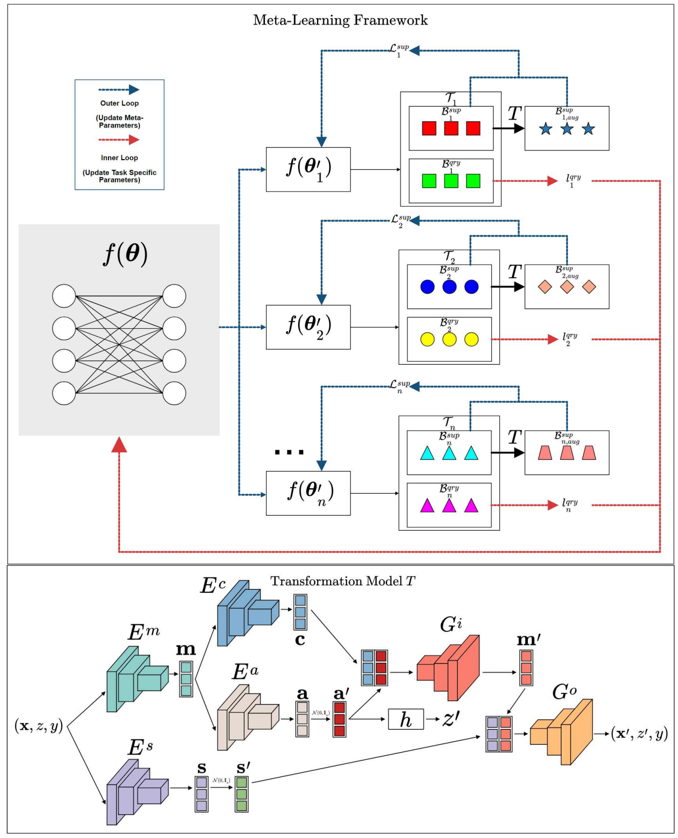

Fig. 3. (Top) An overview of our framework. The red lines and the blue lines correspond to outer loop and inner loop respectively. (Bottom) The transformation model T. It generates an augmented example having the same content factor as the input example but has different style and sensitive factors sampled from their associated distributions that encode a new synthetic domain.

variations in data features across domains into distinct latent spaces. To ensure fairness, data features are categorized into sensitive and non-sensitive components. It is assumed that the dependency of sensitive features on labels might vary across domains, which may not be strictly domain-invariant or domain-specific. This nuanced understanding acknowledges that fairness levels across different domains may differ, enhancing the realism and applicability of the proposed solutions.

Implementation of FEED. Our proposed implementation is shown in Algorithm 1. In lines 14-21, we describe the T 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 content factor as the input example but has different style and sensitive factors sampled from their associated distributions that encode a new synthetic domain as shown in Fig. 3. Line 3-10 show the inner loop updating the task specific parameters for FEED, and line 11-12 show the outer loop udating the meta parameters. In line 6 and line 8, for each example in a data batch $\mathcal{B}$, we apply the procedure T to generate an augmented example from in new synthetic domain. The loss functions are defined in Eqs. (6) to (9) and the Eqs. (10) and (11) show how the hyperparameters λ_1, λ_2 are updated.

Classification loss Given data batch $\mathcal{B} = \{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^{|\mathcal{B}|}$ and classifier f parameterized by $\boldsymbol{\theta}$, the classification loss $\mathcal{L}_{cls}(\boldsymbol{\theta}, \mathcal{B})$ is defined as:

$$\mathcal{L}_{cls}(\boldsymbol{\theta}, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \ell(y_i, f(\mathbf{x}_i, \boldsymbol{\theta})) \tag{6}$$

where we use crossentropy as the distance metric for $d(\cdot)$.

Invariance loss With $\mathcal{B}_{aug}$ whose data points are transformed from $\mathcal{B}$ by T, the invariance loss $\mathcal{L}_{inv}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug})$ is based on the difference between predictions of original and transformed data points:

$$\mathcal{L}_{inv}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug}) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} d[f(\mathbf{x}_i, \boldsymbol{\theta}), f(\mathbf{x}_j, \boldsymbol{\theta})] \tag{7}$$

where $(\mathbf{x}_i, z_i, y_i)$ is a data point from the original data batch $\mathcal{B}$, and $(\mathbf{x}_j, z_j, y_j) = T(\mathbf{x}_i, z_i, y_i)$ is its transformed counterpart in $\mathcal{B}_{aug}$. We consider KL-divergence as the distance metric for $d(\cdot)$.

Fairness loss The fairness loss $\mathcal{L}_{fair}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug})$ is calculated to ensure that the model's predictions remain fairness across different sensitive attributes across source domain and synthetic domain.

$$\begin{aligned}\mathcal{L}_{fair}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug}) = &\left| \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x}_i, z_i) \in \mathcal{B}} g(f(\mathbf{x}_i, \boldsymbol{\theta}), z_i) \right| \\ &+ \left| \frac{1}{|\mathcal{B}_{aug}|} \sum_{(\mathbf{x}_j, z_j) \in \mathcal{B}_{aug}} g(f(\mathbf{x}_j, \boldsymbol{\theta}), z_j) \right|\end{aligned} \tag{8}$$

$$\textit{where} \quad g(f(\mathbf{x}_i, \boldsymbol{\theta}), z_i) = \left| \frac{1}{p_1(1-p_1)} \Big(\frac{z_i + 1}{2} - p_1 \Big) f(\mathbf{x}_i, \boldsymbol{\theta}) \right|$$

where $|\cdot|$ is the absolute function. p_1 is the proportion of samples in group $z = 1$ and correspondingly $1 - p_1$ is the proportion of samples in group $z = -1$.

Total loss The overall loss function $\mathcal{L}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug})$ is then formulated as a weighted sum of the classification, invariance, and fairness losses:

$$\mathcal{L}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug}) = \mathcal{L}_{cls}(\boldsymbol{\theta}, \mathcal{B}) + \lambda_1 \cdot \mathcal{L}_{inv}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug}) + \lambda_2 \cdot \mathcal{L}_{fair}(\boldsymbol{\theta}, \mathcal{B}, \mathcal{B}_{aug}) \tag{9}$$

where λ_1 and λ_2 are hyperparameters that balance the contributions of the invariance and fairness losses, respectively.

