MultiFair: Model Fairness With Multiple Sensitive Attributes

Huan Tian, Bo Liu, *Senior Member, IEEE*, Tianqing Zhu, Wanlei Zhou, *Senior Member, IEEE*, and Philip S. Yu, *Life Fellow, IEEE*

***Abstract*— While existing fairness interventions show promise in mitigating biased predictions, most studies concentrate on single-attribute protections. Although a few methods consider multiple attributes, they either require additional constraints or prediction heads, incurring high computational overhead or jeopardizing the stability of the training process. More critically, they consider per-attribute protection approaches, raising concerns about fairness gerrymandering where certain attribute combinations remain unfair. This work aims to construct a neutral domain containing fused information across all subgroups and attributes. It delivers fair predictions as the fused input contains neutralized information for all considered attributes. Specifically, we adopt mixup operations to generate samples with fused information. However, our experiments reveal that directly adopting the operations leads to degraded prediction results. The excessive mixup operations result in unrecognizable training data. To this end, we design three distinct mixup schemes that balance information fusion across attributes while retaining distinct visual features critical for training valid models. Extensive experiments with multiple datasets and up to eight sensitive attributes demonstrate that the proposed MultiFair method can deliver fairness protections for multiple attributes while maintaining valid prediction results.**

***Index Terms*— Deep learning, domain adaptation, fairness, image processing, multiple sensitive features.**

NOMENCLATURE

γ	Fairness metric.
x, X	Training data.
s^i	ith sensitive attribute.
x^m	Mixed images.
x^i_{sel}	Selected image for the ith attribute
f_{inter}	Interpolation function for x_c.
V	Attribute value vectors.
ϵ	Prediction errors.
u	Extracted features.
e	Domains.
Γ	Discrimination level.
y, $\hat{y}$	Truth and predicted labels.
s_0 and s_1	Subgroup with attribute values of 0, 1.
s^i_x	ith attribute value of x.
$x^i_{s_1}$	Sample with the ith attribute of an attribute value 0.
λ	Weighting parameter.
g	Feature extraction functions.
$\bar{\epsilon}$	Prediction errors for mixed data.
$\bar{u}$	Extracted features for mixed data.
D	Distance measurement.

Manuscript received 8 December 2022; revised 17 August 2023 and 23 February 2024; accepted 25 March 2024. Date of publication 22 April 2024; date of current version 1 March 2025. This work was supported in part by Australian Research Council (ARC) under Grant DP230100246 and Grant DP240100955 and in part by the National Science Foundation (NSF) under Grant III-2106758 and Grant POSE-2346158. *(Corresponding author: Tianqing Zhu.)*

Huan Tian and Bo Liu are with the Centre for Cyber Security and Privacy and the School of Computer Science, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: Huan.Tian@student.uts.edu.au; Bo.Liu@uts.edu.au).

Tianqing Zhu and Wanlei Zhou are with the Faculty of Data Science, City University of Macau, Macau (e-mail: tqzhu@cityu.edu.mo; wlzhou@cityu.edu.mo).

Philip S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA.

Digital Object Identifier 10.1109/TNNLS.2024.3384181

I. INTRODUCTION

FAIRNESS interventions have drawn increasing attention as machine learning models may generate discriminative predictions toward certain subgroups of sensitive attributes [1], [2], [3]. For example, studies have found trained facial recognition models exhibit differing performance across gender subgroups [4], [5]. To mitigate this issue, methods, such as fair constraints [5], [6], [7] or adversarial learning [4], [5], have been proposed. However, most studies focus on protecting single sensitive attributes. In this work, we aim to provide an efficient approach to enhance model fairness performance for multiple sensitive attributes simultaneously.

Two main strategies have emerged for fairness interventions in deep classifiers: fair constraint methods and adversarial learning methods. Fair constraint methods formulate fairness as optimization objectives [7], [8]. They measure prediction gaps across subgroups and minimize these gaps during training to enhance model fairness performance. Adversarial methods aim to remove sensitive attribute information from learned representations [4], [5]. They first predict sensitive attribute values (e.g., male or female) by adding prediction heads. They then remove the attribute information by updating the gradient inversely. Recently, prior studies [9], [10] have extended these techniques for multiattribute protections by introducing more constraints or prediction heads for the multiple attributes. This allows for enhancing model fairness for multiple attributes simultaneously.

However, directly adapting existing methods for multiattribute protection has limitations. First, it incurs substantial computational costs. Current approaches only consider a

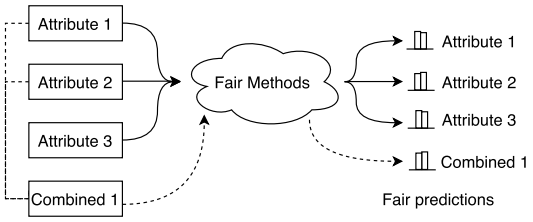


Fig. 1. Fairness-gerrymandering issue [14] for multiattribute protections.

small number of sensitive attributes, dealing with up to three attributes in the experiments [9], [11], [12]. As the number increases, the computational cost becomes impractical. Second, extending these methods risks inheriting or exacerbating their drawbacks. For example, adversarial methods require careful tuning to ensure training convergence due to the inverse gradient updating operation [13]. This issue may worsen under multiattribute settings. Stable training may become more challenging when updating multiple inverse gradients simultaneously. More critically, current methods suffer from the fairness-gerrymandering issue [14], where certain attribute combinations produce unfair predictions even if each attribute is protected by fairness interventions. By enforcing protections attribute-by-attribute, predictions on certain combined subgroups (e.g., young women and old men) may still remain biased. Fig. 1 illustrates this fairness-gerrymandering issue where per attribute fairness interventions risk overlooking unfairness for certain combined subgroups.

Recent studies have utilized mixup operations [15], [16] to blend inputs from different subgroups for single attribute protections [17], [18]. The resulting mixed samples contain "neutralized" features, leading to unbiased predictions. Inspired by this, we propose mixing inputs across all considered attributes for multiattribute protections. The mixed samples construct a neutral domain, where each sample contains neutralized features for all attributes. For instance, a sample can be mixed from inputs with different genders and age groups (e.g., male and female as well as young and old), neutralizing biases of gender and age simultaneously. The method allows models to learn unbiased predictions without costly constraints or additional prediction heads. Moreover, as the mixed samples contain integrated features across all attributes, the method avoids protecting each attribute separately, reducing the risk of fairness gerrymandering.

However, our experiments reveal that directly applying mixup operations for multiple attributes degrades model prediction performance. This is because excessive repeated mixing results in unrecognizable training data. This impedes model training, as the mixed samples no longer contain recognizable features. To resolve the issue, we propose three distinct mixup schemes: *mixup in turn*, *mixup in distance*, and *mixup via interpolations*. The key is to restrict blending on the same pixel locations multiple times, avoiding excessive mixing. By limiting the mixing, our schemes strike a balance between fusing information across attributes while retaining distinguishing visual features critical for training valid models.

In summary, this work contributes the following additions to the literature.

1) We propose a novel method named MultiFair for multiattribute fairness protections. It allows models to learn fair predictions without costly constraints or additional prediction heads. By mixing inputs across attributes, it considers multiple attributes simultaneously rather than per-attribute protections, avoiding fairness gerrymandering issues.
2) We design three distinct mixup schemes—*mixup in turn*, *mixup in distance*, and *mixup via interpolations*—to effectively fuse information across attributes while retaining recognizable visual features critical for training valid models. The key is to restrict repeated mixup operations on the same pixel locations, preventing excessive distortion.
3) Extensive experiments on multiple real-world datasets demonstrate the effectiveness of the proposed MultiFair.

II. Related Works

A. Fairness Protections

1) Single Attribute Protections: Fairness protection techniques can be categorized as preprocessing, in-processing, and postprocessing methods based on when modifications are applied during model training pipelines. Recent work on in-processing fairness for deep classifiers falls under two main approaches: constraint-based methods and adversarial learning methods. Constraint-based methods incorporate fairness metrics directly into the model optimization as constraints or regularization terms. Early work by Zemel et al. [6] proposed demographic parity (DP) constraints on model predictions. Subsequent methods have used approximations [19] or modified training schemes [8] to improve scalability. Adversarial methods learn fair representations by removing sensitive attribute information. They introduce additional prediction heads for attribute subgroup predictions and remove the information through inverse gradient updating [4], [5] or disentangling features [20], [21], [22], [23]. Other fairness methods include balancing dataset with generative methods [24], [25], [26], data augmentations [27], sampling [28], [29], data noising [30], or reweighting mechanisms [31], [32]. More recently, mixup operations [15], [16] have been adopted [17], [18] to enhance fairness by blending inputs across subgroups. However, these studies focus on protecting single sensitive attributes. Multiattribute fairness protection remains relatively unexplored.

2) Multiattribute Protections: There has been limited work on multiattribute fairness protections. For tabular data, methods like reweighting [33], fair constraints [34], [35], and mutual information minimization [36] have been explored. For image inputs, Hwang et al. [9] try balancing subgroups by generating training data with specified attributes. Others [11], [12], [22] aim to learn fair representations by removing sensitive information or perturbing learned features. However, current multiattribute techniques are limited to protecting a small number of attributes, typically two or three. Ensuring fairness considering more sensitive attributes remains an open challenge.

