IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 January 2024, accepted 19 January 2024, date of publication 24 January 2024, date of current version 8 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358275

RESEARCH ARTICLE

GBMix: Enhancing Fairness by Group-Balanced Mixup

SANGWOO HONG, (Graduate Student Member, IEEE), YOUNGSEOK YOON, (Student Member, IEEE), HYUNGJUN JOO, (Student Member, IEEE), AND JUNGWOO LEE, (Senior Member, IEEE)

Communications and Machine Learning Laboratory, Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Jungwoo Lee (junglee@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant 2021R1A2C2014504 (10%); in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Ministry of Science and ICT (MSIT), under Grant 2021-0-00106 (40%), Grant 2021-0-00180 (30%), and Grant 2021-0-02068 (20%); in part by Institute of New Media and Communications (INMAC); and in part by the Brain Korea (BK) 21-Four Program.

ABSTRACT Mixup is a powerful data augmentation strategy that has been shown to improve the generalization and adversarial robustness of machine learning classifiers, particularly in computer vision applications. Despite its simplicity and effectiveness, the impact of Mixup on the fairness of a model has not been thoroughly investigated yet. In this paper, we demonstrate that Mixup can perpetuate or even exacerbate bias presented in the training set. We provide insight to understand the reasons behind this behavior and propose GBMix, a group-balanced Mixup strategy to train fair classifiers. It groups the dataset based on their attributes and balances the Mixup ratio between the groups. Through the reorganization and balance of Mixup among groups, GBMix effectively enhances both average and worst-case accuracy concurrently. We empirically show that GBMix effectively mitigates bias in the training set and reduces the performance gap between groups. This effect is observed across a range of datasets and networks, and GBMix outperforms all the state-of-the-art methods.

INDEX TERMS Mixup, fairness, data augmentation, bias, spurious correlation.

I. INTRODUCTION

Machine learning and deep learning have seen remarkable success in recent years, particularly in computer vision. However, since typical machine learning models aim to optimize for average performance, there can still be significant performance discrepancies between underrepresented groups (which are at the tail of the data distribution) and those with larger populations. This is particularly problematic when spurious correlations exist in the data. A spurious correlation refers to a relationship between attributes that holds in most training examples but not in the test data, and it is typically caused by noise or bias in the sampling procedure

An example of spurious correlation is shown in Figure 1. In Figure 1, the number of samples in group Dark Hair-Male (Blond hair-Female) dominate the samples in group Blond hair-Male (Dark hair-Female), implying a

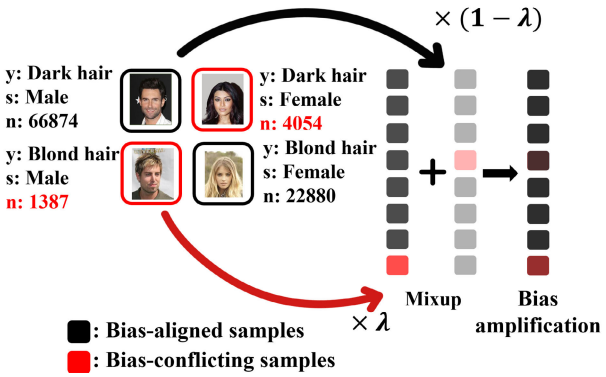


FIGURE 1. Bias amplification of Mixup in CelebA dataset. A spurious correlation between hair color and gender exists in the data and Mixup spreads out the bias to a larger part.

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

spurious correlation between hair color and gender. When the task is to predict hair color, a model may achieve high accuracy by learning spurious correlations (bias) between hair color and gender. However, this can lead to poor validation and test accuracy in groups where such correlations do not exist. Furthermore, learning such correlations on sensitive attributes such as race or gender may lead to the production of an unfair model.

To enhance the performance of the model when the data is imbalanced, various techniques have been suggested, often referred to as imbalanced learning [2], [4], [5], [8], [13], [22], [26], [27], [42], [46]. However, it has been known in [41] and [44] that common data augmentation techniques used for imbalanced data could not eliminate bias and even exacerbate unfairness in real-world data. Mixup [51], a famous and powerful data augmentation strategy, is not an exception. Mixup is an augmentation strategy that is proposed to train with interpolations of data samples. It reduces overfitting to the training set and enhances robustness to noise. However, when the dataset has spurious correlations as in Figure 1, Mixup spreads the spurious correlation to every data sample, resulting in the amplification of bias.

In this paper, we aim to increase the fairness of a model by enhancing the performance of the worst group. We first show that Mixup trained model often shows worse performance than the model trained with empirical risk minimization (ERM), degrading the performance of groups with bias-conflicting features. Then we demonstrate that this problem can be handled by modifying a Mixup strategy and propose GBMix, a group-balancing Mixup strategy that improves the fairness of a model.

A. Contributions: Our main contributions are as follows.

- We empirically demonstrate that Mixup on a biased dataset can aggravate unfairness. We also show that this problem is even more exacerbated as the level of spurious correlation in the dataset increases.
- We propose GBMix, a simple and effective Mixup strategy that mitigates spurious correlation. GBMix improves both the average accuracy and the accuracy of the worst-performing group.
- We compare GBMix with nine baselines across various datasets and architectures, and show that GBMix outperforms state-of-the art (SOTA) methods.

B. Organization: The remainder of the paper is organized as follows. In Section II, we introduce previous research on mixup, machine learning fairness, and neural network debiasing. In Section III, we formulate a classification problem for fair machine learning. In Section IV, we show that naive mixup can exacerbate the bias in the model and explain the reason behind this problem. In Section V, we propose GBMix and explain the overall process of GBMix. In Section VI we demonstrate the performance of GBMix via extensive experiments. In Section VII, we conclude our work.

II. RELATED WORK

A. MIXUP

Mixup [51] is a simple but powerful data augmentation strategy that synthesizes new data samples by linearly interpolating the paths between data samples. Empirical risk minimization (ERM) optimizes the risk

$$R_{ERM} = \mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)], \tag{1}$$

where P is the empirical distribution over the training dataset. On the other hand, Mixup performs a linear interpolation between samples to generate synthesized data as

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \tilde{y} = \lambda y_i + (1-\lambda)y_j, \tag{2}$$

where x_i and x_j are two distinct samples randomly drawn from training distribution and $\lambda \sim \texttt{Beta}(\alpha, \alpha)$. Mixup training minimizes the risk from the interpolated samples

$$R_{Mixup} = \mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim P}[\ell(f(\tilde{x}), \tilde{y})] \tag{3}$$

instead of R_{ERM}. By linearly interpolating and minimizing the risk between data samples, Mixup reduces overfitting to the training data and provides robustness to noise.