The task-specific model parameters are updated using the Adam optimizer, and the meta parameters are updated by using gradient descent on the sum of losses calculated on the query set of each task. Simultaneously, the dual variables λ_1 and λ_2 are adjusted to penalize any violations of the fairness and invariance constraints. The updates for these dual variables are performed as follows:

$$\lambda_1 \leftarrow \max\{[\lambda_1 + \eta_d \cdot (\mathcal{L}_{inv}(\boldsymbol{\theta}, \mathcal{B}) - \gamma_1)], 0\} \tag{10}$$

$$\lambda_2 \leftarrow \max\{[\lambda_2 + \eta_d \cdot (\mathcal{L}_{fair}(\boldsymbol{\theta}, \mathcal{B}) - \gamma_2)], 0\}. \tag{11}$$

where $\gamma_1, \gamma_2 > 0$ are constants.

V. Experiments

We conducted a comprehensive evaluation of our proposed framework, FEED, across a variety of domain generalization benchmarks that encompass both domain characteristics and sensitive attributes. In this assessment, FEED was compared against 11 well-established baselines to illustrate its efficacy. The detailed empirical setup is outlined in Sec. V-A, and the results of these experiments are discussed in Sec. V-B.

A. Experimental Settings

Datasets. We consider four datasets: ccMNIST, FairFace, YFCC100M-DFG, and New York Stop-and-Frisk(NYSF) to evaluate our FEED against state-of-the-art baseline methods, where NYSF is a tabular dataset and the other three are image datasets.

(1) ccMNIST: The ccMNIST is a domain generalization dataset derived from the MNIST dataset [1] by introducing color to the digits and backgrounds. This dataset features images of handwritten digits from 0 to 9, categorized into binary classes with digits 0-4 labeled as 0 and 5-9 labeled as 1, akin to the method used in ColoredMNIST [2]. The ccMNIST includes three distinct domains, each represented by a unique digit color (red, green, blue), encompassing a total of 70,000 images. Notably, each domain exhibits a varying degree of correlation between the class label and the sensitive attribute, the background color, quantified as 0.9, 0.7, and 0 for the red, green, and blue domains, respectively. **(2) FairFace:** The FairFace dataset [3] comprises 108,501 images, portraying a balanced representation across seven racial groups: Black (B), East Asian (E), Indian (I), Latino (L), Middle Eastern (M), Southeast Asian (S), and White (W). For our experimental framework, each racial category is treated as a separate domain, with gender designated as the sensitive attribute and age (either $\geq$ or $<$ 50 years) as the binary class label. **(3) YFCC100M-FDG:** This image dataset, a subset of the YFCC100M [4], curated by *Yahoo Labs*, consists of 90,000 images selected randomly and divided into three domains based on the year of capture: prior to 1999 (d_0), 2000 to 2009 (d_1), and 2010 to 2014 (d_2), with each domain containing 30,000 images. The binary class label is determined by the outdoor or indoor tag of each image, while the latitude and longitude coordinates are translated into a sensitive attribute indicating whether the image was taken in North-America or outside of it. **(4) NYSF:** The NYSF dataset [5] documents police stops in New York City during 2011, focusing on whether pedestrians suspected of weapon possession were indeed carrying a weapon. The data, inherently biased against African Americans, is structured into five sub-city domains: Manhattan (M), Brooklyn (B), Queens (Q), Bronx (R), and Staten (S). Race (black or non-black) is used as the sensitive attribute in this real-world dataset.

Baselines. In our evaluation, the performance of FEED is benchmarked against 15 baseline methods, categorized into three distinct groups based on their primary focus and approach: **(a)** six state-of-the-art *domain generalization* methods, which include ERM [9], IRM [2], GroupDRO [10], Mixup [11], DDG [13], and MBDG [14]; **(b)** three advanced *fairness-aware learning* methods that address variability in environments, namely EIIL [16], FarconVAE [17], and FEDORA [15]; and **(c)** two *naive fairness-aware variants* of existing domain generalization methods, specifically DDG-FC and MBDG-FC, which are adaptations of DDG and MBDG with additional fairness constraints in Eq. (8) integrated into their classification frameworks.

Evaluation metrics. We use three popular evaluation metrics to evaluate the group fairnesses of different methods:

- *Difference in Demographic Parity* (ΔDP) quantifies the disparity in positive prediction rates across groups:

$$\Delta\text{DP} = \Big|\mathbb{P}(\hat{Y} = 1|Z = -1) - \mathbb{P}(\hat{Y} = 1|Z = 1)\Big|$$

  A value of 0 indicates perfect fairness.
- *Difference in Equal Opportunity* (ΔEOPP) measures the difference in true positive rates between groups:

$$\Delta\text{EOPP} = \Big|\mathbb{P}(\hat{Y} = 1|Y = 1, Z = -1) - \mathbb{P}(\hat{Y} = 1|Y = 1, Z = 1)\Big|$$

  A value close to 0 indicates that both groups have an equal chance of receiving a positive outcome.
- *Difference in Equalized Odds* (ΔEO) captures differences across more comprehensive metrics, including both the true positive and false positive rates, ensuring no advantage is given to any group across the decision threshold:

$$\Delta\text{EO} = \frac{1}{2}\Big(\Big|\mathbb{P}(\hat{Y} = 1|Y = 1, Z = -1) - \mathbb{P}(\hat{Y} = 1|Y = 1, Z = 1)\Big|\Big) + \Big(\Big|\mathbb{P}(\hat{Y} = 1|Y = 0, Z = -1) - \mathbb{P}(\hat{Y} = 1|Y = 0, Z = 1)\Big|\Big)$$

  Achieving a ΔEO of 0 is indicative of fair treatment across both outcomes.

TABLE I

PERFORMANCE ON THE NEW-YORK-STOP-AND-FRISK DATASET (BOLD IS THE BEST; UNDERLINE IS THE SECOND BEST).