B. Fairness Protections With Data MIXUP Operations

Data mixup operations have been widely adopted for effective data augmentation and training [15], [16]. For fairness protections, mixup has been applied to safeguard single attributes by blending inputs across subgroups [17], [18]. The mixed samples will contain attribute information across all subgroups. This results in "neutralized" features, which allow models to learn unbiased predictions. We aim to adapt the mixup operations for multiattribute protections. However, direct mixing across multiple attributes risks overly blending inputs and removing distinct visual features. Instead, we design three distinct mixup schemes that strike a balance between fusing information across attributes and retaining distinguishing visual features critical for training valid models.

III. Preliminaries

A. Notations

For clarity, we summarize the main notations used throughout this article in the Nomenclature. Specifically, s_0^i denotes the ith considered sensitive attribute with an attribute value of 0. Similarly, $x_{s_0}^i$ represents a selected sample for the ith attribute with the 0 attribute value. x^m refers to the mixed images composed of x and its counterpart x_c.

B. Fairness Protections

Without the loss of generality, we consider a binary classification task with labels $y = 0, 1$ and multiple sensitive attributes. Given a trained model f and biased training data $x \in X$, the sensitive attributes can be denoted by $S = s^1, s^2, \ldots, s^n$, with subgroups s_0 and s_1 for each attribute. Our goal is to ensure similar prediction results $\hat{y}$ across all subgroups. To measure model fairness performance, we adopt fairness metrics of DP [37] and equalized odds (EOs) [38]. These metrics are widely adopted in the studies.

DP [37] measures the divergence in positive predictions across different subgroups and can be expressed as follows:

$$P(\hat{y} = 1|x, s), \quad s = \{s_0, s_1\}. \tag{1}$$

EOs [38], on the other hand, insist on parity in true positive rates (TPRs) and false positive rates (FPRs) across different groups. The fairness metric can be formalized as follows:

$$P\{\hat{y}|y, s\}; y = \{0, 1\}, \quad s = \{s_0, s_1\}. \tag{2}$$

Given fairness metrics, we measure model fairness performance with the discrimination level of the considered fairness metrics. Formally, the discrimination level Γ can be expressed by

$$\Gamma = |\gamma(f, x, s_0) - \gamma(f, x, s_1)| \tag{3}$$

where γ is the considered fairness metric.

IV. Multiattributes Fairness Protections

In this section, we first explain how mixup operations can enhance model fairness for single attributes. We then focus on protecting multiattributes with the operation: we present the overall structure of MultiFair and introduce the proposed mixup schemes. Finally, we provide theoretical analyses of the proposed schemes.

A. Fairness via Mixup Operations

Studies [17], [18] adapt mixup operations [15], [16] for single attribute protections. They mix inputs across subgroups such that the resulting mixed samples contain information from all subgroups. This allows models trained on the mixed data to learn "neutral" representations of the sensitive attribute, leading to enhanced fair prediction results. Compared with Naive mixup operations in [15] and [16], which mix samples from different classes, the proposed method only mixes *same-class* inputs. The method is named *fair mixup* operation.

Formally, the mixed sample x^m is composed of

$$x^m, y := (\lambda x_{s_0} + (1 - \lambda)x_{s_1}, y), \quad y = \{0, 1\} \tag{4}$$

where λ is a weighting parameter randomly set during training. Notably, as the selected inputs x_{s_0} and x_{s_1} are from the same class $y = \{0, 1\}$, the label y remains unchanged for the resulting mixed sample x^m. While fair mixup operations can enhance model fairness, the method targets single attribute protections. In the following, we explore multiattribute protections with the mixup method.

B. Mixup for Protecting Multiattributes

Given prior works [17], [18], it is natural to adapt the mixup operations for multiattribute protections by blending inputs across all attributes. Intuitively, the resulting samples will contain "neutral" information for all considered attributes. For example, a training sample that is young and male could be mixed with samples that are old and female. The resulting blended sample would then contain combined young, old, male, and female information simultaneously. By mixing across all subgroups, the goal is to generate samples with neutralized features across every sensitive attribute. These mixed samples create a neutral domain where each sample contains neutralized features across all attributes. Crucially, mixing across all attributes avoids separate protections per attribute, mitigating fairness gerrymandering risks. Fig. 2 illustrates the procedure considering three attributes. By mixing across subgroups and attributes, the resulting samples contain neutralized features for every individual and combined attribute.

However, directly mixing inputs across multiple attributes can degrade model prediction performance. Excessive repeated mixing results in unrecognizable training data, yielding unsuitable visual features for training valid models. Table I shows the results considering different numbers of attributes. We consider the CelebA [39] dataset for *smiling* classifications. The 0 attribute indicates that no mixup operations are performed, leading to biased predictions. As more attributes

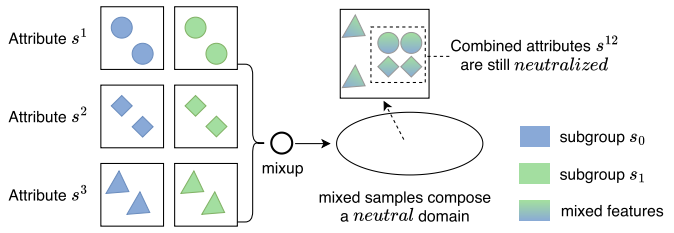


Fig. 2. Mixing different subgroup samples with three attributes, the resulting samples compose a *neutral* domain, where each sample has neutralized features.

TABLE I

WHEN MORE ATTRIBUTES ARE CONSIDERED, MODELS SUFFER DEGRADED PREDICTIONS DUE TO EXCESSIVE MIXUP OPERATIONS

Number of attributes	0	1	2	3	4
Acc	81.4	88.5	65.4	55.3	50.6
DEO	23.6	2.4	1.7	0.3	0.3

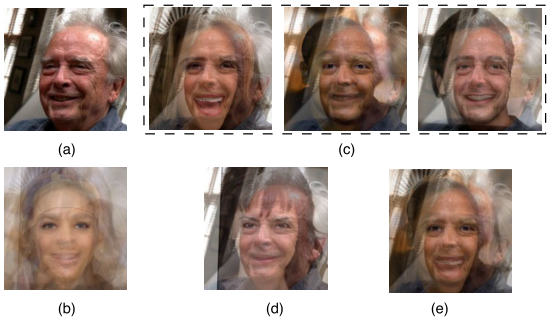


Fig. 3. Resulting mixed sample comparisons with different mixup schemes. (a) Target image. (b) Naive mixup. (c) Mixup in turn. (d) Mixup in distance. (e) Mixup via interpolations.

are considered, the prediction performance declines due to the excessive mixup operations.

To address this issue, we propose three mixup schemes: mixup in turn, mixup in distance, and mixup via interpolations. The key is to restrict blending the same pixel locations multiple times, avoiding excessive distortion. Fig. 3 shows the comparison of the resulting mixed samples with different mixup schemes with attributes of *gender, baldness, and age*. The figure illustrates that the proposed schemes preserve visual features that are better suited for training valid models.

With the designed schemes, we propose MultiFair, which enhances fairness for multiple sensitive attributes. Fig. 4 illustrates the overall structure of the proposed method. It enhances model fairness without modifying the training process. This avoids computational overhead and training instability concerns. Next, we introduce each proposed scheme in detail.

C. Different Mixup Schemes

Fig. 5 presents the proposed mixup schemes: random mixup as a baseline and three designed schemes. In the following, we first present one baseline scheme and then introduce the proposed schemes.

1) Mixup Randomly: One straightforward approach is to *mix inputs between two randomly selected samples*. As only two samples are mixed at each time, the target learning impact is limited. The resulting mixed samples remain recognizable. Fig. 5(a) illustrates the random mixup scheme. Formally, given input x, one counterpart x_c is randomly selected. Then, the mixed sample x^m is composed as follows:

$$x^m := \lambda x + (1-\lambda)x_c. \tag{5}$$

While this baseline mixup scheme retains recognizable visual features for training models, it cannot guarantee fairness. For instance, if the selected counterpart x_c has similar attribute values to x, unfair predictions may still exist. Next, we propose three schemes that strike a balance between fusing cross-attribute information and preserving recognizable visual features.

2) Mixup in Turn: The "mixup in turn" scheme *considers one sensitive attribute to mix samples per training iteration*. Given biased data with multiple sensitive attributes $S = s^1, s^2, \ldots, s^n$, at each iteration only one attribute $s^i \in S$ is selected. Images are mixed across the subgroups s_0, s_1 for the selected attribute.

Formally, with the Mixup in turn scheme, the composed image x^m is

$$x^m := \lambda x^i_{s_0} + (1-\lambda)x^i_{s_1} \tag{6}$$

where i indicates the attribute considered for the current iteration. $x^i_{s_0}$ represents a training sample with the attribute value of s_0 for the ith selected attribute. The value of i iterates through each attribute during training. Fig. 5(b) describes the progress, where attributes are selected in turn.

This scheme enforces multiattribute fairness by enhancing fairness for one attribute per iteration. As the training iterates through each attribute, models learn fair predictions across all attributes.

3) Mixup in Distance: Iterating through all attributes is time consuming as the number of attributes grows. To accelerate training, we propose "mixup in distance"—*mixing inputs with the largest attribute value distance*. We first calculate input distance according to the diverse attribute values. We first define the attribute value distance.

As each sample has attribute values s_0, s_1 for each sensitive attribute $s^i \in S$. We define an attribute value vector V, which considers all attribute values

$$V = [s^1, s^2, \ldots, s^n] \tag{7}$$

where $s^i = \{0, 1\}$, $i = \{1, \ldots, n\}$ indicates the binary attribute value for each attribute. For example, with 3 sensitive attributes, the attribute value distance vector could be $V = [0, 1, 0]$.

