Improving Robustness to Multiple Spurious Correlations by Multi-Objective Optimization

Nayeong Kim [1] **Juwon Kang** [1] **Sungsoo Ahn** [1 2] **Jungseul Ok** [1 2] **Suha Kwak** [1 2]

Abstract

We study the problem of training an unbiased and accurate model given a dataset with multiple biases. This problem is challenging since the multiple biases cause multiple undesirable shortcuts during training, and even worse, mitigating one may exacerbate the other. We propose a novel training method to tackle this challenge. Our method first groups training data so that different groups induce different shortcuts, and then optimizes a linear combination of group-wise losses while adjusting their weights dynamically to alleviate conflicts between the groups in performance; this approach, rooted in the multi-objective optimization theory, encourages to achieve the minimax Pareto solution. We also present a new benchmark with multiple biases, dubbed MultiCelebA, for evaluating debiased training methods under realistic and challenging scenarios. Our method achieved the best on three datasets with multiple biases, and also showed superior performance on conventional single-bias datasets.

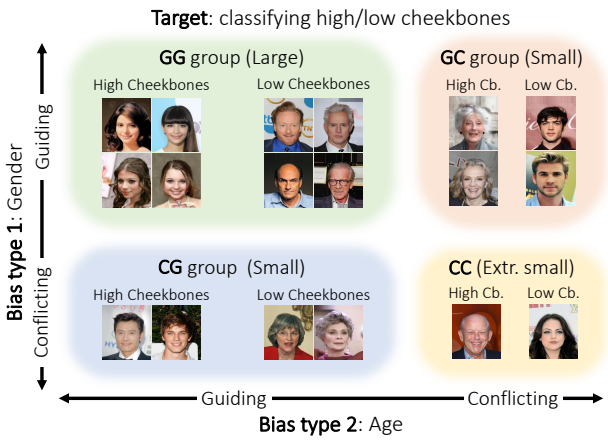


Figure 1: A example of grouping training data with two bias types. Each axis represents each bias type, for which bias-guiding samples make up the majority and bias-conflicting ones hold the minority. In this example, the name of each group indicates if samples of the group has a guiding attribute (G) or a conflicting attribute (C) for `gender` and `age`, in respective order.

1. Introduction

Empirical risk minimization (ERM) (Vapnik, 1999) is currently the gold standard in supervised learning of deep neural networks. However, recent studies (Sagawa et al., 2019; Geirhos et al., 2020) revealed that ERM is prone to taking *undesirable shortcuts* stemming from *spurious correlations* between the target labels and irrelevant attributes arising from subgroup imbalance of training data. For example, Sagawa et al. (2019) showed how much a deep neural network trained to classify bird species relies on the background rather than the bird itself. Such a spurious correlation is often hard to mitigate since the data collection procedure itself is biased towards the correlation.

To resolve this issue, researchers have studied debiased training algorithm, *i.e.*, algorithms training a model while mitigating spurious correlations (Arjovsky et al., 2019; Bahng et al., 2020; Sagawa et al., 2019; Teney et al., 2021; Tartaglione et al., 2021; Lee et al., 2021; Nam et al., 2020; Liu et al., 2021a; Kim et al., 2022). They focus on improving performance on bias-conflicting samples (*i.e.*, samples that disagree with the spurious correlations) to achieve a balance of bias-conflicting and bias-guiding samples (*i.e.*, those agreeing with the spurious correlations). Although these algorithms have shown promising results, they have been evaluated in a limited setting where only a single type of spurious correlation exists in training data.

We advocate that debiased training algorithms should be evaluated under more realistic scenarios with multiple biases. In such scenarios, some samples may align with one bias but may conflict with another, which makes mitigating

[1]Department of Computer Science and Engineering, POSTECH, Pohang, Korea [2]Graduate School of Artificial Intelligence, POSTECH, Pohang, Korea. Correspondence to: Suha Kwak <suha.kwak@postech.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

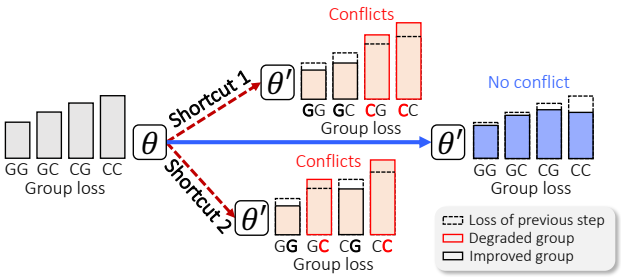


Figure 2: The concept of between-group conflicts and model biases. During model parameter updates (θ), the model risks exploiting spurious correlations as shortcuts for classification (red lines). Updating model parameters toward shortcut results in a reduced group-wise loss in the guiding groups but amplifies the loss in the conflicting groups (*e.g.*, CG and CC for shortcut 1), leading to group conflicts. Updating parameters towards cues directly related to the target classification, free from spurious correlations (blue line), offers the only solution to minimize losses across all groups.

spurious correlations more challenging. If one only considers the intersection of bias-conflicting samples, *i.e.*, *clean* samples that disagree with all the spurious correlations, the resulting group will be extremely small as illustrated in Figure 1 and result in overfitting consequently. Furthermore, mitigation of one bias often promote another as empirically observed by Li et al. (2023). These complexities often lead existing methods to achieve similar or even worse performance compared to simple baselines such as upsampling and upweighting.

We propose a novel debiased training algorithm to tackle the aforementioned challenges. Our algorithm first divides the entire training set into multiple groups where data of the same group have the same impact on training in terms of the model bias, *i.e.*, guiding to or conflicting with each bias type in the same way, as illustrated in Figure 1. With the grouping strategy, all groups share the same target task (*i.e.*, classifying cheekbones), but they have different spurious correlations. If a model exploits a shortcut, its performance on the groups that agree with the associated spurious correlation improves, while that on the groups disagreeing with such a spurious correlation deteriorates, leading to *group conflicts*. We consider a between-group conflict to be an indication that a model is biased towards spurious correlations, which is illustrated in Figure 2.