Methods	Accuracy ↑ / △DP ↓ / △EOPP ↓ / △EO ↓					
	R	B	M	Q	S	**Avg**
ERM [9]	62.97 / 0.032 / 0.019 / 0.028	58.83 / 0.097 / 0.072 / 0.085	62.37 / 0.067 / 0.048 / 0.041	65.01 / 0.070 / 0.077 / 0.072	60.78 / 0.118 / 0.079 / 0.103	61.99 / 0.077 / 0.059 / 0.066
IRM [2]	58.10 / 0.018 / 0.007 / 0.015	56.73 / 0.081 / 0.063 / 0.071	60.68 / 0.038 / 0.024 / 0.030	62.71 / 0.014 / 0.010 / 0.025	59.34 / 0.044 / 0.007 / 0.037	59.51 / 0.039 / **0.022** / 0.036
GroupDRO [10]	62.15 / 0.054 / 0.037 / 0.050	60.08 / 0.103 / 0.080 / 0.088	**62.87** / 0.082 / 0.067 / 0.054	64.55 / 0.073 / 0.079 / 0.076	61.64 / 0.126 / 0.088 / 0.110	62.26 / 0.087 / 0.070 / 0.076
Mixup [11]	63.98 / 0.030 / 0.019 / 0.027	56.30 / 0.073 / 0.055 / 0.065	60.12 / 0.054 / 0.039 / 0.032	65.17 / 0.033 / 0.030 / 0.037	62.96 / 0.101 / 0.085 / 0.086	61.71 / 0.058 / 0.046 / 0.050
DDG [13]	64.78 / 0.020 / 0.010 / 0.025	55.84 / 0.059 / 0.052 / 0.061	59.87 / 0.042 / 0.023 / 0.030	62.97 / 0.028 / 0.011 / 0.039	56.70 / 0.039 / 0.012 / 0.037	60.03 / 0.038 / **0.022** / 0.038
MBDG [14]	62.82 / 0.003 / **0.002** / **0.002**	57.04 / 0.076 / 0.062 / 0.068	61.00 / 0.046 / 0.032 / 0.023	63.39 / **0.001** / 0.016 / 0.018	58.88 / 0.062 / 0.021 / 0.050	60.63 / 0.038 / 0.026 / 0.032
DDG-FC	59.82 / 0.061 / 0.066 / 0.068	56.85 / **0.030** / **0.019** / **0.020**	60.66 / **0.015** / 0.016 / 0.040	57.55 / 0.062 / 0.054 / 0.061	59.95 / **0.018** / **0.002** / **0.008**	58.97 / **0.037** / 0.031 / 0.039
MBDG-FC	62.65 / **0.001** / 0.005 / **0.002**	57.01 / 0.076 / 0.061 / 0.066	60.96 / 0.046 / 0.032 / 0.022	63.38 / **0.001** / 0.016 / **0.017**	58.85 / 0.063 / 0.023 / 0.050	60.57 / **0.037** / 0.027 / 0.032
EIIL [16]	64.36 / 0.019 / 0.021 / 0.019	56.10 / 0.069 / 0.038 / 0.066	60.46 / 0.043 / 0.019 / 0.026	62.82 / 0.013 / 0.019 / 0.020	58.08 / 0.052 / 0.021 / 0.045	60.36 / 0.039 / 0.024 / 0.035
FarconVAE [17]	58.70 / 0.054 / 0.032 / 0.036	**60.80** / 0.076 / 0.027 / 0.039	62.50 / 0.107 / 0.027 / 0.036	65.30 / 0.007 / 0.029 / 0.038	61.20 / 0.056 / 0.027 / 0.031	61.70 / 0.060 / 0.028 / 0.036
FEDORA [15]	63.79 / 0.036 / 0.022 / 0.037	59.19 / 0.132 / 0.117 / 0.113	61.53 / 0.139 / 0.126 / 0.110	62.64 / 0.019 / 0.014 / 0.023	63.19 / 0.076 / 0.087 / 0.065	62.07 / 0.080 / 0.073 / 0.070
FEED (Ours)	**65.81** / 0.006 / 0.003 / 0.006	57.21 / 0.076 / 0.058 / 0.066	60.26 / 0.019 / **0.012** / **0.015**	**65.31** / 0.016 / **0.004** / **0.017**	**63.79** / 0.068 / 0.030 / 0.054	**62.48** / **0.037** / **0.022** / **0.031**

TABLE II

PERFORMANCE ON THE YFCC100M-FDG DATASET (BOLD IS THE BEST; UNDERLINE IS THE SECOND BEST).

Methods	Accuracy ↑ / △DP ↓ / △EOPP ↓ / △EO ↓			
	d_0	d_1	d_2	**Avg**
ERM [9]	89.69 / 0.133 / 0.005 / 0.007	86.92 / 0.049 / 0.005 / 0.017	87.18 / 0.050 / 0.004 / 0.007	87.93 / 0.077 / 0.004 / 0.011
IRM [2]	67.05 / 0.067 / 0.015 / 0.018	65.80 / 0.044 / 0.009 / 0.015	71.01 / 0.040 / 0.002 / 0.012	67.95 / 0.050 / 0.009 / 0.015
GroupDRO [10]	89.20 / 0.138 / **0.001** / 0.026	66.63 / 0.048 / 0.004 / 0.011	85.99 / 0.048 / 0.003 / 0.002	80.61 / 0.078 / 0.002 / 0.013
Mixup [11]	90.00 / 0.130 / **0.001** / **0.004**	86.06 / 0.050 / 0.005 / 0.020	86.70 / 0.049 / 0.002 / 0.007	87.58 / 0.076 / 0.002 / 0.010
DDG [13]	83.74 / 0.093 / 0.032 / 0.067	88.26 / 0.056 / 0.016 / 0.034	89.95 / 0.043 / 0.004 / 0.003	87.32 / 0.064 / 0.018 / 0.035
MBDG [14]	85.70 / 0.136 / 0.029 / 0.024	89.90 / 0.063 / 0.025 / 0.035	87.49 / 0.036 / **0.001** / 0.006	87.70 / 0.079 / 0.019 / 0.022
DDG-FC	86.46 / 0.108 / 0.038 / 0.046	89.32 / 0.067 / 0.030 / 0.038	88.04 / 0.058 / 0.017 / 0.012	87.94 / 0.077 / 0.028 / 0.032
MBDG-FC	**92.12** / **0.057** / 0.032 / 0.154	70.72 / 0.061 / **0.001** / **0.002**	85.56 / 0.054 / **0.001** / 0.008	82.80 / 0.057 / 0.011 / 0.055
EIIL [16]	71.56 / 0.064 / 0.040 / 0.065	68.96 / 0.049 / 0.009 / 0.006	72.20 / 0.042 / **0.001** / **0.001**	70.91 / 0.052 / 0.017 / 0.024
FarconVAE [17]	84.80 / 0.175 / **0.001** / 0.011	72.60 / 0.048 / 0.002 / 0.012	74.50 / 0.071 / 0.004 / 0.012	77.30 / 0.098 / 0.002 / 0.012
FEDORA [15]	87.40 / 0.139 / **0.001** / 0.010	89.50 / **0.020** / 0.002 / 0.008	90.00 / **0.030** / 0.002 / 0.007	88.97 / 0.063 / **0.001** / 0.008
FEED (Ours)	83.96 / 0.060 / **0.001** / 0.008	**91.36** / 0.033 / **0.001** / 0.009	**92.47** / 0.038 / **0.001** / 0.002	**89.26** / **0.044** / **0.001** / **0.006**