Then, the attribute value distance D_{attr} between two samples x and x' can be defined as follows:

$$\begin{aligned} D_{\text{attr}} &= V_x - V_{x'} \\ &= \sum_{i=0}^{n} \left| s^i_x - s^i_{x'} \right|. \end{aligned} \tag{8}$$

The proposed scheme aims to find the counterpart x_c, which is of the largest distance from x

$$x_c = \max_{x' \in X}(V_x - V_{x'}). \tag{9}$$

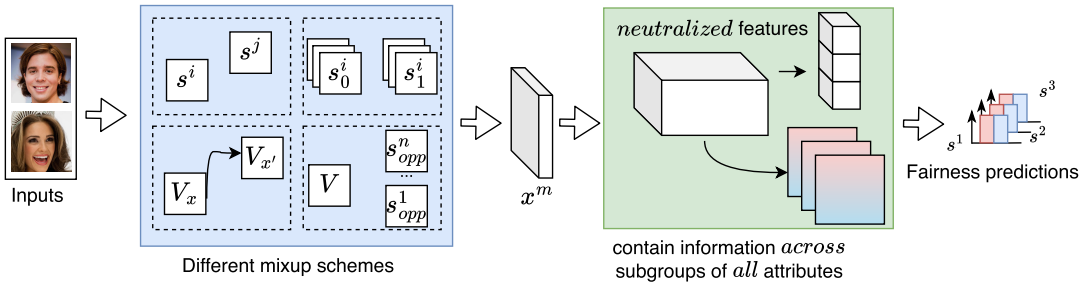


Fig. 4. Overall structure of the proposed MultiFair, colors (red and blue) indicate different subgroups (s_0 and s_1).

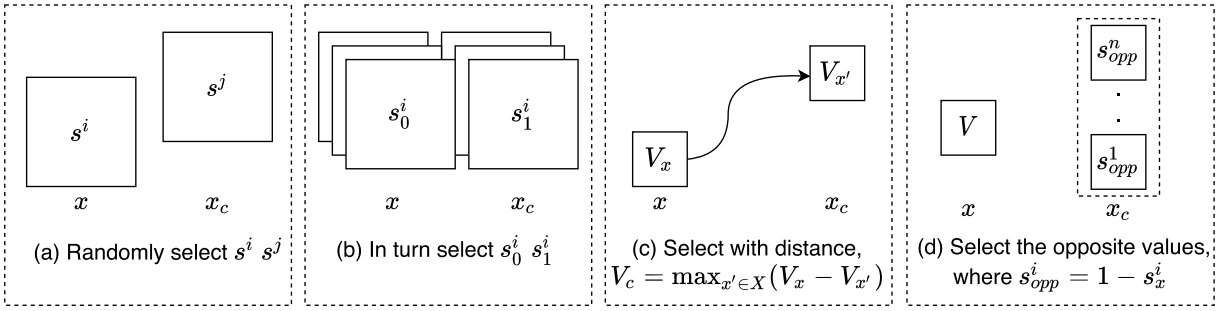


Fig. 5. Different mixup schemes. (a) Mixup randomly. (b) Mixup in turn. (c) Mixup in distance. (d) Mixup via interpolations.

Intuitively, the scheme selects the counterpart that is the furthest in terms of attribute values across all sensitive attributes. Formally, with the selected image x_c, the mixed image x^m can be composed of

$$x^m := \lambda x + (1 - \lambda) x_c. \tag{10}$$

Fig. 5(c) illustrates the proposed scheme where samples with the largest attribute value distances are selected for mixup operations.

This scheme improves model fairness as the mixed images have different attribute values. Ideally, if the values are total opposites, multiple attributes can be viewed as a single attribute. For instance, given three attributes, if the selected samples share the attribute vectors of $[0, 1, 1]$ and $[1, 0, 0]$. Then, the scheme resembles mixup operations for single attribute protections. Thus, when attributes are maximally different, the scheme enhances fairness for multiattributes.

4) Mixup via Interpolations: A limitation of the previous scheme is its reliance on finding optimal counterparts. For multiple attributes, finding x_c with opposite values to x (e.g., $V_c = [1 - s_x^i, \forall s_x^i \in V_x]$) is challenging when attribute number increases. Suboptimal counterpart selection compromises fairness enhancement performance.

To address this, we propose the scheme of "mixup via interpolations": *rather than finding the optimal counterpart, the proposed scheme aims to generate an optimal one by interpolating samples with different attribute values.* Fig. 6 illustrates the procedure. We first select samples of opposite attribute values to x. We then generate x_c with interpolations. In the end, we mix x and x_c to obtain mixed samples.

Formally, we select multiple samples for generating x_c with

$$x_{\text{sel}}^i \leftarrow \{1 - s_x^i, s_x^i \in V_x\} \tag{11}$$

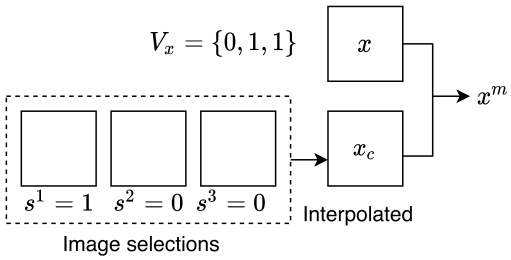


Fig. 6. Mixup via interpolations: we first generate x_c with selected samples, and then mix x with x_c to obtain x^m.

where x_{sel}^i presents the selected sample for the ith attribute. The left arrow indicates the selection procedure with attribute values. With n sensitive attributes, n samples are selected. We then generate the counterpart x_c with

$$x_c := f_{\text{inter}}\left(x_{\text{sel}}^1, x_{\text{sel}}^2, \ldots, x_{\text{sel}}^n\right) \tag{12}$$

where f_{inter} presents the interpolation procedure. Finally, the resulting mixed sample x^m can be obtained via the calculation in (10). Fig. 5(d) illustrates the procedure, where multiple samples are selected for generating the counterpart x_c.

D. Theoretical Analyzes

In this section, we analyze the proposed MultiFair from the perspectives of domain adaptation and causality.

1) Domain Adaptation Interpretations: Studies in domain adaptation demonstrate that prediction differences across domains can be bounded in terms of feature differences. Given $g : x \to u$ as the feature extraction procedures with a function g and output feature u, the proposed MultiFair enforces fairness through learning features that are close across subgroups.

Theorem 1: Given a classification model f with training data $x \in X$. $S = \{s_0^i, s_1^i\}, i = \{1, \ldots, n\}$ present different sensitive attributes with different values. Following (3), the discrimination level Γ' for multiple attributes will be declined with $\Gamma' \leq \Gamma^{\text{org}}$ after enforcing the proposed MultiFair if

$$D\left(\bar{u}_{s_0}^i, \bar{u}_{s_1}^i\right) \leq D\left(u_{s_0}^i, u_{s_1}^i\right), \quad i = \{1, \ldots, n\} \tag{13}$$

where u represents the learned features of models and $\bar{u}$ denotes the features learned from training with mixed data. D measures the distance between features. $u_{s_j}^i$ represents the features from a sample, considering the ith sensitive attribute values.

Proof: Equation (4) shows that the mixed samples contain information across subgroups. The generated training data will be "closer" as they share mixed domain information. Consequently, the features across subgroups will be more similar compared with features from original models, leading to decreased feature distance as illustrated in (13).

Meanwhile, in domain adaption studies [40], [41], the prediction errors between two domains can be bounded with a probability of at least $1 - \delta$

$$\epsilon^{e_1}(f \bullet g) - \epsilon^{e_2}(f \bullet g) \leq D\left(u^{e_1}, u^{e_2}\right) + \lambda + \sqrt{\frac{4}{N}\left(d \log \frac{2n}{d} + d + \log \frac{4}{\delta}\right)} \tag{14}$$

where e_1 and e_2 present the two domains and N represents the number of training data.

From (13), after enforcing the proposed MultiFair, the difference of prediction errors $\bar{\epsilon}$ across subgroups can be formulated as follows:

$$\begin{aligned} \bar{\epsilon}_{s_0}^i(f \bullet g) - \bar{\epsilon}_{s_1}^i(f \bullet g) &\leq D\left(\bar{u}_{s_0}^i, \bar{u}_{s_1}^i\right) + C \\ &\leq D\left(u_{s_0}^i, u_{s_1}^i\right) + C \\ &\leq \epsilon_{s_0}^i(f \circ g) - \epsilon_{s_1}^i(f \circ g) \end{aligned} \tag{15}$$

where C is one constant number. As the discrimination level in (3) is proportional to the gaps of prediction errors, (15) indicates that the discrimination level Γ' will decrease after enforcing the proposed MultiFair. □

2) Causality Interpretation: MultiFair breaks the spurious correlations between sensitive attributes and model prediction targets as the mixed samples contain sensitive information about all subgroups and attributes. Fig. 7 shows the idea, where we take the case of three sensitive attributes as an example. The biased data leads to spurious correlations between attribute values and the prediction classes. With mixed samples, all training data contains information across attribute values, breaking the correlations. Consequently, trained models learn invariant features across subgroups and generate fair predictions.

Formally, with the proposed MultiFair, model fairness performance can be improved with the following theorem.

Theorem 2: Given a model f with training data $x \in X$ considering multiple attributes with different subgroup attribute values, the prediction discrimination level Γ will be minimized if

$$|P(y|x) - P(y|x^m)| \leq \sigma \tag{16}$$

where σ is one small number.