Then, to mitigate multiple biases during training, our algorithm trains a model by alleviating all between-group conflicts at once so that it performs well across all groups. This optimization process, derived from a multi-objective optimization (MOO) algorithm of Désidéri (2012), aims for Pareto optimality, *i.e.*, a state where no group can be further improved without sacrificing others. To be specific, our algorithm optimizes a linear combination of group-wise losses while dynamically adjusting group-wise importance weights so that model parameters converge to the minimax Pareto solution, which is a Pareto-stationary solution that minimizes the maximum group loss.

We also introduce a new multi-bias benchmark along with the new debiased training algorithm. Our benchmark, dubbed MultiCelebA, is a collection of real facial images from CelebA (Liu et al., 2015), and incorporates multiple bias types that are spuriously correlated with target classes. Compared with existing multi-bias datasets composed of synthetic images (Li et al., 2022; 2023), it allows to evaluate debiased training algorithm on more realistic and challenging scenarios.

We extensively evaluated our method on three multi-bias benchmarks including MultiCelebA and three single-bias benchmarks, where it outperformed every prior arts. The main contribution of this paper is four-fold:

- We present a novel debiased training algorithm based on MOO for mitigating multiple biases simultaneously.
- We present a new real-image multi-bias benchmark for evaluating debiased training methods under realistic and challenging scenarios.
- We benchmarked existing methods for debiased training and demonstrated that they struggle when training data exhibit multiple biases.
- Our method achieved the state of the art on three datasets with multiple biases. In addition, it also showed superior performance on conventional single-bias datasets.

2. Related Work

Debiasing in single bias scenarios. A body of research has addressed the bias issue that arises from spurious correlations between target and latent attributes. A group of previous work exploits manual labels for bias attributes (Arjovsky et al., 2019; Bahng et al., 2020; Dhar et al., 2021; Gong et al., 2020; Li & Vasconcelos, 2019; Sagawa et al., 2019; Teney et al., 2021; Tartaglione et al., 2021; Zhu et al., 2021; Yao et al., 2022; Zhang et al., 2022; Wang et al., 2018; Kirichenko et al., 2023). For instance, Sagawa et al. (2019) presented a robust optimization method that weights groups of different bias attributes differently, Dhar et al. (2021) and Gong et al. (2020) employed adversarial training, and Zhang et al. (2022) proposed using contrastive learning. Later on, debiased training algorithms that do not require any bias supervision have been studied to reduce the annotation cost (Darlow et al., 2020; Creager et al., 2021; Kim et al., 2021; Lee et al., 2021; Nam et al., 2020; Liu et al., 2021a; Ahmed et al., 2021; Kim et al., 2022; Hwang et al., 2022; Zhang & Ré, 2022; Yang et al., 2023; Wu et al., 2023).

However, whether directly using the bias labels or not, these methods assume that the bias inherent in data is of a single type. This assumption often does not hold in real-world scenarios, where data exhibit multiple biases, and in practice classifiers can be easily biased to multiple independent biases, as shown in StylEx (Lang et al., 2021).

Debiasing in multiple biases scenarios. Only a few recent studies (Li et al., 2022; 2023) addressed multiple biases with new training algorithms and benchmarks. Li et al. (2022) discovered multiple biases through iterative assignment of pseudo bias labels, while Li et al. (2023) presented an augmentation method that emulates the generation process of bias types. However, their methods are dedicated to handle synthetic images. In contrast, we propose a new algorithm that trains unbiased models regardless of the number and types of biases, along with a new natural image benchmark for evaluating debiased training methods in the presence of multiple biases.

3. Proposed Method

We propose a novel debiased training framework that incorporates a grouping strategy to unveil model biases and an optimization algorithm based on a theory of MOO (Désidéri, 2012), assuming that bias attributes are annotated for training data. This method effectively addresses one or multiple spurious correlations by training a model for all the groups while optimizing importance weights of the groups as well as the model parameter. The rest of this section first introduces the grouping strategy (Section 3.1) and then describes the proposed optimization algorithm with group-wise importance weight adjustment in detail (Section 3.2)

3.1. Grouping Strategy

As illustrated in Figure 1(a), we divide training data into multiple groups so that all data in the same group have the same impact on training in terms of the model bias. To be specific, we consider training a classifier on a dataset $\mathcal{X} = \{(x^{(m)}, t^{(m)})\}_{m=1}^{M}$, where each sample $x^{(m)}$ is associated with a target class $t^{(m)}$ and a list of attributes $\boldsymbol{b}^{(m)} = [b_1^{(m)}, \dots, b_D^{(m)}]^\top$, where D is the number of bias types. We group the samples using a list of binary group labels $\boldsymbol{g}^{(m)} = [g_1^{(m)}, \dots, g_D^{(m)}]$ based on whether each attribute $b_d^{(m)}$ is the *majority attribute* in target class $t^{(m)}$, *i.e.*, $g_d^{(m)} = 1$ if

$$b_d^{(m)} = \operatorname*{argmax}_{b_d} \left| \left\{ m' | t^{(m')} = t^{(m)}, b_d^{(m')} = b_d \right\} \right|,$$

and $g_d^{(m)} = 0$ otherwise. This results in 2^D groups where samples in the same group share the same group labels.

Our grouping policy differs from prior work that uses the target classes and the attributes as the group labels (Sagawa et al., 2019; Kirichenko et al., 2023; Nam et al., 2022; Sagawa et al., 2020; Zhang et al., 2022): each group in our method contains samples from all the target classes, while the existing ones only keep a group of samples with the same target class and the same attributes. Hence, our grouping policy enables to conduct the target classification task on each group. Moreover, since different groups have different combinations of spurious correlations, a model should not rely on any spurious correlation to work on every group; if it is biased to a spurious correlation, its performance will deteriorate on the groups disagreeing with the spurious correlation. Our goal in the following debiased training step is thus to train a model capable of accurately classifying samples of all the groups, *i.e.*, its performance should not be biased towards a certain group.