Due to its simplicity and efficacy, many variations of Mixup have been suggested. For example, Manifold Mixup [39] performs a linear interpolation in the embedding space and provides further generalization by capturing high-level information in the feature space. On the other hand, CutMix [47] randomly samples patches from a sample and pastes them on the other sample. CutMix improves the performance of various tasks such as classification or object detection by combining cutout and Mixup. Numerous studies also proposed variations of Mixup that utilize GAN to generate the label for mixed samples [37], uses saliency information to select the mask for mixup [19], [23], and uses Mixup as a regularizer to improve OOD robustness [30].

Moreover, strategies that incorporate Mixup for handling imbalanced data have also been suggested. In [5], ReMix utilizes different λ for x and y as λ_x and λ_y. It puts more weight on y from minority groups more by controlling λ_y. The authors of [8] suggested instance-based and class-based sampling for Mixup, which improves the performance in a highly imbalanced dataset. More recently, UniMix [43] has been suggested to handle the long-tailed distribution of the dataset and the authors of [53] have proposed a method of mixing the data after oversampling to enhance the performance of Mixup in long-tailed distribution.

On the other hand, FairMix has been suggested in [6] to enhance fairness in machine learning. FairMix divides the data samples based on their attributes and interpolates between data samples from different groups. The Jacobian of the interpolated sample is then used as a regularizer term in addition to R_{ERM}. FairMix shows better generalization for both accuracy and fairness in various datasets. However, to the best of our knowledge, the effects of training a model with the mixed data sample itself on the fairness of a model have not been investigated thoroughly yet.

B. MACHINE LEARNING FAIRNESS

Recently, many methods have emerged to tackle the issue of unfairness within machine learning models. These methods can be classified into different categories based on their approach (pre-processing, in-processing, and post-processing) as well as the specific fairness metrics they target (group fairness, individual fairness, and Rawlsian Max-Min fairness).

Pre-processing methods [7], [15], [16], [50] modify the training set to eliminate inherent biases, thereby enhancing the efficacy of general methods. On the other hand, in-processing methods [18], [34], [49] adjust the training objective to cultivate fair models, while post-processing methods [17], [31] refine the model prediction to ensure fairness.

Group fairness [7], [11] considers the statistical measurement of the whole dataset, while individual fairness [29], [35], [48] measures the similarity of model outputs for similar inputs. Individual fairness aims to ensure that similar individuals get similar outputs. Lastly, Rawlsian Max-Min fairness [3], [12], [21], [33] seeks to improve the performance of the worst-performing group and narrow the performance disparities between groups while maintaining the overall performance.

In this paper, we target **Rawlsian Max-Min fairness**, which aims to enhance the fairness models by *increasing the worst group performance* and narrowing the performance gap between groups, *while maintaining the overall performance*.

C. DISTRIBUTIONALLY ROBUST OPTIMIZATION AND NEURAL NETWORK DEBIASING

In the literature, attempts to enhance the performance of the most disadvantaged group have been addressed through approaches such as Distributional Robust Optimization (DRO) and network debiasing. One famous method is group distributionally robust optimization (GDRO) [36]. GDRO stems from distributionally robust optimization (DRO) [1], [28] which aims to minimize the worst-case loss over the set of possible test distributions. Since it is intractable to minimize the risk over all possible test distributions, GDRO focuses on optimizing the worst group performance over the conditional distribution of data samples associated with groups.

Additionally, several methods have focused on training unbiased neural networks when working with datasets that exhibit spurious correlations, which is called debiasing. The authors of [45] have enhanced the robustness of the model by selectively interpolating the samples, and sample-wise reweighting was used in [10]. However, the relationship between the utilization of Mixup and any potential decline in model performance has yet to be thoroughly examined.

More recently, it has been discovered in [20] that retrains only the last layer with a balanced dataset can remarkably improve the worst group accuracy of a model. This highlights the importance of learning a robust feature space. Furthermore, there have been studies to enhance the worst-group accuracy without accessing sensitive attributes [24], [32], [38], [52]. Nonetheless, these approaches still exhibit inferior performance in comparison to those that leverage sensitive attribute information.

III. PROBLEM STATEMENT AND OBJECTIVE

We now formulate our objective. In this paper, we consider a classification problem that aims to predict target attribute $y \in \mathcal{Y}$ from provided input features $x \in \mathcal{X}$. Each sample has a sensitive attribute $s \in \mathcal{S}$ which should not affect the prediction on the target attribute $\tilde{y}$. Our goal is to train a fair classifier that performs well for all subgroups organized from all possible combinations of y and s. Therefore, we divide the dataset into m ($m = |\mathcal{Z}| = |\mathcal{Y}| \times |\mathcal{S}|$) groups, where $\mathcal{Z} = \mathcal{Y} \times \mathcal{S}$ is a Cartesian product of $\mathcal{Y}$ and $\mathcal{S}$, and aim to minimize the risk of the worst-performing group among m groups as follows

$$R(f, D) = \max_{z \in \mathcal{Z}} \mathbb{E}_{P_z}[\ell(f(x), y)]. \tag{4}$$

Since we consider **Rawlsian Max-Min fairness**, we now define two fairness metrics that have been used in many fairness researches for Rawlsian Max-Min fairness, which are *worst group accuracy* and *robust gap*. Worst group accuracy is the worst-case performance among all groups, and it can be expressed as

$$\min_{z \in \mathcal{Z}} \mathbb{E}_{P_z}[\mathbb{I}(f(x) = y)], \tag{5}$$

where $\mathbb{I}$ denotes the indicator function that evaluates to 1 if the classifier's prediction matches the true label y. The robust gap is the gap between the best and worst-performing groups and can be expressed as

$$\max_{z \in \mathcal{Z}} \mathbb{E}_{P_z}[\mathbb{I}(f(x) = y)] - \min_{z \in \mathcal{Z}} \mathbb{E}_{P_z}[\mathbb{I}(f(x) = y)]. \tag{6}$$

Therefore, it is desirable to maximize the worst group accuracy, and minimize the robust gap while maintaining high average accuracy.

IV. LIMITATION OF MIXUP

In this section, we demonstrate that vanilla Mixup training does not work well when there is a spurious correlation in the training dataset. It fails to improve both the average and the worst group accuracy of neural networks. Moreover, in some cases, even increases the performance gap between groups and degrades fairness. We use the neural networks trained on three datasets (Waterbird [36], CelebA [25], and UTKFace [54]) using ERM and Mixup objectives to demonstrate this tendency.

Table 1 demonstrates that Mixup failed to improve the average accuracy on Waterbird, and degraded the worst group accuracy and robust gap for all three datasets. This implies that Mixup cannot improve the generalization of neural networks if a training set has a spurious correlation.