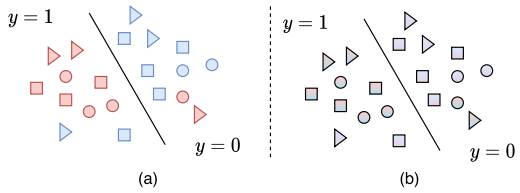


Fig. 7. Spurious correlations due to biased training data. The proposed method breaks the correlations with the mixed samples. (a) Original biased data. (b) Generated data.

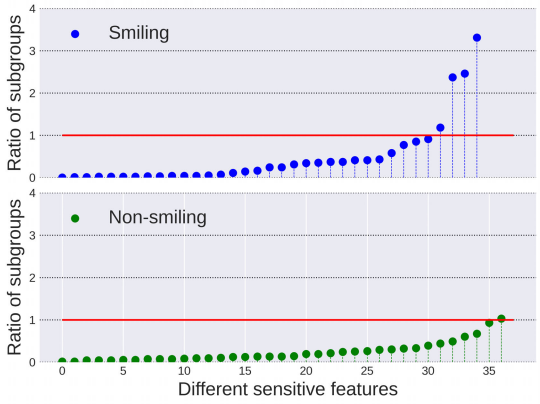


Fig. 8. Data distributions with different attributes. This figure shows that most attributes (37 out of 39) suffer biased distributions.

Proof: The proposed MultiFair requires the same prediction classes for x and x^m. Consequently, the prediction results between the two will be minimized during training, leading to decreased prediction gaps in (16).

As fair mixup operations break spurious correlations, $P(y|x^m)$ should be invariant across different subgroups, where $P(y|x^m, s_0^i) = P(y|x^m, s_1^i)$. Then, with the decreased prediction gaps in (16), we have

$$\left|P\left(y|x, s_0^i\right) - P\left(y|x, s_1^i\right)\right| \leq \sigma. \tag{17}$$

With the prediction error of $\epsilon = y - P(y|x)$, the prediction discrimination level Γ can be expressed with

$$\begin{aligned} \Gamma &= \left|\gamma\left(f, s_0^i\right) - \gamma\left(f, s_1^i\right)\right| \\ &= \left|\left(y - P\left(y|x, s_0^i\right)\right) - \left(y - P\left(y|x, s_1^i\right)\right)\right| \\ &= \left|P\left(y|x, s_0^i\right) - P\left(y|x, s_1^i\right)\right| \leq \sigma. \end{aligned} \tag{18}$$

The equation indicates that the proposed MultiFair will minimize discrimination and improve model fairness performance. □

V. Experiments

We evaluate the proposed method on multiple datasets with up to *eight* sensitive attributes. Our experiments include diverse ablation studies comparing MultiFair to several baselines. We first introduce the experimental settings and then present results demonstrating MultiFair's performance, along with detailed ablation studies. Finally, we provide detailed discussions of the proposed method and its potential applications.

A. Experimental Settings

1) Biased Training Data: We consider a real-world dataset CelebA [39], which contains 30 000 facial images and 40 attribute annotations. In the experiments, we train models for *smiling* classifications and find the training data suffer biased data distributions with multiple attributes. Fig. 8 shows the biased distributions for *smiling* and *nonsmiling* groups per attribute. We compute the training samples per class across subgroups for each attribute and sort the results within a range of $[0, 4]$ (35 out of 39 attributes). The figure shows that only a few attributes have even distributions. Notably, correlated attributes (e.g., gender and lipstick) tend to co-occur in biases. These results indicate simultaneously occurring multiattribute biases, calling for fairness protections with multiple sensitive attributes. In the experiments, we consider a biased ratio of 10% (e.g., 90% data comes from majority subgroups). We first consider protections for single attributes and then protection fairness for multiple ones.

2) Training Models: We adopt MobileNet from [42] for *smiling* classifications. The training is conducted by adopting the Pytorch framework with the compute unified device architecture (CUDA) and CUDA deep neural network (CuDNN) backends. All models are trained with the Adam optimizer and stochastic gradient descent. The batch size is 64, and the learning rate is $1e^{-4}$. We run each experimental setting multiple times and report the mean values.

3) Baseline Methods: In the experiments, we compare the results considering the mixup schemes of Mixup randomly ("Random"), Mixup in turn ("In turn"), Mixup in distance ("Distance"), and Mixup via interpolations ("Interp."). To compare fairness protection results, besides biased ("Biased") and balanced ("Balanced") models, we consider the following baselines.

1) *Preprocessing:* Resampling methods "Resample" adjust data distributions in the training set to reduce bias.
2) *In-Processing:* Adversarial training methods ("Adv") from [43] remove sensitive information in learned features to reduce bias; fair classification orthogonal representation (FCRO) methods ("FCRO") from [10] protect multiattributes based on Adv for medical images. They remove the sensitive information among extracted features with designed constraints. In experiments, we consider one shared model for both target and attribute predictions.
3) *Postprocessing:* EOs postprocessing methods ("EOP," threshold optimization) from [38] adjust decision thresholds across subgroups to reduce bias.

For single attribute protections, we consider baseline methods of biased, balanced, resample, EOP, and Adv. For multiattribute protections, we consider baseline methods of resample, EOP, Adv, and FCRO. Specifically, we resample training samples considering multiple attributes for the resample method. For EOP, we try to identify proper thresholds for multiple attributes. We introduce more prediction heads for the Adv method.

4) Evaluation Metrics: We measure model fairness performance with fairness metrics of DP and EOs. Meanwhile, we adopt accuracy and $F1$ scores to evaluate model prediction performance.

TABLE II
FAIR RESULTS FOR SINGLE ATTRIBUTE PROTECTIONS

Considered attributes	Methods	Mean Acc↑	Δ_{DP}↓	Δ_{EO}↓	s_0 Acc↑	s_1 Acc↑
Gender	Biased	81.5	4.4	23.2	83.6	79.4
	Resample	80.2	3.6	17.4	81.4	79.0
	EOP	80.6	5.6	13.6	82.1	79.1
	Adv	83.9	8.6	16.6	88.2	79.6
	Balanced	88.9	1.4	1.4	89.6	88.2
	Fair mixup	88.5	4.2	4.1	89.8	87.2
Makeup	Biased	82.3	7.5	13.8	86.1	78.6
	Resample	82.4	7.6	9.1	84.1	80.7
	EOP	77.6	5.6	8.6	76.2	79.1
	Adv	85.6	5.6	10.7	82.3	87.9
	Balanced	87.9	2.4	2.4	86.7	89.1
	Fair mixup	89.1	5.7	5.8	92.6	86.6
Age	Biased	87.6	2	7.7	89	86.2
	Resample	80.3	2.2	4.3	82.5	78.1
	EOP	86.9	4.6	4.1	85.3	88.5
	Adv	86.3	1.7	5.6	86.4	86.2
	Balanced	88.8	1.1	1.8	88.3	88.6
	Fair mixup	89.6	1.8	4.4	88.7	90.6

B. Experimental Results

1) Single Attribute Protections: Table II presents the results for single attribute protections. We consider attributes of *gender*, *makeup*, and *age* and enforce fairness interventions for each attribute. "Fair mixup" is the mixup operations in (4). We compare the results with baseline methods. The table shows that the Fair mixup achieved the most effective protections across metrics and attributes. Meanwhile, it maintains competitive accuracy performance compared with the balanced models. While baselines, such as EOP and Adv, also enhanced model fairness, their results are suboptimal. In addition, in some cases, their methods lead to decreased accuracy due to the interventions. The results demonstrate that fair mixup operations can deliver fair and valid prediction results.

2) Three Attribute Protections: In this section, we consider protecting multiattribute protections. Specifically, we consider attributes of *gender, chubby, and goatee* simultaneously. Table III presents the results, where the mixup schemes consistently enhance model fairness across all attributes considered. In contrast, some baselines fail to protect the attributes. Resampling to construct balanced data is challenging and computationally expensive for multiattributes. Meanwhile, as samples contain multiple attributes, adjusting thresholds for one attribute inevitably affects others. This makes it challenging for the EOP method to identify thresholds that promote fairness across all attributes and maintain valid predictions simultaneously. While adversarial methods such as Adv and FCRO can deliver fairness protections, their inherently adversarial design leads to unstable training results.