3.2. Debiased Training with Group Weight Adjustment

Our debiased training algorithm aims to train a model so that it works for all the groups determined by our grouping policy described in Section 3.1. To this end, our algorithm optimizes a linear combination of group-wise losses while adjusting their importance weights dynamically. In this section, we first briefly review a theory of multi-objective optimization, from which our algorithm stems, and deliver the details of our algorithm.

3.2.1. Preliminary: Multi-objective optimization

We consider MOO as a problem of optimizing a parameter θ with respect to a collection of training objectives $L(\theta) = [\mathcal{L}_1(\theta), \dots, \mathcal{L}_N(\theta)]^\top$. To solve such a problem, MOO frameworks aim at finding a solution that achieves Pareto optimality, *i.e.*, a state where no objective can be improved without sacrificing others.

Definition 3.1 (Pareto optimality)**.** A parameter θ^* is Pareto-optimal if there exists no other parameter θ such that $\mathcal{L}_n(\theta) \leq \mathcal{L}_n(\theta^*)$ for $n = 1, \dots, N$ and $L(\theta) \neq L(\theta^*)$.

However, finding the Pareto-optimal parameter is intractable for non-convex loss functions like training objectives of deep neural networks. Instead, one may consider using gradient-based optimization to find a parameter satisfying Pareto stationarity (Désidéri, 2012), *i.e.*, a state where a convex combination of objective-wise gradients equals a zero-vector. Pareto stationarity is a necessary condition for Pareto optimality if the objectives in $L(\theta)$ are smooth (Désidéri, 2012).

Definition 3.2 (Pareto stationarity)**.** A parameter θ^* is Pareto-stationary if there exists an objective-scaling vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_N]^\top$ satisfying the following condition:

$$\boldsymbol{\alpha}^\top \nabla_\theta L(\theta^*) = \mathbf{0}, \quad \boldsymbol{\alpha} \geq \mathbf{0}, \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1, \tag{1}$$

where $\mathbf{0} = [0, \dots, 0]^\top \in \mathbb{R}^N$ and $\mathbf{1} = [1, \dots, 1]^\top \in \mathbb{R}^N$.

Algorithm 1 Debiased training by MOO

while *not converged* **do**
  for $u \leftarrow 1$ *to* $U-1$ **do**
    Update $\theta \leftarrow \theta - \eta_1 \nabla_\theta L_\theta$.
  end for
  Update parameters:
    $\theta \leftarrow \theta - \eta_1 \nabla_\theta L_\theta$,
    $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta_2 \nabla_{\boldsymbol{\alpha}} L_{\boldsymbol{\alpha}}$,
    $\lambda \leftarrow \lambda + \eta_2 \nabla_\lambda L_{\boldsymbol{\alpha}}$.
end while

Désidéri (2012) proposed the multi-gradient descent algorithm (MGDA) to search for a Pareto-stationary parameter. MGDA finds an objective-scaling parameter $\boldsymbol{\alpha}$ which combines the objective-wise gradients $\nabla_\theta L$ to sum to approximately a zero vector by the following optimization:

$$\min_{\boldsymbol{\alpha}} \left\| \boldsymbol{\alpha}^\top \nabla_\theta L \right\|_2^2, \quad \boldsymbol{\alpha} \geq \mathbf{0}, \quad \boldsymbol{\alpha}^\top \mathbf{1} = 1. \tag{2}$$

Given $\boldsymbol{\alpha}$, MGDA performs a gradient-based update on the parameter θ with respect to $\boldsymbol{\alpha}^\top L(\theta)$.

3.2.2. Proposed Training Algorithm

We propose an algorithm to optimize over $N = 2^D$ groups while minimizing the conflicts between group-wise loss functions, inspired by MGDA. Let $L(\theta) = [\mathcal{L}_1(\theta), \ldots, \mathcal{L}_N(\theta)]^\top$ denote the list of empirical risk functions on N groups and consider minimizing their convex combination $\boldsymbol{\alpha}^\top L(\theta)$, where $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\boldsymbol{\alpha}^\top \mathbf{1} = 1$. This is a unique MOO scenario, in which all objectives are of the same loss function but differ in input distribution (*i.e.*, groups). To impose the constraints on $\boldsymbol{\alpha}$ in Eq. (2), we propose applying the softmax function $\sigma(\cdot)$ to $\boldsymbol{\alpha}$. The model parameter θ is thus optimized by minimizing the weighted group-wise losses:

$$L_\theta = \sigma(\boldsymbol{\alpha})^\top L(\theta). \tag{3}$$

To address between-group conflicts, we propose adjusting the group-scaling parameter $\sigma(\boldsymbol{\alpha})$ so that the training converges to a Pareto-stationary solution. To be specific, our goal is to minimize the training objective $\sigma(\boldsymbol{\alpha})^\top L(\theta)$ while simultaneously adjusting the group-scaling parameter to minimize the objective in Eq. (2). To this end, we optimize the following loss function with respect to $\boldsymbol{\alpha}$:

$$L_{\boldsymbol{\alpha}} = \sigma(\boldsymbol{\alpha})^\top L(\theta) + \lambda \left\| \sigma(\boldsymbol{\alpha})^\top (\nabla_\theta L(\theta))_\dagger \right\|_2^2, \tag{4}$$

where $(\cdot)_\dagger$ denotes the stop-gradient operator and λ is the Lagrangian multiplier for the Pareto stationarity objective in Eq. (2). We update the group-scaling parameter with gradient descent and the Lagrangian multiplier λ with gradient ascent every U iteration. The learning process of our method is described in Algorithm 1. Our algorithm encourages a model to achieve the minimax Pareto solution among Pareto-stationary solutions, which minimizes the maximum group loss by emphasizing groups with lower accuracy, *i.e.*, increasing their scaling parameters. This approach enables debiased training across groups. Further details are presented in Section 3.3.1.