TABLE III

PERFORMANCE ON THE FAIRFACE DATASET (BOLD IS THE BEST; UNDERLINE IS THE SECOND BEST).

Methods	Accuracy ↑ / △DP ↓ / △EOPP ↓ / △EO ↓			
	B	E	I	L
ERM [9]	**92.08** / 0.016 / 0.058 / 0.037	92.81 / 0.053 / 0.168 / 0.095	86.94 / 0.041 / 0.109 / 0.064	91.99 / 0.087 / 0.090 / 0.070
IRM [2]	90.78 / **0.001** / 0.022 / **0.001**	68.41 / 0.107 / 0.104 / 0.091	59.31 / 0.037 / 0.074 / 0.047	91.20 / **0.031** / 0.113 / 0.051
GroupDRO [10]	89.78 / 0.022 / 0.086 / 0.053	92.35 / 0.054 / 0.151 / 0.087	89.05 / 0.031 / 0.123 / 0.065	88.59 / 0.062 / 0.060 / 0.042
Mixup [11]	90.46 / 0.003 / 0.040 / 0.021	92.66 / 0.047 / 0.091 / 0.055	89.82 / 0.021 / 0.061 / 0.031	89.66 / 0.054 / **0.056** / **0.039**
DDG [13]	90.49 / 0.026 / 0.009 / 0.007	92.55 / 0.027 / 0.027 / 0.016	89.21 / 0.051 / 0.121 / 0.067	89.13 / 0.047 / 0.081 / 0.047
MBDG [14]	91.84 / 0.041 / 0.073 / 0.040	93.28 / 0.036 / 0.063 / 0.023	88.10 / 0.042 / 0.071 / 0.036	90.31 / 0.055 / 0.074 / 0.041
DDG-FC	90.57 / 0.005 / 0.011 / 0.007	92.62 / 0.003 / **0.013** / **0.008**	90.38 / 0.049 / 0.189 / 0.105	90.97 / 0.074 / 0.187 / 0.113
MBDG-FC	91.12 / 0.032 / 0.056 / 0.038	93.31 / 0.035 / 0.062 / 0.041	87.79 / 0.037 / 0.082 / 0.051	88.77 / 0.032 / 0.077 / 0.049
EIIL [16]	90.71 / 0.038 / 0.050 / 0.032	83.34 / 0.054 / 0.056 / 0.040	83.47 / **0.003** / 0.045 / **0.007**	88.33 / 0.087 / 0.141 / 0.097
FarconVAE [17]	90.30 / 0.032 / 0.092 / 0.053	92.70 / 0.138 / 0.082 / 0.067	87.10 / 0.038 / 0.087 / 0.062	88.30 / 0.109 / 0.088 / 0.058
FEDORA [15]	90.71 / **0.001** / **0.005** / 0.003	**94.72** / 0.032 / 0.156 / 0.083	89.35 / 0.010 / **0.044** / 0.023	**92.56** / 0.035 / 0.110 / 0.059
FEED (Ours)	91.06 / 0.011 / 0.106 / 0.054	94.07 / **0.002** / 0.079 / 0.040	**91.81** / 0.025 / 0.062 / 0.035	91.82 / 0.065 / 0.066 / 0.049

Methods	Accuracy ↑ / △DP ↓ / △EOPP ↓ / △EO ↓			
	M	S	W	**Avg**
ERM [9]	**92.04** / 0.090 / 0.133 / 0.079	89.93 / 0.040 / 0.115 / 0.069	86.73 / 0.083 / 0.192 / 0.104	90.36 / 0.059 / 0.124 / 0.074
IRM [2]	60.95 / **0.049** / 0.043 / 0.032	92.81 / 0.012 / 0.050 / 0.017	89.25 / 0.032 / 0.128 / 0.055	78.96 / 0.038 / 0.076 / 0.042
GroupDRO [10]	89.82 / 0.078 / 0.144 / 0.076	90.73 / 0.031 / 0.040 / 0.030	90.14 / 0.090 / 0.192 / 0.105	90.06 / 0.052 / 0.114 / 0.065
Mixup [11]	89.13 / 0.068 / 0.072 / 0.042	90.19 / 0.034 / 0.017 / 0.021	89.81 / 0.085 / 0.214 / 0.114	90.25 / 0.044 / 0.079 / 0.046
DDG [13]	86.45 / 0.056 / 0.106 / 0.060	90.74 / 0.031 / 0.069 / 0.036	88.89 / 0.112 / 0.218 / 0.129	89.64 / 0.050 / 0.090 / 0.052
MBDG [14]	88.17 / 0.075 / 0.138 / 0.074	91.13 / 0.041 / 0.070 / 0.036	88.26 / 0.056 / 0.109 / 0.056	90.16 / 0.050 / 0.085 / 0.044
DDG-FC	89.00 / 0.107 / 0.210 / 0.130	90.52 / **0.007** / 0.027 / 0.015	88.50 / 0.087 / 0.218 / 0.127	90.37 / 0.047 / 0.122 / 0.072
MBDG-FC	89.40 / 0.066 / 0.045 / 0.036	90.72 / 0.033 / 0.057 / 0.039	89.62 / 0.073 / 0.166 / 0.096	90.10 / 0.044 / 0.078 / 0.050
EIIL [16]	84.77 / 0.137 / 0.122 / 0.113	90.19 / 0.046 / 0.064 / 0.041	86.46 / 0.014 / **0.045** / **0.017**	86.75 / 0.054 / 0.092 / 0.050
FarconVAE [17]	85.30 / 0.154 / 0.092 / 0.066	89.50 / 0.044 / 0.087 / 0.060	86.80 / 0.190 / 0.087 / 0.055	88.57 / 0.101 / 0.088 / 0.060
FEDORA [15]	91.09 / 0.079 / 0.252 / 0.131	93.62 / 0.020 / 0.057 / 0.032	**92.48** / 0.097 / 0.232 / 0.127	**92.08** / 0.039 / 0.122 / 0.065
FEED (Ours)	91.47 / 0.087 / **0.032** / **0.027**	**94.10** / 0.030 / **0.004** / **0.012**	86.89 / **0.010** / 0.105 / 0.053	91.60 / **0.033** / **0.065** / **0.039**