For the proposed method, the interpolation scheme achieves the best protection results with the lowest fairness metric values. However, its accuracy is slightly lower than Mixup in turn and Mixup in distance schemes. This is because these schemes generate more realistic mixed samples that preserve

TABLE III
FAIR RESULTS CONSIDERING THREE SENSITIVE ATTRIBUTES

Considered attributes	Methods	Mean Acc↑	Δ_{DP}↓	Δ_{EO}↓	s_0 Acc↑	s_1 Acc↑
The presence of chubby	Biased	75.5	8.2	14.6	71.4	79.6
	Resample	74.3	8.4	12.4	72.2	76.4
	EOP	72.1	4.4	10.7	72	73.7
	Adv	78.7	3	11.8	78.6	78.8
	FCRO	79.2	2.9	7.6	79.4	79.0
	Random	80	9.2	12.8	75.4	84.6
	Distance	**83.1**	6.6	9.9	79.9	86.4
	In turn	82	6.5	10	78.7	85.2.4
	Interp.	80.3	**2.7**	**5.9**	78.9	80.3
The presence of goatee	Biased	77.6	10.4	26	72.4	82.8
	Resample	79.1	7.3	15.8	75.3	82.9
	EOP	80.1	12.9	17.5	78.6	81.6
	Adv	81.8	7	18	79.4	84.2
	FCRO	81.6	6.1	**13.2**	79.6	83.6
	Random	80.8	10	18	75.8	85.8
	Distance	83.8	9.8	19.2	78.9	88.8
	In turn	83.1	10.3	18.6	77.8	88.2
	Interp.	**86.3**	**6.1**	15.1	80.2	86.3
Gender	Biased	80.1	5.4	28.6	77.4	82.8
	Resample	79.3	4.7	23.1	78.5	80.1
	EOP	77.3	5.7	17.2	76.8	77.8
	Adv	82.6	3	21.6	82.2	83
	FCRO	83.6	2.5	19.3	83.0	84.2
	Random	85.7	5.8	20.6	82.8	88.6
	Distance	**87.7**	3.9	**16.2**	85.7	89.6
	In turn	86.9	5.8	18.2	84.1	89.8
	Interp.	85.5	**2.4**	18	85.8	85.1

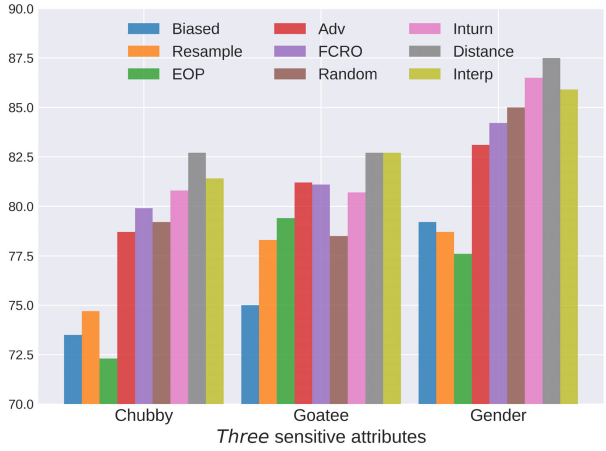


Fig. 9. $F1$ score comparisons for different methods.

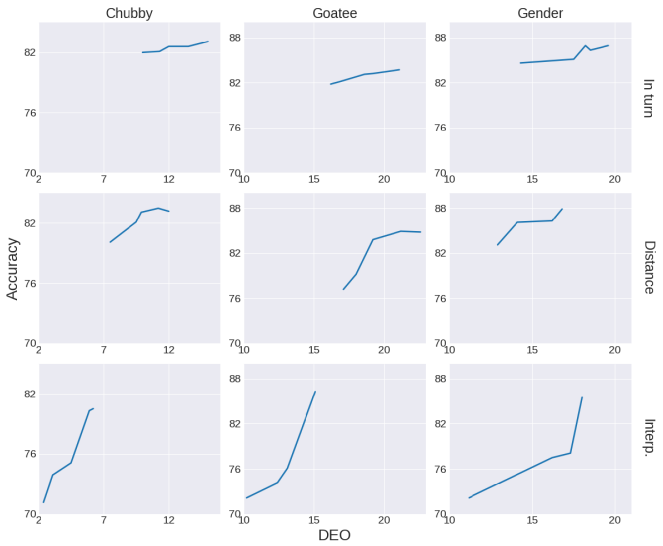


Fig. 10. Fairness-accuracy trade-off across different attributes for the proposed schemes.

distinguishing visual features for predictions. However, they require longer training with more iterations and attribute distance calculations.

In addition, comparing the results in Table II, fairness protections for multiple attributes prove more challenging. This may stem from the biases that can occur in combinations of attributes. When multiple attributes are considered, the combinations of these attributes may introduce extra biases for model predictions, leading to suboptimal protection results.

In Section V-B2a–d, we further explore the protection results regarding model utility, fairness gerrymandering concerns, fair representations, and computational costs

a) Model utility: We compare $F1$ score results to explore trained model prediction performance in Fig. 9. In the figure, the schemes of Mixup in turn and Mixup in distance achieved better results than the other methods, which are consistent with previous observations. The resulting mixed samples from these two schemes are much more realistic than other schemes, contributing to enhanced prediction performance.

We further explore the fairness-accuracy trade-off for the proposed schemes in Fig. 10. The Pareto fronts show trained models that shift the balance between achieving higher accuracy and ensuring fairness. We observe that the turn method yields the most consistent accuracy performance as it does not require repeating mixup operations, resulting in more realistic mixed samples. In contrast, the interpolation schemes achieve the best fairness results but high variability in accuracy. This is likely due to the generation of more complex blended samples, which contain neutralized information across all attributes. The distance scheme strikes a balance. It mixes samples with the largest attribute distance while avoiding excessive mixup operations, enabling valid training results. Despite differing instability, all three schemes demonstrate enhanced fairness with valid model predictions.

b) Fairness for combined attributes: Fairness studies [14] raise concerns about fairness gerrymandering in multiattribute protections: combinations of the protected attributes may still encounter unfairness even if fairness interventions have been enforced per attribute. In this section, we address this by further exploring MultiFair's performance on combined attributes. With the three attribute protection results in Table III, we examine the combined attributes of *chubby and goatee*. Table IV reports the results.

In the table, the binary values represent subgroups for each attribute, and number combinations denote the combined attributes. For example, 0 denotes the s_0 subgroup, and 00 presents the combined subgroups $\{s_0^1, s_0^2\}$. The first number

TABLE IV
PROTECTION RESULTS FOR ATTRIBUTE COMBINATIONS OF *Chubby* AND *Goatee*

Methods	00-01	00-10	00-11	01-10	01-11	10-11	Mean
Biased	23.4	13.1	26.2	10.3	7.1	13.1	18.2
Resample	26.3	15.6	16.2	13.1	4.9	17.6	14.5
EOP	20.7	14.2	19.6	10.7	3.2	10.4	23.2
Adv	27.8	17.1	30.3	10.6	2.4	13.1	20.4
FCRO	17.4	**7.1**	16.4	9.8	1.4	11.2	13.6
Random	20.9	9.3	21.8	11.5	3.4	12.4	16.6
Distance	20.6	8.4	20.6	14	**1.2**	14	17.3
In turn	19.6	9	22.1	10.6	3.1	13.1	16.3
Interp.	**14.3**	10.6	**14.3**	**9.3**	2.4	**9.3**	**11.8**

indicates the attribute of *chubby*, and the second indicates *goatee*. The column "00–01" shows the disparity results of fairness metrics between the "none-chubby-and-none-goatee" and "none-chubby-and-goatee." We measure fairness using EOs and report the mean values following previous settings.

The table shows biased models lead to unfairness for combined attributes. Similar trends can also be observed with baseline methods of Resample, EOP, and Adv. The results confirm concerns from [14]. They cannot deliver consistent protections for the combined attributes due to per-attribute fairness interventions. However, the proposed mixup schemes exhibit lower metric values, demonstrating efficient protections for different attribute combinations. Compared with baselines, MultiFair achieves lower values in most cases. These results highlight that MultiFair can enhance protections for combined attributes, mitigating fairness gerrymandering risks.

c) Feature visualizations: We further examine the learned representations under MultiFair using TSNE visualizations from [44]. Specifically, we extract features from the last model layer from the results of the interpolation scheme and compare them with the biased models.

Fig. 11 shows the feature visualizations for attributes of *chubby, goatee, and gender*. Each plot presents results for each attribute. In the plot, we show biased model features in the upper area while the MultiFair features in the lower area. Different colors denote subgroups (e.g., blue/red for male/female). The facial images are samples from different subgroups. For biased models, data points across subgroups are separable. In contrast, MultiFair blends the points across subgroups together. This indicates that attribute information has been adopted as classification clues for the biased models. Hence, the extracted features are separable in terms of the attribute values. MultiFair, however, encourages models to learn invariant features across subgroups, contributing to fair predictions. The results demonstrate that MultiFair produces consistently fair predictions for the considered multiple attributes.

d) Computational complexity: Table V provides computational complexity analyses. Our schemes involve data preprocessing, incurring light computational costs. Specifically, they entail counterpart selection with complexity $\mathcal{O}(n_b)$ for batch size n_b, and mixup complexity of $\mathcal{O}(n_i^2)$ with image size n_i. In addition, the interpolation scheme mixes each sample twice, while the distance scheme requires attribute distance calculations. In contrast, in-processing baselines like Adv and

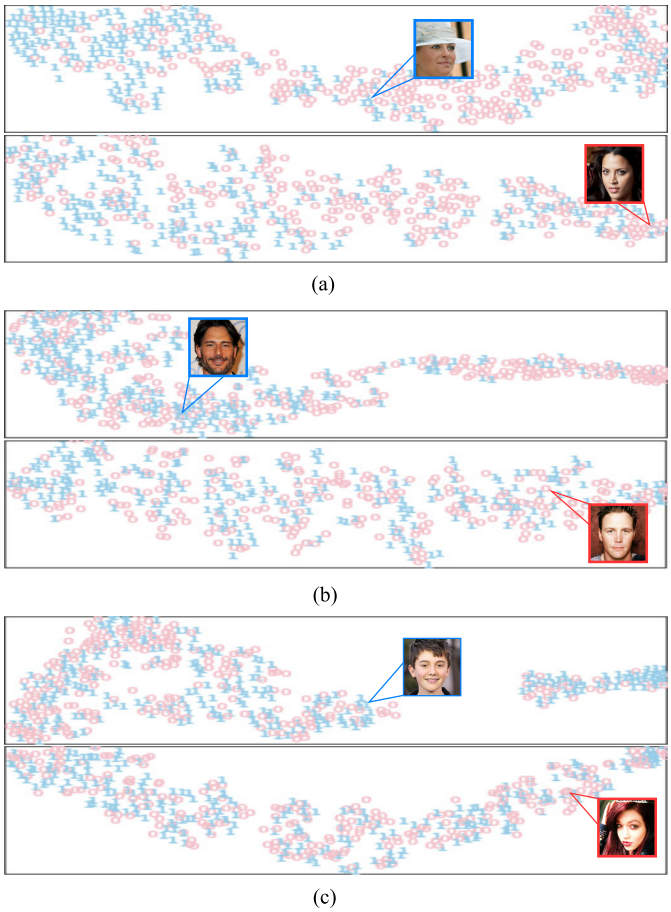


Fig. 11. TSNE results for the attributes *chubby, goatee, and gender*. 1 (blue) and 0 (red) indicate different subgroups. Data points from MultiFair blend into each other, indicating fair predictions for multiple attributes. (a) TSNE plot for the attribute of the presence of chubby. (b) TSNE plot for the attribute of the presence of goatee. (c) TSNE plot for the attribute of the presence of male.