We also note that our method can be interpreted as curvature aware training (Li & Gong, 2021), where the group-scaling parameter $\boldsymbol{\alpha}$ is adjusted for better generalization on each group. Specifically, Li & Gong (2021) consider adjusting the training objective $\left\| \boldsymbol{\alpha}^\top (\nabla_\theta L(\theta)) \right\|_2^2$ so that gradient-based optimization of the parameter θ converges to a *flat minimum* with a small curvature, *i.e.*, a parameter with a small trace of the Hessian matrix with respect to the training objective. It has been reported in the literature that a model converging to such a flat minimum in training has better generalization capability (Keskar et al., 2017; Dziugaite & Roy, 2017; Jiang et al., 2020). Since the number of samples that disagree with all the spurious correlations is extremely small and prone to overfitting, improving generalization capability is particularly beneficial in multiple biases scenarios.

3.3. Discussion

3.3.1. Why our algorithm achieves the minimax Pareto solution

During training, our algorithm more emphasizes groups with worse accuracy by increasing their scaling parameters, which encourages achieving the minimax Pareto solution. This behavior of our algorithm is attributed to the following two factors of $L_{\boldsymbol{\alpha}}$ in Eq. (4).

Regarding the first term of $L_{\boldsymbol{\alpha}}$: Minimizing this term substantially increases the group-scaling parameter of the CC group that exhibits the worst performance in testing, consequently improving the worst accuracy. To be specific, as $\sigma(\boldsymbol{\alpha})$ holds the sum-to-one constraint, minimizing the first term increases the scaling parameter for groups with lower loss magnitudes while decreasing the parameter for groups with higher loss magnitudes. Here, the CC group shows the lowest training loss scale since its cardinality is extremely small and the model is easily overfitted to the group (which causes the worst accuracy on the group in testing). Further empirical analysis on this term is provided in Appendix A.3.

Regarding the second term of $L_{\boldsymbol{\alpha}}$: This term, originated from MGDA, also increases the scaling parameter for the CC group. MGDA is known to be biased towards tasks with low loss magnitudes in multi-task learning, a phenomenon known as *task impartiality* (Javaloy & Valera, 2022; Liu et al., 2021b). In our setting, the task with the lowest loss

magnitude corresponds to the CC group as discussed above. Hence, the task impartiality of MGDA leads to the increased scaling parameter for the CC group, leading to the minimax Pareto solution.

3.3.2. Comparison with Multi-task Learning

Multi-task learning (MTL) is a research area that aims to develop a model capable of performing multiple tasks simultaneously. Considering that MTL incorporates MOO to address task conflicts (Sener & Koltun, 2018), MTL exhibits similarities with our method. Nevertheless, our work is clearly distinct from MTL in multiple aspects. Firstly, our work addresses a single classification task and thus the group-wise losses have the same form. However, their input distributions differ, with each group-wise loss calculated using samples from its respective group. In contrast, MTL assumes different loss functions for different tasks. Secondly, since all the groups aim to solve the same target task, an optimal solution that fits perfectly across all the groups exists for our debiased training setting in principle. On the contrary, MTL rarely has the perfect solution for all the tasks since task conflicts are almost inevitable. Third, our method does not employ task-specific network parameters unlike MTL, which in general distinguishes task-specific and task-shared parameters. Lastly, we present a novel loss tailored to the debiased training.

3.3.3. On the Use of Bias Labels

Bias attribute labels would be expensive, particularly in the multi-bias setting. However, regarding that debiasing in this setting has been rarely studied so far and is extremely challenging, we believe it is premature to tackle the task in an unsupervised fashion at this time. As in the single-bias setting where the society has first developed supervised debiasing methods and then unsupervised counterparts, our algorithm will be a cornerstone of follow-up unsupervised methods in the multi-bias setting. Moreover, the annotation cost for bias labels can be substantially reduced by incorporating existing techniques for pseudo labeling of bias attributes (Jung et al., 2022; Nam et al., 2022).

4. MultiCelebA Benchmark

We present a new benchmark, dubbed MultiCelebA, for evaluating debiased training algorithms under the presence of multiple biases. Unlike Multi-Color MNIST (Li et al., 2022) and UrbanCars (Li et al., 2023) built for the same purpose using synthetic images, MultiCelebA is composed of natural facial images, making it more suitable for simulating real-world scenarios.

MultiCelebA is built upon CelebA (Liu et al., 2015), a large-scale collection of facial images each with 40 attribute

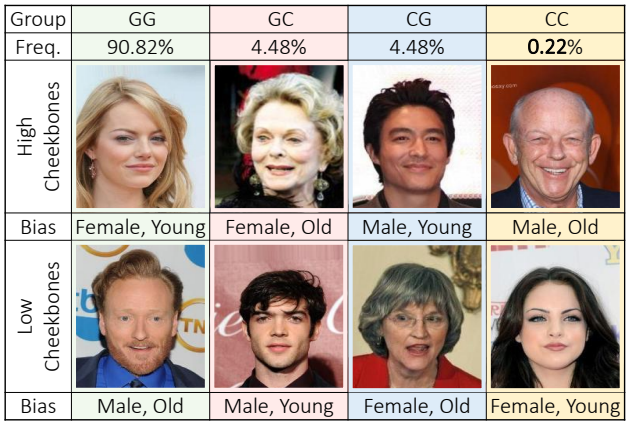


Figure 3: Training set configuration of MultiCelebA in the two-bias setting.

annotations. Among these attributes, `high-cheekbones` is chosen as the target class, while `gender`, `age`, and `mouth slightly open` are used as bias attributes that are spuriously correlated with `high-cheekbones` and thus cause undesirable shortcuts during training. Note that these bias attributes are not randomly chosen but identified by adapting the empirical analysis procedure of Scimeca et al. (2022) to CelebA, which revealed that these attributes are strongly correlated with the target class; details of the analysis are presented in Appendix A.2.

Based on MultiCelebA, we present two different benchmark settings: one with two bias attributes `gender` and `age`, and the other with all three bias attributes. To simulate challenging scenarios where training data are extremely biased, we set the bias-guiding samples for both `gender` and `age` to 95.3% by subsampling from the CelebA training set, so that only 0.22% of training samples are free from spurious correlations in the two-bias setting and 0.07% for the three-bias settings. Example images and the frequency of each attribute in the two-bias setting are presented in Figure 3.