Architectures. In the construction of the semantic encoder E^m and the content encoder E^c, both are designed with four strided convolutional layers, each followed by Instance Normalization [19] and ReLU activation functions, as utilized in various image datasets such as ccMNIST, FairFace, and YFCC100M-FDG [14], [18]. The style encoder E^s and the sensitive encoder E^a are configured with 6 strided convolutional layers, which utilize ReLU activation, succeeded by an adaptive average pooling layer and a trio of fully connected (FC) layers. The architecture for the inner level decoder G^i and the outer level decoder G^o includes an upsampling layer followed by 4 convolutional layers. The sensitive classifier at the inner level incorporates an FC layer equipped with 2 neurons employing a Sigmoid activation function. The outer level discriminator D^o employs a multi-scale structure as proposed by [20] to ensure that G^o yields realistic details and accurate global structure. In contrast, the inner level discriminator D^i is composed of a straightforward FC layer with 112 neurons, activated by ReLU. The stage 2 classifier utilizes a ResNet-50 architecture [19]. For the NYSF dataset, following the

TABLE IV
PERFORMANCE ON THE CCMNIST DATASET (BOLD IS THE BEST; UNDERLINE IS THE SECOND BEST).

Methods	Accuracy ↑ / ΔDP ↓ / ΔEOPP ↓ / ΔEO ↓			
	R	G	B	**Avg**
ERM [9]	98.69 / 0.793 / 0.065 / 0.046	97.68 / 0.393 / 0.014 / 0.012	97.81 / 0.020 / 0.006 / 0.008	98.06 / 0.402 / 0.028 / 0.022
IRM [2]	97.55 / 0.785 / 0.115 / 0.075	97.36 / 0.396 / 0.030 / 0.019	97.14 / 0.021 / 0.009 / 0.009	97.35 / 0.401 / 0.052 / 0.034
GroupDRO [10]	99.03 / 0.800 / 0.085 / 0.052	97.97 / 0.399 / 0.023 / 0.017	97.63 / 0.010 / 0.011 / 0.013	98.21 / 0.403 / 0.040 / 0.027
Mixup [11]	98.92 / 0.796 / 0.050 / 0.045	97.13 / 0.398 / 0.021 / 0.024	97.70 / 0.014 / 0.006 / **0.004**	97.92 / 0.403 / 0.026 / 0.024
DDG [13]	98.99 / 0.794 / 0.040 / 0.039	97.04 / 0.421 / 0.059 / 0.052	97.81 / 0.013 / 0.010 / 0.011	97.95 / 0.409 / 0.036 / 0.034
MBDG [14]	98.87 / 0.787 / 0.036 / 0.025	98.23 / 0.411 / 0.033 / 0.029	**98.75** / 0.017 / 0.006 / **0.004**	**98.62** / 0.405 / 0.025 / 0.019
DDG-FC	98.40 / 0.784 / 0.064 / 0.036	**98.74** / 0.400 / 0.005 / 0.012	97.87 / 0.023 / 0.005 / 0.011	98.33 / 0.403 / 0.025 / 0.020
MBDG-FC	95.74 / 0.867 / 0.360 / 0.380	87.72 / 0.480 / 0.184 / 0.146	79.53 / 0.414 / 0.413 / 0.406	87.66 / 0.587 / 0.319 / 0.311
EIIL [16]	89.65 / 0.999 / 0.999 / 0.999	70.01 / 0.999 / 0.998 / 0.999	55.60 / 0.749 / 0.637 / 0.754	71.75 / 0.916 / 0.878 / 0.917
FarconVAE [17]	94.30 / 0.797 / **0.021** / **0.011**	86.80 / 0.405 / 0.003 / 0.022	93.70 / 0.013 / 0.041 / 0.021	91.60 / 0.405 / 0.022 / 0.018
FEDORA [15]	96.95 / **0.736** / 0.027 / 0.021	98.08 / 0.389 / 0.005 / **0.004**	96.65 / 0.013 / 0.011 / 0.021	97.23 / **0.379** / 0.014 / 0.015
FEED (Ours)	**99.09** / 0.784 / 0.025 / 0.017	97.81 / **0.385** / **0.001** / **0.004**	98.47 / **0.004** / **0.004** / **0.004**	98.46 / 0.391 / **0.010** / **0.008**

guidelines from [17], all networks are exclusively formed from FC layers, including the stage 2 classifier, which comprises 4 FC layers.

Model selection. In our approach to model selection within the domain generalization framework, we adhere to the leave-one-domain-out validation criteria, a methodology supported by [14] and identified as one of the three prominent methods by [21]. This involves evaluating FEED on a training domain that is withheld during the training process and averaging the performance across the remaining $|\mathcal{E}_{train}| - 1$ domains.