FCRO require heavy computations for model modifications. While resampling to balance the data distribution avoids heavy computational demands, the resulting enlarged training data leads to increased training time. The EOP method involves identifying thresholds that both enforce fairness for all attributes and maintain model prediction performance simultaneously. This requires repeated model training to find the valid thresholds. Compared with baseline methods, our method exhibits efficient fairness protections with multiple attributes.

C. Ablation Studies

In this section, we consider diverse settings to further evaluate each module of the proposed model.

1) Model Structures: While previous experiments consider MobileNet [42] to train target models, we further assess model fairness performance considering structures of: ResNet18 [45] and VGG [46]. Table VI shows the results. The table shows that the fair mixup achieved enhanced fairness performance with both structures. Notably, compared with VGG models, the results show higher fairness metric values with lighter models like ResNet18. They tend to be more vulnerable to the effects

TABLE V
COMPUTATIONAL COMPLEXITY ANALYSES AND COMPARISONS

Methods	Model structure modifications	Additional training epochs	Pre/Post-processing	
			Resample	Threshold
Resample	-	-	✓	-
EOP	-	-	-	✓
Adv	✓	-	-	-
FCRO	✓	-	-	-
			Selections	Generations
Random	-	-	$\mathcal{O}(n_b)$	$\mathcal{O}(n_i^2)$
Distance	-	-	$\mathcal{O}(n_b)$	$\mathcal{O}(n_i^2)$
In turn	-	✓	$\mathcal{O}(n_b)$	$\mathcal{O}(n_i^2)$
Interp.	-	-	$\mathcal{O}(n_b)$	$\mathcal{O}(n_i^2)$

TABLE VI
PROTECTIONS WITH DIFFERENT MODEL STRUCTURES

Structures	Models	Acc ↑	$\Delta_{\text{DP}}\downarrow$	$\Delta_{\text{EO}}\downarrow$
Res18	Biased	82.3	4.5	22.7
	Fair mixup	89.2	3.7	4.0
VGG	Biased	87.4	4.0	20.1
	Fair mixup	92.1	2.3	3.5

TABLE VII
RESULTS FOR UTKFACE WITH TWO SENSITIVE ATTRIBUTES

Considered attributes	Methods	Mean Acc↑	$\Delta_{\text{DP}}\downarrow$	$\Delta_{\text{EO}}\downarrow$	s_0 Acc↑	s_1 Acc↑
Race	Biased	76.0	11.3	36.1	75.2	76.8
	Resample	74.3	12.4	17.5	75.6	73.0
	EOP	75.7	9.2	20.1	76.1	75.3
	Adv	75.0	7.2	23.6	74.4	75.6
	FCRO	75.8	9.1	**16.6**	74.8	76.8
	Random	74.2	10.4	30.5	71.6	76.8
	Distance	**77.3**	8.6	18.8	79.2	75.4
	In turn	76.1	**7.2**	23.2	75.6	76.6
	Interp.	75.6	7.7	20.4	74.0	77.2
Gender	Biased	74.8	14.5	41.6	76.0	73.6
	Resample	74.4	12.8	35.1	73.9	74.9
	EOP	72.1	8.2	21.4	71.9	72.3
	Adv	74.6	9.2	23.6	77.1	72.1
	FCRO	73.8	9.5	20.3	74.0	73.6
	Random	73.9	11.5	47.4	73.6	74.2
	Distance	78.6	7.7	25.2	78.8	78.4
	In turn	**79.2**	**7.6**	22.4	82.4	76.0
	Interp.	77.8	9.4	**18.6**	81.2	74.4

of data imbalance, leading to more biased predictions. The results demonstrate that fair mixup operations can enhance fairness consistently with different prediction models.

2) Datasets: This section evaluates the proposed methods with the real-world datasets: UTKFace [47] and FairFace [5]. The UTKFace dataset [47] consists of over 20 000 facial images of different ethnicities. The images are annotated with age, race, and gender information. The dataset includes five ethnicity groups, namely, *White, Black, Asian, Indian*, and *others* (including *Hispanic* and *Latino*). As the *White* dominates the dataset, we group the minority into *Others*. Similarly, ages are grouped into *young* and *old*. In the experiments, we consider *race* and *gender* as sensitive attributes and age as the prediction target.

Table VII shows the results for the UTKFace dataset, where we highlight the best protection results. The table shows that

TABLE VIII
RESULTS FOR FAIRFACE WITH TWO SENSITIVE ATTRIBUTES

Considered attributes	Methods	Mean Acc↑	$\Delta_{\text{DP}}\downarrow$	$\Delta_{\text{EO}}\downarrow$	s_0 Acc↑	s_1 Acc↑
Race	Biased	73.8	6.2	14.8	74.4	73.2
	Resample	73.4	5.7	11.3	73.9	72.9
	EOP	72.1	3.2	7.3	71.3	72.9
	Adv	73.2	6.1	14.4	73.6	72.8
	FCRO	73.4	2.3	4.6	74.2	72.5
	Random	74.8	6.8	11.2	77.2	72.4
	Distance	**79.0**	1.6	**1.1**	78.8	79.2
	In turn	77.0	2.2	5.2	77.2	76.8
	Interp.	78.2	**1.1**	2.8	78.8	77.6
Gender	Biased	80.6	5.0	13.2	79.2	82
	Resample	80.4	4.2	12.5	79.6	81.2
	EOP	78.6	4.4	7.9	77.1	80.1
	Adv	80.1	5.6	13.6	78.0	82.1
	FCRO	80.6	3.3	6.6	78.8	82.4
	Random	77.0	4.8	18.0	77.6	76.4
	Distance	**84.0**	3.0	**6.4**	82.4	85.6
	In turn	83.2	3.5	8.8	81.6	84.8
	Interp.	79.6	**3.2**	10.8	80.8	78.4

the proposed mixup schemes consistently improved model fairness performance with multiple attributes. Specifically, compared to the baselines, the proposed schemes achieved the lowest fairness metric values for the gender attribute and competitive results for the race attribute. Meanwhile, the baselines either achieve suboptimal protection performance or suffer degraded accuracy results. In contrast, the proposed schemes strike a better balance between improving model fairness performance and maintaining valid predictions.

The FairFace dataset [5] comprises over 100 000 facial images. The images are annotated with attribute information of age, race, and gender. Similar to the UTKFace dataset, we group the data according to the annotations. We then consider *gender* and *race* as attributes and age as the prediction target. Table VIII shows the experiment results.

The table shows that the proposed methods consistently deliver fairness protections for biased models. They achieved the best protection performance with the considered attributes and metrics compared with the baselines. Notably, compared with the results of the three attribute protections, fairness interventions generally achieve better protection results. For example, the interpolation and the distance schemes can achieve a value of 1.1 for the metrics of DP and EO for the two attribute protections. This improvement may stem from the fact that the mixup schemes generate more "neutral" representations when fewer attributes are involved, avoiding excessive mixing with more attributes. Furthermore, consistent with previous findings, these schemes improve model fairness performance while maintaining valid predictions. This highlights the effectiveness and robustness of the proposed mixup schemes in protecting fairness for multiattributes.

3) More Attributes: In this section, we simultaneously consider more sensitive attributes to evaluate the proposed MultiFair. Fig. 12 presents the results for protecting four attributes: *clock shadow, arched eyebrows, baldness, and attractiveness*. Fig. 13 shows the results for protecting eight

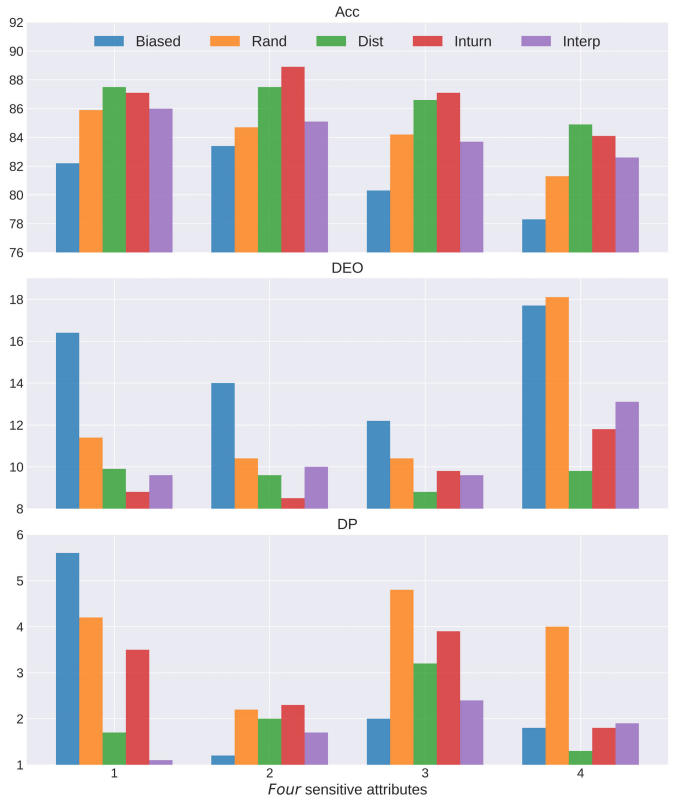


Fig. 12. Fairness protecting results with *four* sensitive attributes of clock shadow, arched eyebrows, baldness, and attractiveness.