To further analyze the relationship between Mixup and fairness, we adjust the CelebA dataset to provide artificial

TABLE 1. Average accuracy, worst group accuracy, and robust gap of ERM and Mixup trained models on three datasets.

ResNet-18		Waterbird	CelebA	UTKFace
Average Acc. ↑ (%)	ERM	86.1	95.4	91.2
	Mixup	80.4	95.9	92.2
Worst group Acc. ↑ (%)	ERM	63.6	49.7	70.2
	Mixup	57.6	39.1	67.9
Robust Gap ↓ (%)	ERM	30.8	49.2	25.2
	Mixup	41.3	60.6	29.0

imbalance ratios. We undersample the CelebA dataset to have different ratios between gender and hair color from 1:2 to 1:8, making different degrees of spurious correlation in a training set. We train a neural network in these artificial scenarios, and indicate the results in Table 2.[1] Although Mixup can improve the models when the imbalance ratio is relatively low (1 : 2), it works poorly as the imbalance ratio increases.

TABLE 2. Average accuracy, worst group accuracy, and robust gap of ERM and Mixup on three undersampled CelebA datasets.

Imbalance ratio in gender		1 : 2	1 : 4	1 : 8
Average Acc. ↑ (%)	ERM	91.9	92.2	91.9
	Mixup	91.8	92.3	91.6
Robust Acc. ↑ (%)	ERM	78.2	70.5	63.9
	Mixup	84.7	68.9	59.4
Robust Gap ↓ (%)	ERM	17.4	26.0	33.7
	Mixup	5.9	28.2	38.9

When there is a severe spurious correlation in the dataset, Mixup synthesizes more data samples inheriting this correlation. Therefore, most of the mixed samples would contain spurious correlation, which implies Mixup spreads spurious correlation to the majority of data. If the dataset does not contain spurious correlation, i.e., if the dataset is balanced for all groups, mixup would not exacerbate bias and improve generalizability. However, there may be limited amounts of data for certain classes or categories depending on the specific domain, so constructing a balanced dataset for all groups may be impractical. Therefore, we propose GBMix, a simple and effective mixup strategy that provides generalizability while mitigating bias as well.

V. GBMix

To handle the bias-exacerbating problem of Mixup on biased datasets, we now propose GBMix. The key idea of GBMix is to control the mixup target and ratio between groups. By generating mixed samples focusing on the worst-performing group, GBMix can mitigate spurious correlation in the dataset and learn a robust feature space. In this section, we explain the process of GBMix in detail and the overall process of GBMix is summarized in Figure 2.

A. TARGET GROUP DECISION

GBMix first decides the target group and conducts group-balanced mixup based on the target group. The target group is decided by the average loss of each group at each mini-batch. GBMix first makes a mini-batch by sampling an equal number of samples from each group to compute the loss of each group accurately. Then, it divides the mini-batch into groups based on attributes y and s and computes the average loss for each group.

Since our goal is to boost the performance of the worst-performing group, GBMix chooses the worst group (the group with the largest loss among $m = |\mathcal{Z}|$ groups in the mini-batch) as the target group and denotes it as

$$g_0 = \arg\max_{z \in \mathcal{Z}} \mathbb{E}_{P_z}[\ell(f(x), y)]. \tag{7}$$

The loss for every group is calculated in each mini-batch iteration and g_0 is determined at every mini-batch. Let us denote the target attribute and sensitive attribute of data samples in g_0 as y_t and s_t.

B. GROUP CATEGORIZATION

To boost the performance of the worst-performing group by Mixup, GBMix uses group categorization before mixing samples. The group categorization helps the network focus on g_0. After selecting the target group g_0, the remaining $m - 1$ groups are categorized into three larger groups as in Figure 2. The groups with $y = y_t$ and $s \neq s_t$ are regrouped as g_1, and groups with $y \neq y_t$ and $s = s_t$ are regrouped as g_2. The other groups with $y \neq y_t$ and $s \neq s_t$ are regrouped as g_3. Then possible Mixup sample combinations from the four groups can be expressed as

$$\begin{aligned} \tilde{x}_{g_{(i,j)}} &= \lambda x_{g_i} + (1-\lambda) x_{g_j}, \\ \tilde{y}_{g_{(i,j)}} &= \lambda y_{g_i} + (1-\lambda) y_{g_j}, \\ R_{g_{(i,j)}} &= \mathbb{E}_P[\ell(f(\tilde{x}_{g_{(i,j)}}), \tilde{y}_{g_{(i,j)}})] \\ &\qquad \forall i,\ j \in [0:3] \ \text{ and } \ i > j, \end{aligned} \tag{8}$$

where $\lambda \sim \texttt{Beta}(\alpha, \alpha)$ and x_{g_i} and x_{g_j} are two samples randomly drawn from g_i and g_j, respectively. By categorizing the groups into four larger groups based on their attributes, GBMix can manage non-binary attributes (when $|\mathcal{Y}| > 2$ or $|\mathcal{S}| > 2$) effectively and allows the model to concentrate on the worst-performing group.

C. BALANCING MIXUP

Let us assume that g_0 is an underrepresented group that suffers from spurious correlation in the training data. This implies that there exists a spurious correlation between y_t and $\{s_t\}^c$ or s_t and $\{y_t\}^c$. To alleviate the effect of spurious correlation during training, GBMix balances the number of mixed samples between groups based on g_0, and it can be formulated as

$$|\tilde{x}_{g_{(i,j)}}| = k, \quad \forall i,\ j \in [0:3] \ \text{ and } \ i \geq j, \tag{9}$$

where k is a constant number that limits the number of samples generated from Mixup. For large k, the samples from the minority are required to be resampled for Mixup. On the other hand, using small k will discard samples from

[1] More detailed experimental settings are given in the Appendix.

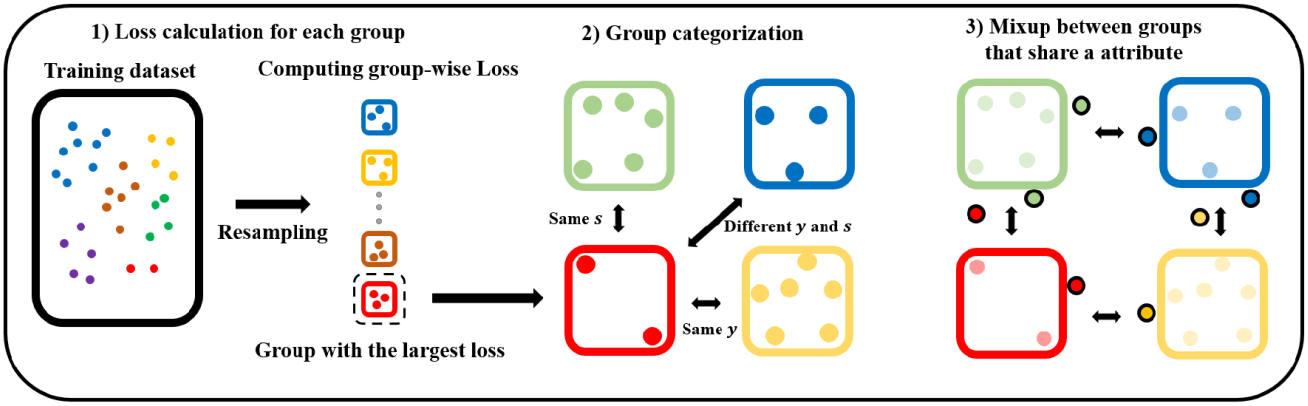


FIGURE 2. Step-wise process of GBMix.