B. Results

Quantitative results. For all tables in the paper, the results shown in each column represent performance on the test domain, using the rest as training domains.

Our method FEED demonstrates superior performance in maintaining fairness across different datasets, significantly outperforming both traditional domain generalization methods and state-of-the-art fairness-aware approaches. For instance, in the York-Stop-and-Frisk dataset (Table I), FEED achieves top fairness metrics (0% for ΔDP, 0% for ΔEOPP, and 0.1% for ΔEO) and shows a notable accuracy improvement of 0.22% over the best baseline. This trend is consistently observed across other datasets as well.

In the YFCC100M-FDG dataset (Table II), FEED not only upholds the highest fairness levels (0.6% for ΔDP, 0% for ΔEOPP, 0.2% for ΔEO) but also achieves a comparable accuracy improvement of 0.29%. These results underline the effectiveness of FEED in handling domain-specific variations while ensuring robust fairness across domains.

The datasets such as ccMNIST and NYSF further validate FEED's performance. For the FairFace dataset (Table III), our method reports better fairness metrics (0.5% for ΔDP, 1.1% for ΔEOPP, 0.3% for ΔEO) with a slight trade-off in accuracy (0.48% lower than the best baseline). Similarly, in the ccMNIST dataset (Table IV), FEED maintains competitive fairness metrics and accuracy, demonstrating its adaptability and efficiency across varying experimental settings.

Our observations indicate that FEED consistently delivers strong performance on fairness metrics while maintaining competitive accuracy, affirming its potential for widespread applicability in real-world settings that demand fairness outcomes across diverse populations. This consistent performance is particularly notable in the context of challenging datasets such as York-Stop-and-Frisk and YFCC100M-FDG, where FEED excels in achieving top-tier results in fairness, a critical quality for models deployed in sensitive applications.

The analysis extends to datasets like ccMNIST and NYSF, FEED shows only marginal discrepancies in accuracy, yet continues to uphold superior fairness metrics. This ability to balance fairness with accuracy underpins the versatility of FEED, making it a robust solution for scenarios that extend beyond traditional domain applications. Moreover, the integration of FEED with domain-specific requirements showcases its adaptability and readiness to tackle the intrinsic variability and unpredictability of real-world data.

In conclusion, FEED stands out as a formidable framework in the landscape of domain generalization and fairness-aware meta-learning, offering significant improvements over both conventional and state-of-the-art methods. Its dual strengths in maintaining high classification accuracy while excelling in fairness across varied domains position FEED as a transformative tool for deploying robust and fairness model in diverse real-world settings. This generalizability, coupled with the method's inherent flexibility to adapt to various data characteristics and domain shifts.

Ablation studies. We conducted two ablation studies. (1) The difference between the FEED and the first ablation study (Abs1, w/o inner loop) is that the latter does not update the task-specific parameters based on the support set for the inner loop. In other words, the meta-parameters are directly updated based on the query loss which is calculated based on the meta-parameters. Without updating the task-specific parameters, it makes the ablation study hard to train good initial parameters, leading to poor generalization performance. Experimental results show that the first ablation study performs worse than FEED on all four datasets on both accuracy and fairness metrics. (2) The second study (Abs2, w/o augment) does not use the transformation model T to generate augmented support set and augmented query set. The parameters are updated only based on the support set and the query set. Similar to Abs1, without generating the augmented support set and the augmented query set in synthetic domains, it is much harder to learn good initial parameters. Our results demonstrate that Abs2 performs worse on all the datasets. We include the performance of such ablation studies in Fig. 4.

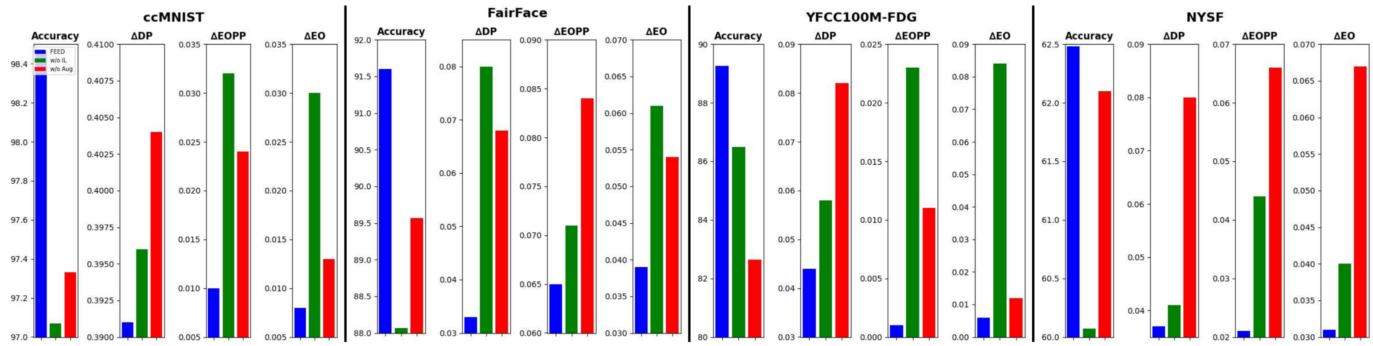

Fig. 4. Ablation study on four datasets. Results are plotted as averages across all domains.

VI. Conclusion

In this paper, we have introduced a novel framework for fairness-aware meta-learning aimed at enhancing domain generalization across diverse environments. By disentangling latent factors into content, style, and sensitive vectors, our approach ensures that the fairness, even in the face of domain shifts. The proposed fairness-aware invariance criterion plays a crucial role in maintaining fairness across different domains.

Our extensive experimental evaluation demonstrates that the proposed method not only achieves superior accuracy but also significantly improves fairness compared to existing state-of-the-art approaches. These results underscore the importance of incorporating fairness considerations into domain generalization frameworks.

Future work will explore the extension of our framework to handle multiple sensitive attributes and its application to more complex, real-world datasets. We aim to investigate the integration of our method with other fairness-aware learning paradigms to further enhance its fairness and generalizability.