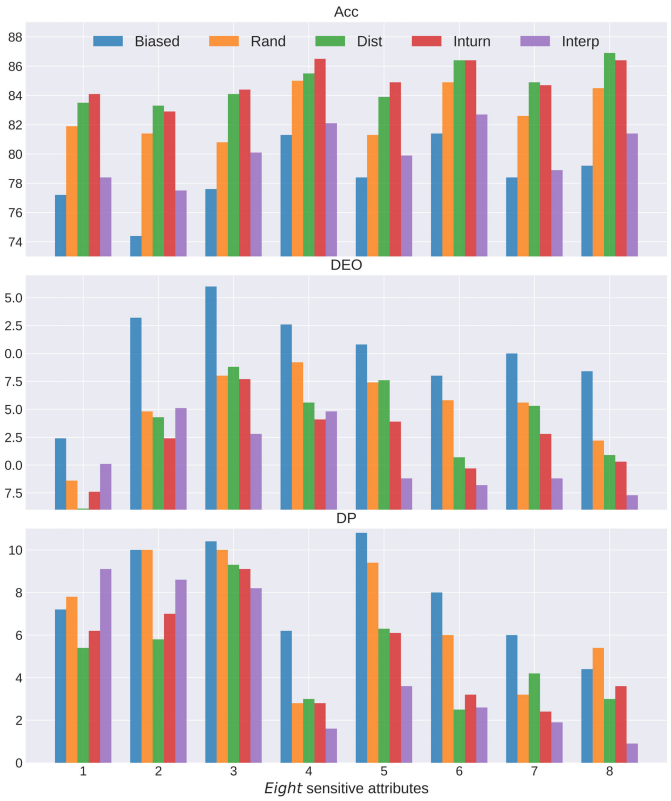


Fig. 13. Fairness protecting results with *eight* sensitive attributes of double chin, eyeglasses, goatee, heavy makeup, sideburns, earrings, hat, and necktie.

attributes: *double chin, eyeglasses, goatee, heavy makeup, sideburns, earrings, hat, and necktie*.

Notably, as more attributes are considered, baselines such as Adv and FCRO struggle to achieve valid predictions. The inverse gradient and constraint operations lead to unstable training and converging difficulty, degrading model accuracy performance. This aligns with previously raised concerns about those methods. While previous results show the feasible protections using resample and EOP methods, the methods are impractical when more attributes are considered. The resample method requires sample data to balance the dataset. The EOP method involves repeatedly training models to find the optimal threshold. As the number of attributes considered increases, balancing the dataset and determining thresholds for all attributes becomes increasingly challenging. Moreover, the soaring computational cost makes them impractical for real-world applications.

In contrast, our proposed schemes maintain effectiveness even with eight considered attributes. The proposed schemes sustain valid predictions while improving fairness compared with biased models. Of the proposed schemes, the interpolation scheme provides the strongest fairness protections but slightly lower accuracy than the Mixup in turn and Mixup in distance schemes. We also observed that, occasionally, the biased model tends to have low DP values. In these cases, MultiFair maintains the low DP value while decreasing the EOs value substantially, indicating superior fairness performance. The results show that the proposed mixup schemes are efficient in improving model fairness for multiattributes.

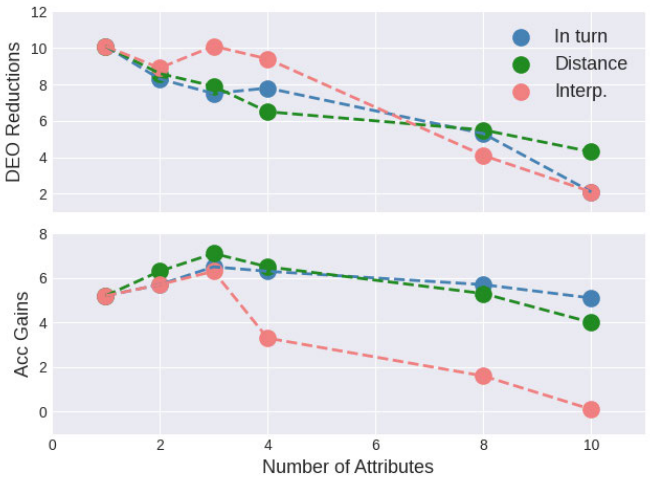


Fig. 14. Impact of the considered attribute number on fairness interventions and model prediction performance. We measure the accuracy gains (Acc gains) and the deceased DEO values (DEO reductions) after fairness interventions.

We further explore the impact of the considered attribute number on model performance in Fig. 14. We adopt the CelebA dataset and consider different attribute numbers, ranging from 1 to 10 attributes. After enforcing fairness interventions with the proposed schemes, we measure the accuracy improvements (Acc gains) and reductions in disparity of equalized odds (DEO) values (DEO reductions) with the mean values across considered attributes. We find that the

gains tend to decrease as more attributes are considered. This trend is more pronounced with the interpolation scheme, though it achieves better protection results with a smaller number of attributes (e.g., three). The results are consistent with previous observations and indicate that with a larger number of attributes (e.g., over ten), the gains in fairness protection and model prediction performance tend to decrease. However, the results demonstrate that the proposed methods can still deliver fairness protections when accounting for the given attributes.

D. Discussion

1) Proposed Schemes: Among the proposed different mixup schemes, results (Table III and Figs. 12 and 13) have shown that the Mixup via interpolations scheme delivered the best fair prediction results with the lowest metric results. This is because the scheme interpolates the counterpart containing information on all attributes, leading to more similar features across subgroups. The results are consistent with analyses in Theorem 2, where similar features contribute to invariant $P(y|x^m)$ and $P(y|x)$ across subgroups. However, their accuracy results are suboptimal compared to other schemes. Schemes of Mixup in turn and Mixup in distance achieve better balances between fairness and accuracy. This is because they only require blending two samples, leading to more realistic mixed samples. Nevertheless, all proposed schemes can deliver protections for multiattributes.

2) Generalization for Other Tasks: The proposed methods construct a neutral domain, which encourages models to learn fair representations. The operation can benefit various other learning tasks. For example, in domain adaptation studies, it is essential to have representations that are similar across different domains. The neutral domain allows models to learn similar representations for multiple domains, facilitating the transfer of knowledge between them. In data augmentations, the proposed mixup schemes can generate diverse samples. This diversity enhances the variety of augmented datasets and makes the trained models more robust. More generally, in feature representation studies, the proposed methods encourage trained models to learn representations with diverse features. This enables the models to gain a more comprehensive understanding of complex patterns and characteristics in the data.

3) Practical Applications: The proposed methods provide fairness interventions for classifiers, which can be applied to multiple areas. For example, for algorithmic decision-making processes, such as credit scoring, hiring, and parole decisions, the methods can be applied to avoid severe discrimination against certain subgroups of considered sensitive attributes such as race, gender, or age. The methods can also be applied in educational settings to monitor and address biases in grading, admissions, and access to educational resources, striving for an equal learning experience for all students. Meanwhile, fairness studies have mitigated biases in image recognition, facial recognition, and natural language processing tasks like sentiment analysis or language translations.

4) Limitations: While our results validate the efficiency of the proposed schemes in protecting multiattributes, there are some limitations that holds the following.

a) Considering more attributes: First, providing protections for a large number of attributes remains challenging. Prior work has generally focused on fewer than three sensitive attributes. While the proposed schemes demonstrate efficient protections for up to eight attributes, as shown in Fig. 14, the gains from fairness protections and model accuracy tend to diminish as more attributes are considered. This suggests that more delicate designs are still needed when enforcing fairness across a large number of attributes (e.g., more than ten attributes). More work is still required to deliver efficient fairness protections for a large number of attributes while maintaining model prediction performance.

b) Protecting unknown attributes: Second, providing protections for hidden attributes remains a significant challenge. There are two types of hidden attributes to consider: combinations of known attributes and attributes that are unknown in the training data. While the proposed method helps protect combinations of known attributes by blending inputs across different attributes, it cannot ensure fairness for unknown attributes. Protecting unknown attributes is difficult without additional information. Some prior works [3], [48], [49] have tried to circumvent the difficulty by requiring partly annotated data in a semisupervised approach. With the partly annotated data, they can identify the attributes for the rest and provide fairness protections with existing intervention methods. However, fully protecting unknown attributes across all data remains an open challenge. Further research is still needed to identify and protect unknown attributes to provide more comprehensive fairness protections.