TABLE 3. Performance of GBMix and compared schemes on binary classification tasks.

Methods	Waterbirds			CelebA		
	Worst (%)	Gap (%)	Average (%)	Worst (%)	Gap (%)	Average (%)
ERM	77.9±0.8	21.8±0.8	91.1±0.3	48.8±3.4	50.3±3.6	95.7±0.3
SUBG	89.1±1.1	-	-	85.6±2.3	-	-
Mixup	76.8±1.0	20.2±0.6	91.3±0.3	46.5±1.6	52.8±1.6	**95.8**±0.04
FairMix	86.6±0.2	10.7±0.8	91.9±0.4	84.7±1.4	9.5±1.0	92.5±0.2
OBMix	87.4±0.3	8.3±0.9	91.6±0.5	85.7±0.9	8.4±1.0	92.5±0.6
UniMix	78.3±0.2	21.2±0.6	90.2±0.5	43.7±2.3	55.7±2.4	95.8±0.1
GDRO	91.4	-	93.5	88.9	-	92.9
JTT	86.7	-	93.3	81.1	-	88
DFR	92.9±0.2	-	94.2±0.4	88.3±1.1	-	91.3±0.3
GBMix (proposed)	**94.5**±0.3	**4.0**±0.7	**95.8**±0.2	**90.6**±0.2	**5.0**±0.6	91.9±0.2

the majority, and it corresponds to undersampling. In this paper, we choose k as the average number of data samples from each group.

GBMix mixes samples between groups that either share y or s. By mixing samples that share one identical attribute and one different attribute, GBMix can mitigate spurious correlations between attributes that might exist in the data and provide generalization. The loss for GBMix can be expressed as

$$R_{GBMix} = R_{g(0,1)} + R_{g(0,2)} + R_{g(1,3)} + R_{g(2,3)}. \quad (10)$$

D. CLASSIFIER FINE TUNING

After training using R_{GBMix} as a loss, we retrain the last fully connected layer using a small and balanced set-aside training dataset as in [20]. It was first observed in [20] that retraining the last layer using a small-size dataset that does not contain spurious correlation can mitigate the bias of the neural networks, even when trained on datasets with spurious features. This implies that if the network has learned good feature space, bias can be easily mitigated by retraining the classifier, pointing to the importance of learning a good feature extractor. The balanced mixup strategy of GBMix facilitates the acquisition of robust feature space, and by fine-tuning the last layer, GBMix outperforms the state-of-the-art method.

VI. EXPERIMENT

We now provide experimental results that demonstrate the effectiveness of GBMix in mitigating bias. GBMix was evaluated on four datasets, Waterbird, CelebA, UTKFace, and FairFace. Waterbird and CelebA are binary datasets which are commonly used as benchmarks for evaluating the bias of models. We also used two non-binary facial datasets, UTKFace and FairFace, which are commonly used for fairness evaluation. For network architecture, we use ResNet-50 initialized with weights pre-trained on ImageNet following prior works [14], [20], [24]. We also experimented with two additional architectures (ResNet-18 and MobilenetV2) and provide the results in the Appendix A-A.

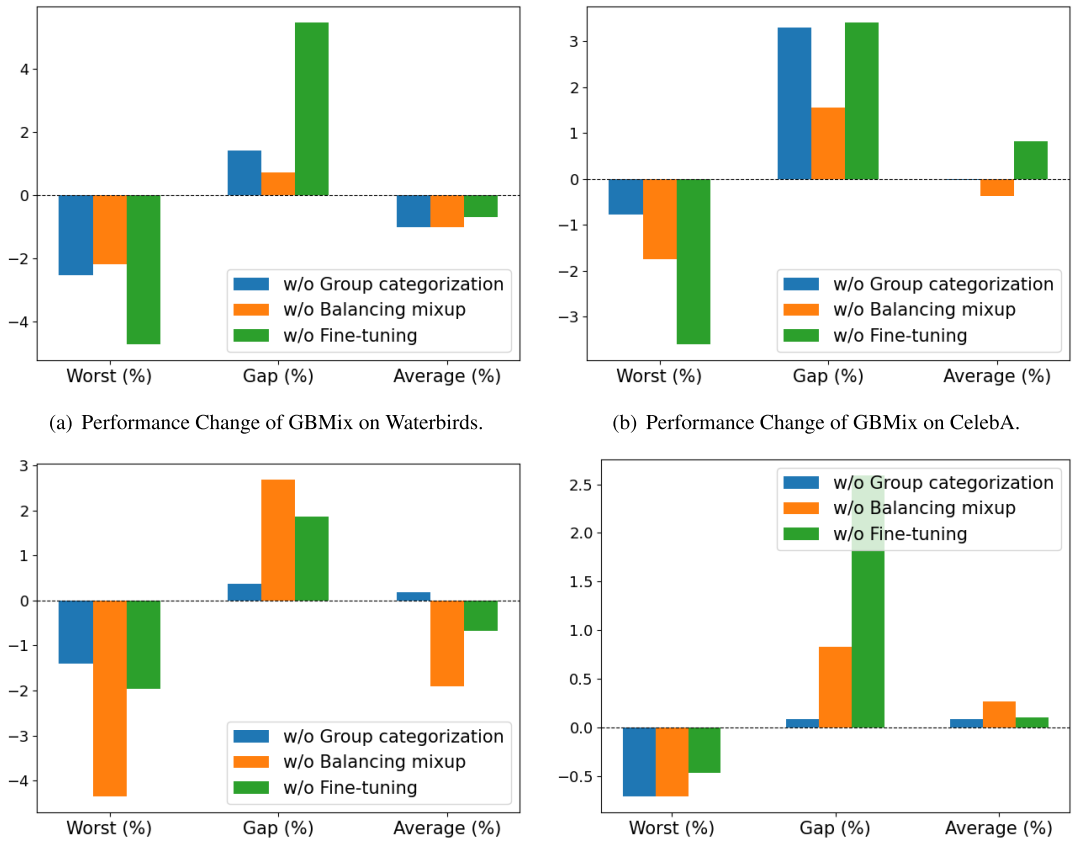
We compared GBMix with nine baselines, including ERM, Mixup [51], SUBG [14], GDRO [36], JTT [24], DFR [20], UniMix [43], OBMix [53], and FairMix [6]. Additionally, we provide an ablation study to demonstrate the role of each step in GBMix.