References

[1] LeCun, Yann and Bottou, Léon and Bengio, Yoshua and Haffner, Patrick, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] Arjovsky, Martin and Bottou, Léon and Gulrajani, Ishaan and Lopez-Paz, David, “Invariant risk minimization,” arXiv preprint arXiv:1907.02893, 2019.

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

[4] Thomee, Bart and Shamma, David A and Friedland, Gerald and Elizalde, Benjamin and Ni, Karl and Poland, Douglas and Borth, Damian and Li, Li-Jia, “YFCC100M: The new data in multimedia research,” Communications of the ACM, vol. 59, no. 2, pp. 64–73, 2016.

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

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

[7] Zhao, Chen and Chen, Feng and Thuraisingham, Bhavani, “Fairness-Aware Online Meta-learning,” ACM SIGKDD, 2021.

[8] Zhao, Chen and Mi, Feng and Wu, Xintao and Jiang, Kai and Khan, Latifur and Chen, Feng, “Adaptive Fairness-Aware Online Meta-Learning for Changing Environments,” Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2565–2575, 2022.

[9] Vapnik, Vladimir, “The nature of statistical learning theory,” Springer science & business media, 1999.

[10] Sagawa, Shiori and Koh, Pang Wei and Hashimoto, Tatsunori B and Liang, Percy, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” arXiv preprint arXiv:1911.08731, 2019.

[11] Yan, Shen and Song, Huan and Li, Nanxiang and Zou, Lincan and Ren, Liu, “Improve unsupervised domain adaptation with mixup training,” arXiv preprint arXiv:2001.00677, 2020.

[12] Li, Da and Yang, Yongxin and Song, Yi-Zhe and Hospedales, Timothy, “Learning to generalize: Meta-learning for domain generalization,” Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.

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

[14] Robey, Alexander and Pappas, George J and Hassani, Hamed, “Model-based domain generalization,” Advances in Neural Information Processing Systems, vol. 34, pp. 20210–20229, 2021.

[15] C. Zhao, K. Jiang, X. Wu, H. Wang, L. Khan, C. Grant, and F. Chen, “Algorithmic Fairness Generalization under Covariate and Dependence Shifts Simultaneously,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 4419–4430.

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

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

[18] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[20] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.

[21] I. Gulrajani and D. Lopez-Paz, “In search of lost domain generalization,” 2020, arXiv:2007.01434. [Online]. Available: https://arxiv.org/abs/2007.01434

[22] R. Volpi, D. Larlus, and G. Rogez, “Continual adaptation of visual representations via domain randomization and meta-learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4443–4453.

[23] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.
[24] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "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, 2020, pp. 561–578.
[25] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in Neural Information Processing Systems*, vol. 24, 2011.
[26] S. Alonso et al., "Ordering artificial intelligence based recommendations to tackle the sdgs with a decision-making model based on surveys," *Sustainability*, vol. 13, no. 11, pp. 6038, 2021, MDPI.
[27] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.
[28] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 1929–1938.
[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[30] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, PMLR, 2017, pp. 1126–1135.
[31] Y. Zeng, X. Yang, L. Chen, C. C. Ferrer, M. Jin, M. I. Jordan, and R. Jia, "Fairness-Aware Meta-Learning via Nash Bargaining," arXiv preprint arXiv:2406.07029, 2024.

VII. APPENDIX.

A. Transformation Model Training

Our proposed framework involves disentangling an input sample from training domains into three factors in distinct latent spaces, using a series of encoders $E = \{E^m, E^s, E^c, E^a\}$ and decoders $G = \{G^i, G^o\}$. These are parameterized respectively by $\boldsymbol{\theta}_m, \boldsymbol{\theta}_s, \boldsymbol{\theta}_c, \boldsymbol{\theta}_a \in \Theta$ and $\boldsymbol{\phi}_i, \boldsymbol{\phi}_o \in \Phi$. The framework operates through two hierarchical levels: an outer level and an inner level, each with its own auto-encoder.

In the outer level, an input datapoint undergoes encoding into a semantic factor $\mathbf{m} \in \mathcal{M}$ and a style factor $\mathbf{s} \in \mathcal{S}$, achieved via the encoders $E^m : \mathcal{X} \times \Theta \to \mathcal{M}$ and $E^s : \mathcal{X} \times \Theta \to \mathcal{S}$. Progressing to the inner level, the semantic factor $\mathbf{m}$ is further decomposed into a content factor $\mathbf{c} \in \mathcal{C}$ and a sensitive factor $\mathbf{a} \in \mathcal{A}$ through the encoders $E^c : \mathcal{M} \times \Theta \to \mathcal{C}$ and $E^a : \mathcal{M} \times \Theta \to \mathcal{A}$. The corresponding decoders in these levels are $G^i : \mathcal{C} \times \mathcal{A} \times \Phi \to \mathcal{M}$ for the inner level and $G^o : \mathcal{M} \times \mathcal{S} \times \Phi \to \mathcal{X}$ for the outer level, facilitating the reconstruction of the original data. Inspired by image-to-image translation in computer vision [18], [27], Our total loss function of learning such encoders and decoders comprises three components: a bidirectional reconstruction loss, a sensitive label prediction loss, and an adversarial loss.

Reconstruction loss Considering a datapoint $\mathbf{x}$ sampled from $p(\mathbf{x})$, encoders and decoders in outer loop are able to reconstruct it by minimizing the reconstruction loss:

$$\mathcal{L}^x_{recon} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \left[\| G^o \left(\hat{\mathbf{m}}, E^s(\mathbf{x}) \right) - \mathbf{x} \|_1 \right]$$

where $\hat{\mathbf{m}} = G^i(\mathbf{c}, \mathbf{a}) = G^i \left(E^c(E^m(\mathbf{x})), E^a(E^m(\mathbf{x})) \right)$. For the inner level, the semantic factor $\mathbf{m} = E^m(\mathbf{x})$ encoded from the outer level is required to be reconstructed:

$$\mathcal{L}^{m_d}_{recon} = \mathbb{E}_{\mathbf{m}\sim p(\mathbf{m})} \left[\left\| G^i \left(E^c(\mathbf{m}), E^a(\mathbf{m}) \right) - \mathbf{m} \right\|_1 \right]$$

with $p(\mathbf{m})$ determined by the mapping $\mathbf{m} = E^m(\mathbf{x})$ and $\mathbf{x} \sim p(\mathbf{x})$.