VI. Conclusion

In this article, we presented a method named MultiFair for multiple sensitive attribute fairness protections. We construct a neutral domain that contains information across subgroups of considered multiple attributes. We proposed three different schemes to mix samples to ensure better fairness protections while maintaining model accuracy performance. We analyze the proposed MultiFair from the perspective of domain adaptation and causality. We undertake extensive experiments with up to eight considered attributes to demonstrate that MultiFair consistently enhances model fairness performance for multiple attributes. The proposed methods will encourage fairness predictions in diverse areas, such as decision-making algorithms. They can also benefit other tasks, such as data augmentation or domain adaptation studies. In future research, we plan to focus on proposing more delicate schemes to construct neutral domains, which will lead to enhanced fairness protections under diverse settings.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019, *arXiv:1908.09635*.

[2] S. Caton and C. Haas, "Fairness in machine learning: A survey," 2020, *arXiv:2010.04053*.

[3] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and P. S. Yu, "Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1763–1774, Apr. 2022.

[4] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9319–9328.

[5] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1548–1558.

[6] R. S. Zemel, L. Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. ICML*, 2013, pp. 325–333.

[7] X. Xu et al., "Consistent instance false positive improves fairness in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 578–586.

[8] P. Manisha and S. Gujar, "FNNC: Achieving fairness through neural networks," in *Proc. IJCAI*, 2020, pp. 1–7.

[9] S. Hwang, S. Park, D. Kim, M. Do, and H. Byun, "Fair-FaceGAN: Fairness-aware facial image-to-image translation," 2020, *arXiv:2012.00282*.

[10] W. Deng, Y. Zhong, Q. Dou, and X. Li, "On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations," in *Proc. Int. Conf. Inf. Process. Med. Imag.* San Carlos de Bariloche, Argentina: Springer, 2023, pp. 158–169.

[11] E. Creager et al., "Flexibly fair representation learning by disentanglement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1436–1445.

[12] W. Deng, Y. Zhong, Q. Dou, and X. Li, "On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations," in *Information Processing in Medical Imaging*, A. Frangi, M. de Bruijne, D. Wassermann, and N. Navab, Eds. Cham, Switzerland: Springer, 2023, pp. 158–169.

[13] F. Schafer and A. Anandkumar, "Competitive gradient descent," in *Proc. NIPS*, 2019, pp. 1–11.

[14] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2564–2572.

[15] V. Verma et al., "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6438–6447.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[17] C.-Y. Chuang and Y. Mroueh, "Fair mixup: Fairness via interpolation," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–15.

[18] M. Du, S. Mukherjee, G. Wang, R. Tang, A. H. Awadallah, and X. Hu, "Fairness via representation neutralization," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 12091–12103.

[19] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *Proc. ICML*, 2021, pp. 11492–11501.

[20] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12110–12119.

[21] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," 2019, *arXiv:1905.13662*.

[22] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 746–761.

[23] D. Guo, C. Wang, B. Wang, and H. Zha, "Learning fair representations via distance correlation minimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2139–2152, Feb. 2024.

[24] S. Hwang and H. Byun, "Unsupervised image-to-image translation via fair representation of gender bias," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 1953–1957.

[25] J. Joo and K. Kärkkäinen, "Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation," in *Proc. 2nd Int. Workshop Fairness, Accountability, Transparency Ethics Multimedia*, 2020, pp. 1–5.

[26] V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9297–9306.

[27] S. Mo, H. Kang, K. Sohn, C.-L. Li, and J. Shin, "Object-aware contrastive learning for debiased scene representation," 2021, *arXiv:2108.00049*.

[28] Y. Roh, K. Lee, S. Euijong Whang, and C. Suh, "FairBatch: Batch selection for model fairness," 2020, *arXiv:2012.01696*.

[29] M. Mahdi Khalili, X. Zhang, and M. Abroshan, "Fair sequential selection using supervised learning models," 2021, *arXiv:2110.13986*.

[30] T. Zhang, T. Zhu, K. Gao, W. Zhou, and P. S. Yu, "Balancing learning model privacy, fairness, and accuracy with early stopping criteria," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5557–5569, Sep. 2023.

[31] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13208–13217.

[32] S. Gong, X. Liu, and A. K. Jain, "Mitigating face recognition bias via group adaptive classifier," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3413–3423.

[33] A. Roy, V. Iosifidis, and E. Ntoutsi, "Multi-fairness under class-imbalance," in *Discovery Science*. Nice, France: Springer, 2022, pp. 286–301.

[34] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 2737–2778, 2019.

[35] N. Martinez, M. Bertran, and G. Sapiro, "Minimax Pareto fairness: A multi objective perspective," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6755–6764.

[36] J. Kang, T. Xie, X. Wu, R. Maciejewski, and H. Tong, "InfoFair: Information-theoretic intersectional fairness," in *Proc. IEEE Int. Conf. Big Data*, May 2022, pp. 1455–1464.

[37] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, Jan. 2012, pp. 214–226.

[38] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3315–3323.

[39] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5548–5557.

[40] J. Wang et al., "Generalizing to unseen domains: A survey on domain generalization," 2021, *arXiv:2103.03097*.

[41] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, p. 137.

[42] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[43] Z. Wang et al., "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8916–8925.

[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

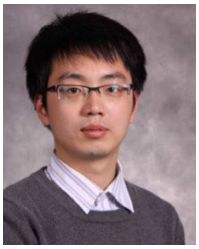
[47] J. Geralds, "Utkface large scale face dataset," Univ. Tennessee, Knoxville, TN, USA, 2017.

[48] C. Chen, Y. Liang, X. Xu, S. Xie, Y. Hong, and K. Shu, "When fairness meets privacy: Fair classification with semi-private sensitive attributes," in *Proc. NIPS*, 2022, pp. 1–13.

[49] F. Zhang, K. Kuang, L. Chen, Y. Liu, C. Wu, and J. Xiao, "Fairness-aware contrastive learning with partially annotated sensitive attributes," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022, pp. 1–18.

Huan Tian received the B.Sc. degree from the University of Shanghai for Science and Technology, Shanghai, China, in 2011, and the M.Sc. degree from TU Dortmund, Dortmund, Germany, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer Science, University of Technology Sydney, Ultimo, NSW, Australia.

His research interests include fairness and privacy, computer vision, and deep learning.



Bo Liu (Senior Member, IEEE) received the B.Eng. degree from the Department of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004, and the M.Eng. and Ph.D. degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2007 and 2010, respectively.

He is currently an Associate Professor with the University of Technology Sydney, Ultimo, NSW, Australia. His research interests include cybersecurity, privacy, location privacy, image privacy, privacy protection, and machine learning.

Tianqing Zhu received the B.Eng. and M.Eng. degrees from Wuhan University, Wuhan, China, in 2000 and 2004, respectively, and the Ph.D. degree from Deakin University, Sydney, Australia, in 2014.

She was a Lecturer with the School of Information Technology, Deakin University, from 2014 to 2018, and an Associate Professor with the University of Technology Sydney, Ultimo, NSW, Australia. She is currently a Professor with the Faculty of Data Science, City University of Macau, Macau, SAR, China. Her research interests include cyber security and privacy in artificial intelligence (AI).

Wanlei Zhou (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in computer science and engineering from Harbin Institute of Technology, Harbin, China, in 1982 and 1984, respectively, and the Ph.D. degree in computer science and engineering from The Australian National University, Canberra, ACT, Australia, in 1991, and the D.Sc. degree (Hons.) from Deakin University, Australia, in 2002.

He held various positions including the Head of the School of Computer Science, University of Technology Sydney, Ultimo, NSW, Australia; and the Alfred Deakin Professor, the Chair of Information Technology, the Associate Dean, and the Head of the School of Information Technology, Deakin University. He also served as a Lecturer with the University of Electronic Science and Technology of China, Chengdu, China; as a System Programmer with HP, MA, USA; as a Lecturer with Monash University, Melbourne, VIC, Australia; and as a Lecturer at the National University of Singapore, Singapore. He is currently the Vice Rector (Academic Affairs) and the Dean of the Institute of Data Science, City University of Macau, Macau, SAR, China. He has published more than 400 articles in refereed international journals and refereed international conference proceedings, including many articles in IEEE TRANSACTIONS and journals. His main research interests include security, privacy, and distributed computing.

Philip S. Yu (Life Fellow, IEEE) received the Bachelor of Science degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1970, the Master of Science and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1973 and 1976, respectively, and the Master of Business Administration degree from New York University, New York, NY, USA, in 1980.

He is currently a Distinguished Professor in computer science at the University of Illinois at Chicago, Chicago, IL, USA, and also holds the Wexler Chair in Information Technology. He has published more than 1600 articles in refereed journals and conferences. He holds or has applied for more than 300 U.S. patents. His research interests are on big data, and artificial intelligence (AI), including data mining, database, and privacy.

Dr. Yu is a fellow of the Association for Computing Machinery (ACM). He was a recipient of the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) 2016 Innovation Award, the IEEE Computer Society's 2013 Award, and the Research Contributions Award from International Conference on Data Mining (ICDM) in 2003 for his pioneering contributions to the field of data mining. He also received the Very Large Data Bases (VLDB) 2022 Test of Time Award, ACM Special Interest Group on Spatial Information (SIGSPATIAL) 2021 10-Year Impact Award, Web Search and Data Mining (WSDM) 2020 Honorable Mentions of the Test of Time Award, ICDM 2013 10-Year Highest-Impact Paper Award, and the Extending Database Technology (EDBT) Test of Time Award (2014). He was the Editor-in-Chief of *Transactions on Knowledge Discovery from Data* (journal) (TKDD) and *Transactions on Knowledge and Data Engineering* (TKDE).