5. Experiments

5.1. Setup

Datasets. We adopt three multi-bias benchmarks, MultiCelebA, UrbanCars (Li et al., 2023), and Multi-Color MNIST (Li et al., 2022), and three single-bias datasets, Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2015), and BFFHQ (Lee et al., 2021) for evaluation.

Evaluation metrics. For multi-bias benchmarks, the quality of debiased training algorithms is measured mainly by Unbiased, the average of group average accuracy scores. For the benchmarks with two bias types, we also adopt average accuracy for each of the four groups categorized by the

Table 1: Performance in INDIST, GG, GC, CG, CC, and UNBIASED (%) on MultiCelebA in two-bias setting. The first element of each of the four combinations {GG, GC, CG, CC} is about the bias type `gender`, while the second is about the bias type `age`. We mark the best and the second-best performance in **bold** and underline, respectively.

Method	Bias label	INDIST	GG	GC	CG	CC	UNBIASED
ERM	✗	**97.0**$_{\pm 0.2}$	**98.2**$_{\pm 0.7}$	**89.2**$_{\pm 2.6}$	58.2$_{\pm 3.0}$	19.0$_{\pm 1.8}$	63.8$_{\pm 1.2}$
LfF	✗	81.9$_{\pm 3.1}$	79.8$_{\pm 2.6}$	71.7$_{\pm 2.2}$	80.2$_{\pm 1.7}$	71.5$_{\pm 3.3}$	75.8$_{\pm 0.5}$
JTT	✗	78.7$_{\pm 6.5}$	76.1$_{\pm 5.2}$	60.8$_{\pm 5.2}$	65.1$_{\pm 10.7}$	51.9$_{\pm 1.6}$	64.7$_{\pm 3.2}$
DebiAN	✗	66.8$_{\pm 34.1}$	64.4$_{\pm 30.4}$	63.6$_{\pm 22.2}$	49.8$_{\pm 7.6}$	45.5$_{\pm 13.2}$	55.8$_{\pm 11.7}$
Upsampling	✓	82.6$_{\pm 0.8}$	79.8$_{\pm 1.5}$	81.0$_{\pm 1.3}$	76.7$_{\pm 1.1}$	75.6$_{\pm 1.2}$	78.3$_{\pm 0.8}$
Upweighting	✓	83.4$_{\pm 5.9}$	79.0$_{\pm 4.1}$	79.2$_{\pm 6.0}$	80.8$_{\pm 0.0}$	78.7$_{\pm 3.6}$	79.4$_{\pm 3.4}$
GroupDRO	✓	83.5$_{\pm 0.7}$	81.2$_{\pm 1.0}$	81.2$_{\pm 1.2}$	76.7$_{\pm 1.5}$	74.6$_{\pm 0.4}$	78.4$_{\pm 0.7}$
SUBG	✓	80.3$_{\pm 1.1}$	77.1$_{\pm 1.0}$	78.4$_{\pm 0.7}$	77.5$_{\pm 1.7}$	78.0$_{\pm 1.2}$	77.7$_{\pm 0.6}$
LISA	✓	84.5$_{\pm 1.7}$	82.8$_{\pm 1.3}$	83.2$_{\pm 0.5}$	79.8$_{\pm 0.8}$	77.6$_{\pm 2.6}$	80.9$_{\pm 0.2}$
DFR$^{tr}_{tr}$	✓	85.5$_{\pm 6.2}$	91.3$_{\pm 3.5}$	83.6$_{\pm 4.0}$	46.7$_{\pm 3.8}$	28.5$_{\pm 4.6}$	62.5$_{\pm 0.6}$
Ours	✓	84.3$_{\pm 0.9}$	82.4$_{\pm 0.9}$	85.1$_{\pm 0.4}$	**81.7**$_{\pm 0.4}$	**82.6**$_{\pm 1.0}$	**82.9**$_{\pm 0.2}$

Table 2: Performance in INDIST, CCC, and UNBIASED (%) on MultiCelebA in three biases for evaluation two biases for training setting. We mark the best and the second-best performance in **bold** and underline, respectively.

Method	INDIST	CCC	UNBIASED
ERM	**96.7**$_{\pm 0.2}$	11.1$_{\pm 2.8}$	63.8$_{\pm 1.2}$
LfF	81.8$_{\pm 3.1}$	60.1$_{\pm 1.7}$	71.8$_{\pm 0.7}$
Upsampling	82.6$_{\pm 0.8}$	63.0$_{\pm 3.5}$	73.6$_{\pm 0.8}$
Upweighting	85.4$_{\pm 9.5}$	63.4$_{\pm 3.5}$	75.8$_{\pm 4.8}$
GroupDRO	83.4$_{\pm 0.6}$	61.4$_{\pm 2.9}$	73.7$_{\pm 0.7}$
SUBG	80.3$_{\pm 1.1}$	65.8$_{\pm 4.0}$	72.6$_{\pm 1.3}$
LISA	84.5$_{\pm 1.7}$	63.1$_{\pm 1.0}$	75.9$_{\pm 0.6}$
DFR$^{tr}_{tr}$	85.5$_{\pm 6.1}$	26.2$_{\pm 6.2}$	61.3$_{\pm 0.6}$
Ours	84.3$_{\pm 1.0}$	**70.0**$_{\pm 0.6}$	**78.4**$_{\pm 0.0}$

Table 3: Performance in INDIST and CC (%) on UrbanCars. We mark the best and the second-best in **bold** and underline, respectively.