The latent factors $\mathbf{c}, \mathbf{s}, \mathbf{a}$, extracted from the datapoint $\mathbf{x}$ are encouraged to be reconstructed through some latent factors randomly sampled from the prior distributions.

$$\mathcal{L}^c_{recon} = \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{a}\sim \mathcal{N}(0, \mathbf{I}_a)} \left[\left\| E^c \left(G^i(\mathbf{c}, \mathbf{a}) \right) - \mathbf{c} \right\|_1 \right]$$
$$\mathcal{L}^a_{recon} = \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{a}\sim \mathcal{N}(0, \mathbf{I}_a)} \left[\left\| E^a \left(G^i(\mathbf{c}, \mathbf{a}) \right) - \mathbf{a} \right\|_1 \right]$$

where $p(\mathbf{c})$ is given by $\mathbf{c} = E^c(E^m(\mathbf{x}))$, and $\mathbf{a} = E^a(E^m(\mathbf{x}))$. Considering the dual-role of $\mathbf{m}$, as both a latent factor from the inner level and an input to the outer level, $\mathbf{s}$ can be reconstructed by two reconstruction losses:

$$\mathcal{L}^{s_{in}}_{recon} = \mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}), \mathbf{s}\sim \mathcal{N}(0, \mathbf{I}_s)} \left[\| E^s(G^o(\mathbf{m}, \mathbf{s})) - \mathbf{s} \|_1 \right]$$
$$\mathcal{L}^{s_{out}}_{recon} = \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{s}\sim \mathcal{N}(0, \mathbf{I}_s), \mathbf{a}\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_a)} [\| E^s \left(G^o(G^i(\mathbf{c}, \mathbf{a}), \mathbf{s}) \right) - \mathbf{s} \|_1]$$

and for reconstructing $\mathbf{m}$ as a latent factor:

$$\mathcal{L}^{m_f}_{recon} = \mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}), \mathbf{s}\sim \mathcal{N}(0, \mathbf{I}_s)} \left[\| E^m \left(G^o(\mathbf{m}, \mathbf{s}) \right) - \mathbf{m} \|_1 \right]$$

The reconstruction loss is defined as follows:

$$\mathcal{L}_{recon} = \mathcal{L}^x_{recon} + \mathcal{L}^{m_d}_{recon} + \mathcal{L}^c_{recon} + \mathcal{L}^a_{recon} + \mathcal{L}^{s_{in}}_{recon} + \mathcal{L}^{s_{out}}_{recon} + \mathcal{L}^{m_f}_{recon}$$

Sensitive prediction loss The sensitive attributes encoded from the datapoint $\mathbf{x}$ underpin the training of a classifier $h : \mathcal{A} \times \Theta \to \mathcal{Z}$. This classifier is then employed to predict the sensitive label associated with the attribute vector $\mathbf{a}$. Specifically, the prediction is formulated as:

$$\hat{\mathbf{z}} = h(\mathbf{a}, \boldsymbol{\theta}_z) = h(E^a(E^m(\mathbf{x})), \boldsymbol{\theta}_z)$$
$$\mathcal{L}^z_{cls} = \text{CrossEntropy}(\mathbf{z}, \hat{\mathbf{z}})$$

Adversarial loss Inspired by the effectiveness of Generative Adversarial Networks (GANs) [29], define discriminators $D = \{D^i, D^o\}$, where $D^o : \mathcal{X} \times \Psi \to \mathbb{R}$ is the discriminator for the outer level, parameterized by $\boldsymbol{\psi}_o \in \Psi$, and $D^i : \mathcal{M} \times \Psi \to \mathbb{R}$ is the discriminator for the inner level, parameterized by $\boldsymbol{\psi}_i \in \Psi$. The discriminators are tasked with differentiating between real and constructed data with random factors.

$$\begin{aligned}\mathcal{L}^x_{GAN} = &\mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{s}\sim \mathcal{N}(0, \mathbf{I}_s), \mathbf{a}\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_a)} \big[\log \big(1 \\ &- D^o(G^o(\hat{\mathbf{m}}, \mathbf{s})) \big) \big] + \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \big[\log D^o(\mathbf{x}) \big] \\ &+ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{s}\sim p(\mathbf{s}), \mathbf{a}\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_a)} \big[\log \big(1 - D^o(G^o(\hat{\mathbf{m}}, \mathbf{s})) \big) \big] \\ &+ \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \big[\log D^o(\mathbf{x}) \big] + \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{s}\sim \mathcal{N}(0, \mathbf{I}_s), \mathbf{a}\sim p(\mathbf{a}))} \big[\log \big(1 \\ &- D^o(G^o(\hat{\mathbf{m}}, \mathbf{s})) \big) \big] + \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \big[\log D^o(\mathbf{x}) \big]\end{aligned}$$

where $\hat{\mathbf{m}}$ is as defined in $\mathcal{L}^x_{recon}$.

$$\begin{aligned}\mathcal{L}^m_{GAN} = &\mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{a}\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_a)} \big[\log \big(1 - D^i(G^i(\mathbf{c}, \mathbf{a})) \big) \big] \\ &+ \mathbb{E}_{\mathbf{m}\sim p(\mathbf{m})} \big[\log D^i(\mathbf{m}) \big]\end{aligned}$$

The adversarial loss is defined as:

$$\mathcal{L}_{GAN} = \mathcal{L}^x_{GAN} + \mathcal{L}^m_{GAN}$$

Total loss We jointly train the encoders, decoders, and discriminators to optimize the final objective:

$$\begin{aligned}&\min_{E,G} \max_{D} \mathcal{L}_{total}(E, G, D) \\ &= \mathcal{L}_{recon} + \beta_z \mathcal{L}^z_{cls} + \beta_g \mathcal{L}_{GAN}\end{aligned}$$

The $\beta_z, \beta_g > 0$ modulate the relative significance of each loss term within this formula.