Datasets. First, we will briefly explain four datasets. The detailed experimental settings are given in Table 5.

CelebA. CelebA is a dataset composed of faces of celebrities with various attributes. We choose hair color $y \in \{\texttt{blond}, \texttt{dark hair}\}$ as a target attribute y and gender $s \in \{\texttt{male}, \texttt{female}\}$ as a sensitive attribute s as in [9]

TABLE 4. Performance of GBMix and compared schemes on non-binary classification tasks.

Methods	UTKFace			FairFace		
	Worst (%)	Gap (%)	Average (%)	Worst (%)	Gap (%)	Average (%)
ERM	60.1±0.2	36.1±5.1	91.1±4.5	85.1±0.2	9.2±0.2	91.8±0.1
SUBG	69.1±1.1	23.2±1.0	87.9±0.2	86.5±1.0	8.6±1.0	92.0±0.1
Mixup	61.3±4.7	36.0±5.2	92.1±0.1	87.1±1.1	7.3±1.0	92.3±0.04
FairMix	70.8±3.4	22.4±2.9	89.2±0.4	85.6±0.5	8.7±0.7	91.6±0.2
OBMix	75.6±3.4	17.4±3.8	88.7±0.2	87.2±0.5	7.0±0.6	92.3±0.1
UniMix	65.0±4.6	32.5±4.8	**92.3**±0.2	87.1±0.4	7.5±0.7	**92.5**±0.2
GDRO	75.7±5.3	17.5±5.5	87.1±1.3	86.3±1.5	8.0±1.6	91.9±0.1
JTT	55.5±3.1	40.4±2.9	88.9±0.6	84.8±0.9	8.5±0.9	90.8±0.4
DFR	68.6±2.4	27.0±3.5	84.5±1.7	86.6±0.6	8.0±0.6	92.1±0.2
GBMix (proposed)	**80.9**±0.3	**13.5**±0.5	88.8±1.0	**87.3**±0.1	**6.8**±0.7	91.9±0.2



(a) Performance Change of GBMix on Waterbirds.

(b) Performance Change of GBMix on CelebA.

(c) Performance Change of GBMix on UTKFace.

(d) Performance Change of GBMix on FairFace.

FIGURE 3. Performance gap between GBMix and ablation baselines.

TABLE 5. Experimental settings for the datasets.

Dataset	Waterbirds	CelebA	UTKFace	FairFace
Target attribute (y)	{waterbird, landbird}	{blond, non-blond}	{-19,20-59,60-}	{White, Black, Others}
Sensitive attribute (s)	{water background, land background}	{male, female}	{White, Black, Others}	{-19,20-59,60-}
# of groups ($\lvert\mathcal{Y}\rvert \times \lvert\mathcal{S}\rvert$)	4 ($=2\times2$)	4 ($=2\times2$)	& 9 ($=3\times3$)	& 9 ($=3\times3$)

and [36]. It should be noted that CelebA primarily focuses on images of celebrities from the entertainment industry, which may not represent a diverse range of appearances and backgrounds. This lack of diversity can limit the generalization of models trained on the dataset to other demographics.

Waterbird. Waterbird is a dataset for simulating spurious correlation created in [36] by combining bird

photograph from CUB [40] dataset and background images from Places [55] dataset. In waterbird, a bird $y \in \{\texttt{waterbird}, \texttt{landbird}\}$ is placed in background $s \in \{\texttt{water background}, \texttt{land background}\}$ and Waterbirds (Landbirds) appear more frequently against water background (land background) than land background (water background).

UTKFace. UTKFace is a facial dataset widely used for multi-group classification benchmark. We set age as target attribute y and race as the sensitive attributes s. We use a multi-class classification setting (where *target attributes are non-binary*) on the UTKFace dataset. The age distribution in the dataset is not perfectly balanced, with more samples available for certain age groups than others. This class imbalance can affect the performance of machine learning models, particularly when attempting to classify less-represented age categories.

FairFace. FairFace is a balanced facial dataset that includes a similar number of samples for attributes. The dataset includes images of people from different racial and ethnic groups, and it is balanced for gender and race. We use race as target attribute y and age as sensitive attribute s. We also use a multi-class classification setting on the FairFace dataset.

Considering that both the UTKFace and FairFace datasets are non-binary, achieving high performance on these datasets is crucial for applying the debiasing method in real-world scenarios, where most of the lables are non-binary.

Training information. We use ResNet-50 with pre-trained ImageNet weights, and all models are trained with Stochastic Gradient Descent (SGD). We train for 200 epochs for Waterbird, 25 epochs for CelebA, and 80 epochs for UTKFace and FairFace. More detailed experimental settings are given in the Appendix.

A. COMPARISON

We first conduct experiments on binary-classification tasks, using Waterbirnd and CelebA datasets. The results are presented in Table 3. In Waterbirds and CelebA, we report the results from the original paper for SUBG, LISA, GDRO, JTT, and DFR, and we report the mean±std over three independent runs for Mixup, OBMix, UniMix, and GBMix.

GBMix demonstrates significant performance improvement on both datasets, improving worst group accuracy and narrowing the performance gap. More specifically, GBMix achieves the highest worst group accuracy and the lowest performance gap among all the baselines in both datasets and also achieves the best average accuracy on Waterbirds. On CelebA, Mixup achieves better average accuracy (3.9% higher) than GBMix, but shows much lower worst group accuracy (44.1% lower) and higher performance gap (47.8% higher) than GBMix.

We also conduct experiments on multi-group classification tasks using UTKFace and FairFace. The results are summarized in Table 4. Unlike Waterbird or CelebA where y is binary and the number of groups is relatively small (4 in Waterbird and CelebA), UTKFace and FairFace have a non-binary attribute and thus have a larger number of groups. Unlike other mixup schemes that only focus on balancing the data subpopulation, GBMix focuses on the worst group and providing diverse samples for them. By interpolating and synthesizing the samples adaptively based on the worst-performing group, GBMix can enhance the worst group accuracy effectively. Specifically, GBMix achieves the highest worst group accuracy and the lowest performance gap in both of the datasets.