Method	Bias label	INDIST	CC	GAP
ERM	✗	**97.6**	28.4	-69.2
Upsampling	✓	92.8	76.0	-16.8
Upweighting	✓	93.4	80.0	-13.4
GroupDRO	✓	91.6	75.2	-16.4
SUBG	✓	71.1	64.8	-6.3
LISA	✓	94.6	80.8	-13.8
DFR$^{tr}_{tr}$	✓	89.7	44.5	-45.2
Ours	✓	91.8	**87.6**	**-4.2**

guiding or conflicting nature of the biases: {GG, GC, CG, CC}, where G and C indicate whether a group includes bias-guiding or bias-conflicting samples for each bias type, respectively. Conceptually, the GG metric can be high regardless of whether a model is biased or not. However, the CC metric can be high only when a model is debiased from all spurious correlations. Similarly, the GC accuracy can be high only when a model is debiased from the spurious correlation of the second bias type. We also report INDIST, the weighted average of group accuracy scores where the weights are proportional to group sizes of training data (Sagawa et al., 2019). For Waterbirds and CelebA, we adopt WORST, the minimum of group accuracy scores, following Sagawa et al. (2019).

Baselines. We compare our algorithm with a large body of existing debiased training algorithms. Among them, GroupDRO (Sagawa et al., 2019), EnD (Tartaglione et al., 2021), SUBG (Sagawa et al., 2020), LISA (Yao et al., 2022), and DFR (Kirichenko et al., 2023) as well as simple upsampling and upweighting strategies demand true bias labels of training data like ours, while HEX (Wang et al., 2019), ReBias (Bahng et al., 2020), LfF (Nam et al., 2020), JTT (Liu et al., 2021a), EIIL (Creager et al., 2021), PGI (Ahmed et al., 2021), DisEnt (Lee et al., 2021), LWBC (Kim et al., 2022), SelecMix (Hwang et al., 2022), CNC (Zhang et al., 2022), and DebiAN (Li et al., 2022) do not.

Implementation details. Following previous work, we conduct experiments using different neural network architectures for different datasets: a three-layered MLP for Multi-Color MNIST and ResNet18 for MultiCelebA and BFFHQ, ResNet50 for UrbanCars, Waterbirds, and CelebA. The group-scaling parameter $\boldsymbol{\alpha}$ is initialized to $\frac{1}{N}\mathbf{1}$ where N is the number of groups, and the Lagrangian multiplier λ is initialized to 0. For mini-batch construction during training, group-balanced sampling is used to compute each loss for multiple groups. For MultiCelebA, we tuned hyperparameters in the two-bias setting and performed both training and evaluation in the two-bias setting, and conducted evaluation only in the three-bias setting without training. We report the average and standard deviation of each metric calculated from three runs with different random seeds. More implementation details are provided in Appendix A.4.

5.2. Quantitative Results

MultiCelebA in two-bias setting. In Table 1, we present the results of our experiments evaluating the performance of various baselines and existing debiased training methods

Table 4: Performance in GG, GC, CG, CC, and UNBIASED (%) on Multi-Color MNIST. The first element of each of the four combinations {GG, GC, CG, CC} is about the bias type `left-color`, while the second is about the bias type `right-color`. We mark the best and the second-best performance in **bold** and underline, respectively.

Method	Bias label	GG	GC	CG	CC	UNBIASED
ERM	✗	$100.0_{\pm 0.0}$	$96.5_{\pm 1.2}$	$79.5_{\pm 2.5}$	$20.8_{\pm 1.1}$	$74.2_{\pm 1.1}$
LfF	✗	$99.6_{\pm 0.5}$	$4.7_{\pm 0.5}$	$\mathbf{98.6}_{\pm 0.4}$	$5.1_{\pm 0.4}$	$52.0_{\pm 0.1}$
EIIL	✗	$100.0_{\pm 0.0}$	$\mathbf{97.2}_{\pm 1.5}$	$70.8_{\pm 4.9}$	$10.9_{\pm 0.8}$	$69.7_{\pm 1.0}$
PGI	✗	$98.6_{\pm 2.3}$	$82.6_{\pm 19.6}$	$26.6_{\pm 5.5}$	$9.5_{\pm 3.2}$	$54.3_{\pm 4.0}$
DebiAN	✗	$100.0_{\pm 0.0}$	$95.6_{\pm 0.8}$	$76.5_{\pm 0.7}$	$16.0_{\pm 1.8}$	$72.0_{\pm 0.8}$
Upsampling	✓	$99.4_{\pm 0.6}$	$89.8_{\pm 1.4}$	$81.3_{\pm 2.6}$	$42.0_{\pm 1.7}$	$78.1_{\pm 1.4}$
Upweighting	✓	$100.0_{\pm 0.0}$	$90.0_{\pm 2.5}$	$83.4_{\pm 2.1}$	$37.1_{\pm 2.8}$	$77.6_{\pm 1.0}$
GroupDRO	✓	$98.0_{\pm 0.0}$	$87.2_{\pm 4.3}$	$77.3_{\pm 7.5}$	$\mathbf{52.3}_{\pm 2.6}$	$78.7_{\pm 2.7}$
Ours	✓	$99.7_{\pm 0.6}$	$90.4_{\pm 3.4}$	$81.8_{\pm 4.0}$	$48.1_{\pm 0.3}$	$\mathbf{80.0}_{\pm 2.0}$

Table 5: WORST and INDIST metrics (%) evaluated on Waterbirds, and CelebA. We mark the best and the second-best performance of WORST in **bold** and underline, respectively.

Method	Bias label	Waterbirds WORST	Waterbirds INDIST	CelebA WORST	CelebA INDIST
ERM	✗	$63.7_{\pm 1.9}$	$97.0_{\pm 0.2}$	$47.8_{\pm 3.7}$	$94.9_{\pm 0.2}$
LfF	✗	78.0	91.2	70.6	86.0
EIIL	✗	$77.2_{\pm 1.0}$	$96.5_{\pm 0.2}$	$81.7_{\pm 0.8}$	$85.7_{\pm 0.1}$
JTT	✗	$83.8_{\pm 1.2}$	$89.3_{\pm 0.7}$	$81.5_{\pm 1.7}$	$88.1_{\pm 0.3}$
LWBC	✗	-	-	$85.5_{\pm 1.4}$	$88.9_{\pm 1.6}$
CNC	✗	$88.5_{\pm 0.3}$	$90.9_{\pm 0.1}$	$88.8_{\pm 0.9}$	$89.9_{\pm 0.5}$
Upweighting	✓	$88.0_{\pm 1.3}$	$95.1_{\pm 0.3}$	$83.3_{\pm 2.8}$	$92.9_{\pm 0.2}$
GroupDRO	✓	$89.9_{\pm 0.6}$	$92.0_{\pm 0.6}$	$88.9_{\pm 1.3}$	$93.9_{\pm 0.1}$
SUBG	✓	$89.1_{\pm 1.1}$	-	$85.6_{\pm 2.3}$	-
SSA	✓	$89.0_{\pm 0.6}$	$92.2_{\pm 0.9}$	$\mathbf{89.8}_{\pm 1.3}$	$92.8_{\pm 0.1}$
LISA	✓	$89.2_{\pm 0.6}$	$91.8_{\pm 0.3}$	$89.3_{\pm 1.1}$	$92.4_{\pm 0.4}$
DFR^{tr}_{tr}	✓	$90.2_{\pm 0.8}$	$97.0_{\pm 0.3}$	$80.7_{\pm 2.4}$	$90.6_{\pm 0.7}$
Ours	✓	$\mathbf{91.8}_{\pm 0.3}$	$95.6_{\pm 0.3}$	$\mathbf{89.8}_{\pm 1.3}$	$91.4_{\pm 1.2}$

Table 6: UNBIASED metric (%) evaluated on BFFHQ. We mark the best and the second-best performance in **bold** and underline, respectively.

Method	Bias label	BFFHQ UNBIASED
ERM	✗	$56.2_{\pm 0.4}$
HEX	✗	$52.8_{\pm 0.9}$
ReBias	✗	$56.8_{\pm 1.6}$
LfF	✗	$65.6_{\pm 1.4}$
DisEnt	✗	$61.6_{\pm 2.0}$
SelecMix	✗	$71.6_{\pm 1.9}$
SelecMix	✓	$75.0_{\pm 0.5}$
EnD	✓	$56.5_{\pm 0.6}$
LISA	✓	$65.2_{\pm 0.5}$
GroupDRO	✓	$85.1_{\pm 0.9}$
Ours	✓	$\mathbf{85.7}_{\pm 0.3}$

on MultiCelebA. One can observe how our method outperforms the baselines by a significant margin in UNBIASED, CG, and CC metrics. Our method even achieves a second-best accuracy in the GC metric and a moderate accuracy in the GG metric. This highlights how our method successfully prevents performance degradation by simultaneously removing multiple spurious correlations. Intriguingly, we observe that algorithms like JTT, DebiAN, and DFR exhibit UNBIASED metric similar to or even lower than the vanilla ERM algorithm. Our hypothesis is that this performance degradation stems from conflicts between the removal of different spurious correlations. To be specific, JTT (Liu et al., 2021a) exhibits varying accuracy across the GG, GC, CG, and CC groups, indicating that the model is biased towards both `gender` and `age` biases. DebiAN (Li et al., 2022) shows high accuracy in the GG and GC groups, but low accuracy in the CG and CC groups, indicating that the algorithm partially mitigates `age` bias but still suffers from `gender` bias. We also observe that DFR (Kirichenko et al., 2023) achieves lower CC and CG metrics than ERM, suggesting that an ERM-based feature representation alone is insufficient in multi-bias setting. The remaining algorithms, *e.g.*, Upsampling, GroupDRO (Sagawa et al., 2019), and LISA (Yao et al., 2022) show overall decent performance, but the GG and GC metrics are slightly higher than that in CG and CC groups, indicating that the model is still biased towards the `gender` attribute. Surprisingly, the upweighting baseline achieved the second-best performance in CG and CC metrics on MultiCelebA, surpassing all the existing debiased training methods.

MultiCelebA in three-bias setting. Results of the evaluation with three bias types are reported in Table 2, where only `gender` and `age` labels are visible during training. ERM exhibits lower CCC accuracy in the three-bias setting compared to the two-bias setting. This arises as the number of bias types increases, resulting in a substantially reduced size of the smallest group, demonstrating a more challenging setting. In contrast, our method substantially outperformed existing methods and baselines in UNBIASED and CCC. This demonstrates the scalability of our method to more than two bias types.

UrbanCars. In Table 3, we present the results of debiased training algorithms that exploit bias labels and share the

Table 7: Comparisons among different strategies for adjusting the group-scaling parameter $\boldsymbol{\alpha}$ on MultiCelebA in two biases setting. (a) Fixing $\sigma(\boldsymbol{\alpha})$ by $\frac{1}{N}\mathbf{1}$. (b) Minimizing $\sigma(\boldsymbol{\alpha})^\top L(\theta)$. (c) MGDA. (d) GradNorm. (e) MoCo, the latest MOO method. (f) Ours minimizing $L_{\boldsymbol{\alpha}}$.

	INDIST	GG	GC	CG	CC	UNBIASED
(a) No optimization	$78.5_{\pm 5.7}$	$79.6_{\pm 2.9}$	$80.0_{\pm 2.2}$	$79.0_{\pm 1.9}$	$78.4_{\pm 1.3}$	$79.2_{\pm 1.4}$
(b) Minimizing group losses	$81.3_{\pm 3.6}$	$76.4_{\pm 2.2}$	$77.8_{\pm 0.4}$	$77.1_{\pm 2.2}$	$78.0_{\pm 1.7}$	$77.3_{\pm 0.6}$
(c) MGDA	$82.7_{\pm 3.7}$	$81.6_{\pm 3.5}$	$\underline{85.1}_{\pm 2.1}$	$\underline{80.1}_{\pm 1.3}$	$\underline{82.3}_{\pm 3.2}$	$\underline{82.3}_{\pm 0.4}$
(d) GradNorm	$\mathbf{86.5}_{\pm 5.4}$	$\mathbf{85.9}_{\pm 5.8}$	$\mathbf{86.9}_{\pm 2.4}$	$78.1_{\pm 3.5}$	$76.6_{\pm 6.5}$	$81.9_{\pm 0.6}$
(e) MoCo	$83.8_{\pm 1.6}$	$81.7_{\pm 1.3}$	$81.8_{\pm 2.6}$	$77.2_{\pm 0.9}$	$74.9_{\pm 1.2}$	$78.9_{\pm 1.5}$
(f) Ours	$\underline{84.3}_{\pm 0.9}$	$\underline{82.4}_{\pm 0.9}$	$\underline{85.1}_{\pm 0.4}$	$\mathbf{81.7}_{\pm 0.4}$	$\mathbf{82.6}_{\pm 1.0}$	$\mathbf{82.9}_{\pm 0.2}$