To further evaluate the performance of GBMix and compare with the baselines, we make a ranking for the average accuracy, worst group accuracy, and worst gap for four datasets and compute the average ranking for binary and non-binary datasets. The average rankings are indicated in Table 6. It should be noted that, while SOTA methods such as GDRO and DFR show good performance on binary classification tasks, their effectiveness substantially diminishes when applied to non-binary classification tasks, where most previous studies have not focused on. In contrast, mixup-based strategies such as OBMix and UniMix demonstrate superior performance compared to DFR and GDRO in non-binary classification tasks by synthesizing samples for underrepresented groups. However, their performance diminishes in binary classification tasks. On the other hand, GBMix consistently achieves the best-worst group accuracy and the lowest performance gap across both binary and non-binary datasets, all while maintaining a high average accuracy, demonstrating the superiority of GBMix.

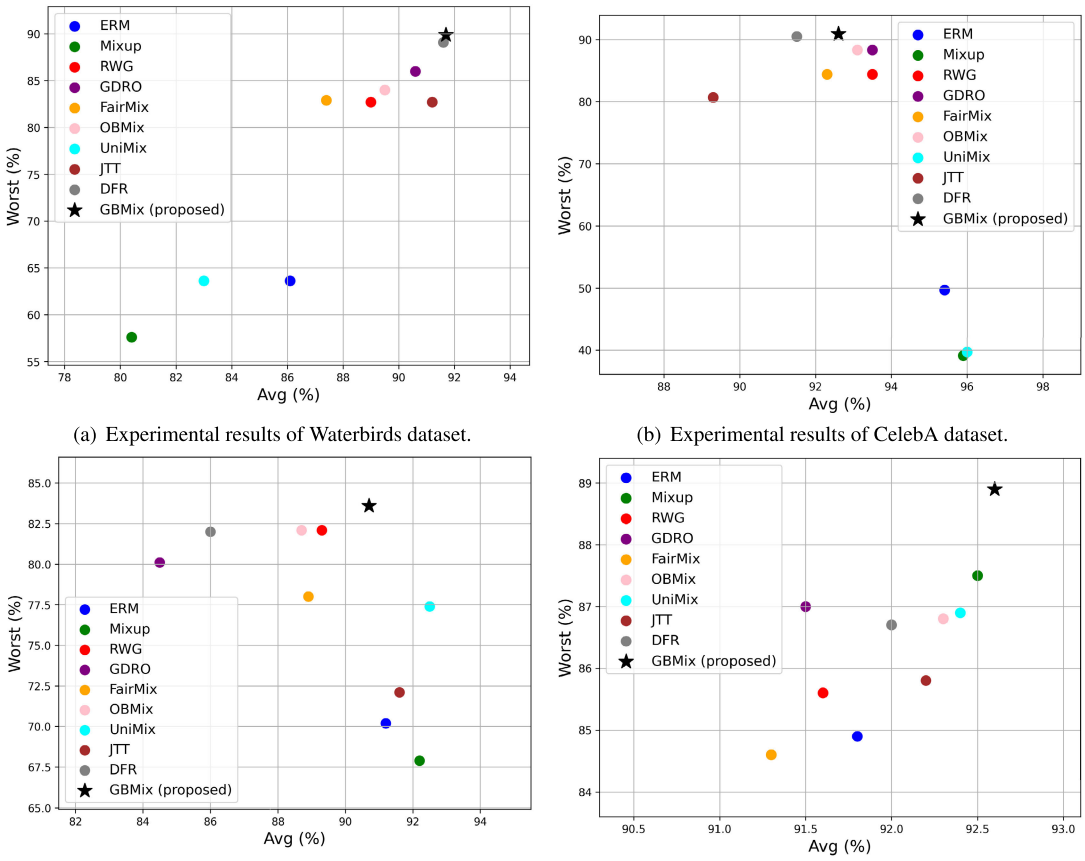
TABLE 6. Average ranking of GBMix and baselines on binary and non-binary datasets.

Methods	Ranking	
	Binary datasets	Non-binary datasets
ERM	6.3	8.0
RWG	4.5	6.2
Mixup	6.0	4.0
FairMix	5.0	6.3
OBMix	4.0	3.3
UniMix	6.7	4.0
GDRO	3.0	5.7
JTT	6.5	8.7
DFR	3.8	6.0
GBMix	**2.0**	**2.8**

B. ABLATION STUDIES

In order to gain a deeper understanding of the role of GBMix, we conduct an ablation study. GBMix has three major components, which are group categorization, balancing mixup, and fine-tuning the last FC-layer. To isolate the benefit of each step in GBMix, we compare GBMix with **i) without group categorization** which uses vanilla Mixup without group categorization into four larger groups, **ii) without balancing mixup** that uses group categorization but does not use mixup training, and **iii) without fine-tuning** that does not fine-tune the last FC-layer.

We set the x-axis as the performance of GBMix to show the performance change of GBMix, and indicate



(a) Experimental results of Waterbirds dataset.

(b) Experimental results of CelebA dataset.

(c) Experimental results of UTKFace dataset.

(d) Experimental results of FairFace dataset.

FIGURE 4. Performance of GBMix and baselines on ResNet-18 network.

the performance gap between GBMix and the ablated methods in Figure 3. **Negative** average accuracy change or worst group accuracy change indicates that the considered method has deteriorated (smaller) average or worst group accuracy compared to GBMix, and a **Positive** robust gap change indicates that the considered method has deteriorated (larger) robust gap compared to GBMix. Therefore, if an ablated method has a larger change than others, it implies that the missing component in the ablated method is the essential component for the performance of GBMix.

On binary classification tasks (Waterbirds or CelebA), fine-tuning is the most important component that affects the performance of GBMix, while balanced mixup and fine-tuning are the primary source of performance enhancement on non-binary classification tasks. By using both balanced mixup and fine-tuning, GBMix can outperform SOTA methods in both cases. This finding is similar to the results in Section VI-A, where DFR (which uses fine-tuning) shows relatively low performance in non-binary datasets and mixup based strategies show relatively low performance in binary datasets. Notably, Group categorization enhances the worst group accuracy to a large extent on non-binary datasets, effectively making the model focus on the worst-performing group.

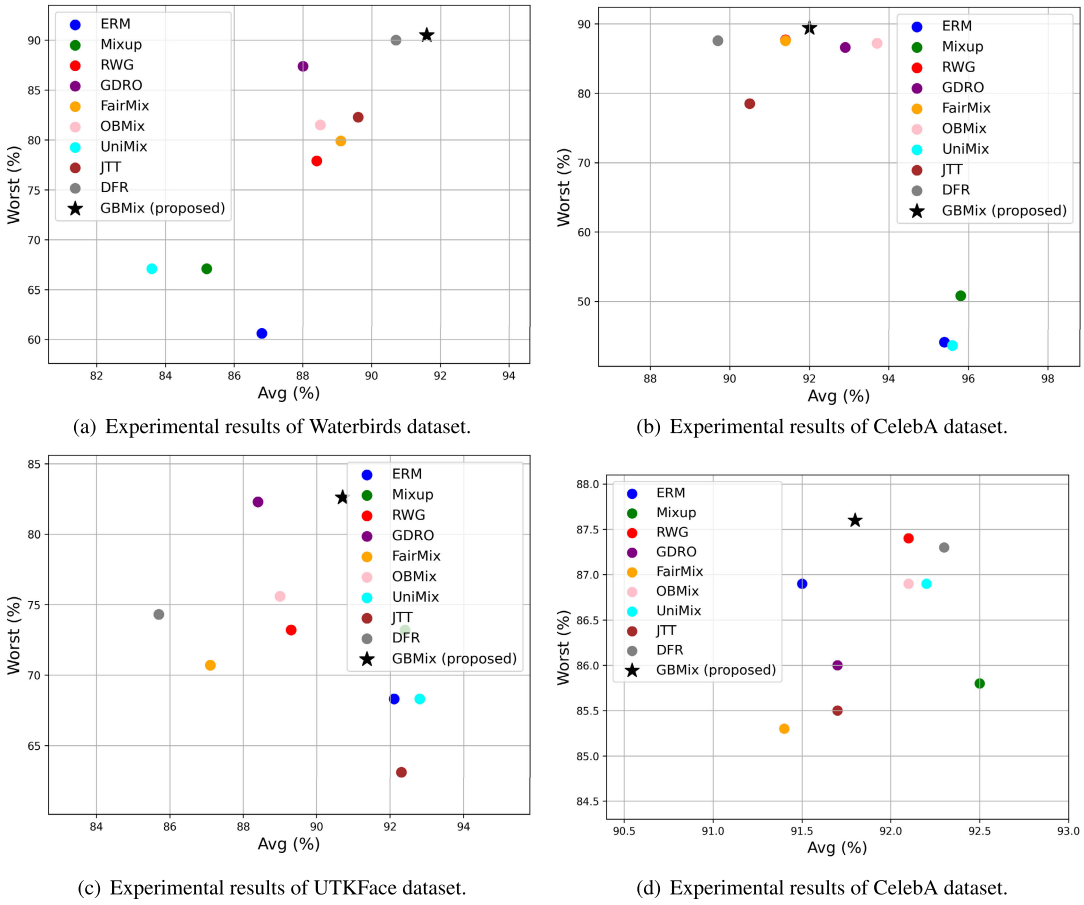
VII. CONCLUSION

In this paper, we demonstrate that Mixup may exacerbate bias in training data and result in unfair models. To address this issue, we propose GBMix, a group-balanced Mixup strategy that effectively mitigates bias and narrows the performance gap between groups. GBMix categorizes groups into four larger groups and balances the mixup ratio between them. GBMix mitigates bias and improves performance by balancing the mixup ratio for categorized groups and preventing overfitting in minority groups. Our experimental results in various scenarios demonstrate the effectiveness of GBMix, and an ablation study is provided to isolate and explore the role of each component in GBMix. Future work includes extending GBMix to settings where multiple sensitive attributes exist, and using the mixup ratio adaptively depending on the imbalance structure of the data. Moreover, using GBMix when the sensitive attribute s is not known would be an interesting research topic.

APPENDIX A
SUPPLEMENTARY EXPERIMENTAL RESULTS

A. EXPERIMENTS ON VARIOUS NETWORK ARCHITECTURES

In Section VI-A, we only presented the experimental results for Resnet-50. We now provide additional experimental



(a) Experimental results of Waterbirds dataset.

(b) Experimental results of CelebA dataset.

(c) Experimental results of UTKFace dataset.

(d) Experimental results of CelebA dataset.

FIGURE 5. Performance of GBMix and baselines on Mobilenet-V2 network.

TABLE 7. Experimental settings for binary classification task in Table 3.

Dataset	Waterbird	CelebA
Target attribute y	{waterbird, landbird}	{blond, non-blond}
Sensitive attribute s	{water background, land background}	{male, female}
# of groups ($\lvert\mathcal{Y}\rvert \times \lvert\mathcal{S}\rvert$)	$4\ (=2\times 2)$	$4\ (=2\times 2)$

results of four datasets across two additional network architectures (ResNet-18 and MobilenetV2). We use the same settings as Section VI-A, which are given in Appendix B-C, and indicate the results in Figure 4 and 5. Our experimental results indicate that GBMix consistently achieves the best performance across various network architectures, indicating the superiority of GBMix in enhancing the fairness of a model.

APPENDIX B
EXPERIMENTAL DETAILS

A. COMMON SETTINGS

For every method, we sweep over learning rates {1e-3, 1e-4, 1e-5}. However, since GDRO requires additional tuning of weight decay and adjustment parameters, we set the weight decay to 1 for Waterbird, 0.1 for CelebA following [27], and swept over {0.1, 1} for UTKFace and FairFace. For all other methods except GDRO, we used a weight decay of 1e-4. In addition, we conducted a sweep over adjust parameters {1, 3, 5} for GDRO, and a parameter $\gamma_{FairMix}$ (which balances R_{ERM} and $R_{FairMix}$) {1, 3, 5} for FairMix method.

We kept the batch size fixed to 64 and repeated all experiments three times. To evaluate the effectiveness of the different methods, we measured their average accuracy, worst group accuracy, and robust gap across all datasets. We select the model with the best worst group accuracy on a valid dataset as a criterion.

B. SETTINGS FOR EXPERIMENTS IN SECTION IV

For ERM and Mixup comparison experiments in Section IV, we use the target attribute and sensitive attribute following Table 7. Furthermore, we undersample the CelebA dataset to make a spurious correlation between hair color and gender.

TABLE 8. Number of training data samples in each group of undersampled CelebA dataset used for the experiments in Table 2.

Imbalance ratio	dark hair, female	dark hair, male	blond hair, female	blond hair, male
1 : 2	4054	8359	2680	1387
1 : 4	4054	66874	22880	1387
1 : 8	4054	33437	11440	1387

We also indicate the distribution of the undersampled CelebA training dataset in Table 8.

C. SETTINGS FOR EXPERIMENTS IN SECTION VI-A

In Section VI-A, we demonstrate the performance of GBMix on binary classification tasks and multi-label classification tasks. For binary classification tasks (Waterbirds and CelebA), we used the same target and sensitive attribute settings with experiments in Section IV, which are summarized in Table 7. For multi-label classification tasks (UTKFace and FairFace), we set age as target attribute y and use race as sensitive attribute s and race as target attribute y and use age as sensitive attribute s, respectively.

REFERENCES

[1] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Manage. Sci.*, vol. 59, no. 2, pp. 341–357, Feb. 2013.

[2] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.

[3] J. Chai, T. Jang, and X. Wang, "Fairness without demographics through knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 19152–19164.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[5] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: Rebalanced mixup," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 95–110.

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

[7] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 259–268.

[8] A. Galdran, G. Carneiro, and M. A. G. Ballester, "Balanced-mixup for highly imbalanced medical image classification," in *Medical Image Computing and Computer Assisted Intervention*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham, Switzerland: Springer, 2021, pp. 323–333.

[9] K. Goel, A. Gu, Y. Li, and C. Re, "Model patching: Closing the subgroup performance gap with data augmentation," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[10] Z. Han, Z. Liang, F. Yang, L. Liu, L. Li, Y. Bian, P. Zhao, B. Wu, C. Zhang, and J. Yao, "UMIX: Improving importance weighting for subpopulation shift via uncertainty-aware mixup," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 37704–37718.

[11] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2016.

[12] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 1929–1938.

[13] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5375–5384.

[14] B. Y. Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz, "Simple data balancing achieves competitive worst-group-accuracy," in *Proc. 1st Conf. Causal Learn. Reasoning*, vol. 177, pp. 336–351.

[15] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 702–712.

[16] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, Oct. 2012.

[17] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 924–929.

[18] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, pp. 35–50.

[19] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5275–5285.

[20] P. Kirichenko, P. Lzmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," in *Proc. Int. Conf. Learn. Represent.*, 2023.

[21] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 728–740.

[22] F. Last, G. Douzas, and F. Bacao, "Oversampling for imbalanced learning based on K-Means and SMOTE," 2017, *arXiv:1711.00837*.

[23] H. Li, X. Zhang, Q. Tian, and H. Xiong, "Attribute Mix: Semantic data augmentation for fine grained recognition," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2020, pp. 243–246.

[24] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 6781–6792.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[26] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 181–196.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[28] Y. Oren, S. Sagawa, T. Hashimoto, and P. Liang, "Distributionally robust language modeling," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4227–4237.

[29] M. Peychev, A. Ruoss, M. Balunović, M. Baader, and M. Vechev, "Latent space smoothing for individually fair representations," in *Computer Vision—ECCV 2022*. Springer, 2022, pp. 535–554.

[30] F. Pinto, H. Yang, S. N. Lim, P. Torr, and P. Dokania, "Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 14608–14622.

[31] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[32] S. Qiu, A. Potapczynski, P. Izmailov, and A. G. Wilson, "Simple and fast group robustness by automatic feature reweighting," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, 2022, pp. 28448–28467.

[33] J. Rawls, *Justice As Fairness: A Restatement*. Cambridge, MA, USA: Harvard Univ. Press, 2001.

[34] Y. Roh, K. Lee, S. Whang, and C. Suh, "FR-train: A mutual information-based approach to fair and robust training," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 8147–8157.
[35] A. Ruoss, M. Balunović, M. Fischer, and M. Vechev, "Learning certified individually fair representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7584–7596.
[36] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[37] J.-Y. Sohn, L. Shang, H. Chen, J. Moon, D. Papailiopoulos, and K. Lee, "GenLabel: Mixup relabeling using generative models," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, Jul. 2022, pp. 20278–20313.
[38] N. Sohoni, J. Dunnmon, G. Angus, A. Gu, and C. Ré, "No subclass left behind: Fine-grained robustness in coarse-grained classification problems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19339–19352.
[39] V. Verma, A. Lamb, C. Beckham, A. Najafi, L. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6438–6447.
[40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep., 2011.
[41] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5309–5318.
[42] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[43] Z. Xu, Z. Chai, and C. Yuan, "Towards calibrated model for long-tailed visual recognition from prior perspective," in *Advances in Neural Information Processing Systems*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 7139–7152.
[44] S. Yan, H.-T. Kao, and E. Ferrara, "Fair class balancing: Enhancing model fairness without observing sensitive attributes," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 1715–1724.
[45] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 25407–25437.
[46] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5718–5727, Apr. 2009.
[47] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.
[48] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ml models with sensitive subspace robustness," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[49] 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. 75, pp. 1–42, 2019.
[50] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.
[51] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[52] M. Zhang, N. S. Sohoni, H. R. Zhang, C. Finn, and C. Re, "Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 26484–26516.
[53] S. Zhang, C. Chen, X. Zhang, and S. Peng, "Label-occurrence-balanced mixup for long-tailed recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3224–3228.
[54] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4352–4360.
[55] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

SANGWOO HONG (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University, South Korea, in 2020, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include distributed computing, fairness in machine learning, and pruning nerual networks.

YOUNGSEOK YOON (Student Member, IEEE) received the B.S. and master's degrees in electrical and computer engineering from Seoul National University, South Korea, in 2021 and 2023, respectively. He focuses on how models could and would learn. His research interests include machine learning and deep learning. Accordingly, he is interested in meta-learning and fairness in machine learning. Also, he is interested in distributed learning to train the model concerning privacy effectively.

HYUNGJUN JOO (Student Member, IEEE) received the B.S. degree from the School of Electrical Engineering, KAIST, South Korea, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University, South Korea. His research interests include deep learning and explainable AI.

JUNGWOO LEE (Senior Member, IEEE) received the B.S. degree in electronics engineering from Seoul National University, Seoul, South Korea, in 1988, and the M.S.E. and Ph.D. degrees in electrical engineering from Princeton University, in 1990 and 1994, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, Seoul National University. He was a member of technical staff, working on multimedia signal processing with SRI (Sarnoff), from 1994 to 1999, where he was a Team Leader (PI) of $18M NIST ATP Program. He has been with the Wireless Advanced Technology Laboratory, Lucent Technologies Bell Labs, since 1999, and worked on W-CDMA base station algorithm development as a Team Leader, for which he received two Bell Labs Technical Achievement Awards. He holds 21 U.S. patents. His research interests include wireless communications, information theory, distributed storage, and machine learning. He is a member of the National Academy of Engineering of Korea. He received the Qualcomm Dr. Irwin Jacobs Award, in 2014, for his contributions in wireless communications. He was a co-recipient of the 2020 IEEE Communications Society Fred W. Ellersick Prize. He has been the General Chair of JCCI'19; the Track Chair of IEEE ICC SPC (2016–2017); and a TPC/OC Member of ICC'15, ITW'15, VTC'15s, ISIT'09, PIMRC'08, ICC'05, and ISITA '05. He was an Editor of IEEE Wireless Communications Letters (WCL), from 2017 to 2021. He was an Associate Editor of IEEE Transactions on Vehicular Technology (2008–2011) and *Journal of Communications and Networks* (JCN) (2012–2016). He has also been a Chief Editor of *KICS* and an Executive Editor of *ICT Express* (Elsevier-KICS), since 2015.