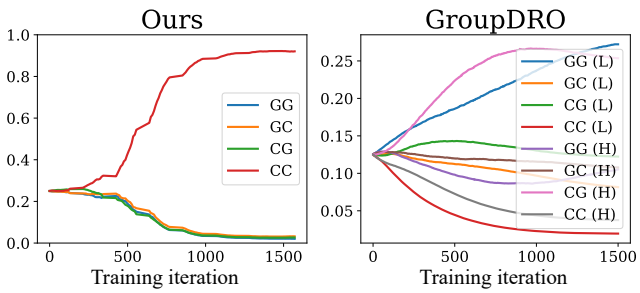
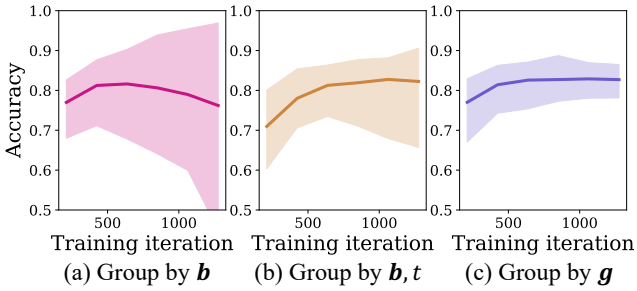


Figure 4: Change of the group-scaling parameter over time on MultiCelebA in two-bias settings. In the case of GroupDRO, (H) and (L) denote `High-cheekbones` and `Low-cheekbones`, respectively.



(a) Group by $\boldsymbol{b}$ (b) Group by $\boldsymbol{b}, t$ (c) Group by $\boldsymbol{g}$

Figure 5: Group-wise test accuracy of three different grouping strategies. Lines indicate UNBIASED performance, and shaded regions show the lowest and the highest accuracy among the group-wise scores. To ensure fair comparison, the test data are grouped by $\mathbf{b}$ and t.

identical network structure. Our method achieved significantly superior CC accuracy when compared to methods using bias labels, demonstrating a substantial difference.

Multi-Color MNIST. In Table 4, we report the evaluation results for the Multi-Color MNIST dataset. Note that we reuse the performance of LfF (Nam et al., 2020), EIIL (Creager et al., 2021), PGI (Ahmed et al., 2021), and DebiAN (Li et al., 2022) reported by Li et al. (2022). Overall, our method demonstrates the best performance along with GroupDRO. In particular, our algorithm exhibits the highest UNBIASED accuracy and the second-best CC accuracy.

Single-bias datasets. In Table 5 and 6, our method achieves the best WORST accuracy on Waterbirds and CelebA, and the best UNBIASED on BFFHQ, indicating that our method is effective not only for multi-bias settings but also for single-bias settings.

5.3. In-depth Analysis

Comparisons among different strategies for adjusting the group-scaling parameter. We first verify the impact of our strategy for adjusting the group-scaling parameter. In Table 7, we compare our training strategy with five alternatives: (a) Using a fixed uniform group-scaling parameter $\sigma(\boldsymbol{\alpha}) = \frac{1}{N}\mathbf{1}$ (*i.e.*, no optimization), (b) minimizing group losses $\sigma(\boldsymbol{\alpha})^\top L(\theta)$, (c) MGDA that minimizes $\left\|\sigma(\boldsymbol{\alpha})^\top(\nabla_\theta L(\theta))_\dagger\right\|_2^2$, (d) GradNorm (Chen et al., 2018), (e) MoCo, the latest technique for MOO method (Fernando et al., 2023), and (f) our method that minimizes $L_{\boldsymbol{\alpha}}$ in Eq. (4). Intriguingly, (b) leads to worse performance compared to (a) that uses a fixed value for $\boldsymbol{\alpha}$. We found that utilizing a learnable group-scaling parameter based solely on the weighted sum of group-wise losses resulted in worse performance in all metrics except INDIST when compared with training without it. The results in (c), (d), and (e) demonstrate that blindly applying an existing MOO method as-is with our grouping strategy falls short of the desired level of unbiased performance during training on a biased dataset. This highlights the superiority of our method in scenarios involving multiple spurious correlations.

Change of the group-scaling parameter over time. We compare the trend of the group-scaling parameter in our method with that of GroupDRO (Sagawa et al., 2019) on MultiCelebA in the two-bias setting, as illustrated in Figure 4. Our method shows an increasing trend for the weight of the CC group, while those of the other groups decrease during training. This indicates that the model initially learns a shared representation that incorporates information from all the groups, but later focuses more on the minority group. On the other hand, GroupDRO exhibits a decreasing weight

Table 8: Ablation study on the grouping strategy on MultiCelebA in two biases setting: Grouping by bias attribute $\boldsymbol{b}$, grouping by both bias attribute and target class $(\boldsymbol{b}, t)$, and our strategy using the list of binary group labels $\boldsymbol{g}$. SUBG and GroupDRO with our grouping strategy are indicated by †.

Method	Group by	INDIST	GG	GC	CG	CC	UNBIASED
ERM	-	**97.0**$_{\pm 0.2}$	**98.2**$_{\pm 0.7}$	**89.2**$_{\pm 2.6}$	58.2$_{\pm 3.0}$	19.0$_{\pm 1.8}$	63.8$_{\pm 1.2}$
SUBG	$\boldsymbol{b}, t$	80.3$_{\pm 1.1}$	77.1$_{\pm 1.0}$	78.4$_{\pm 0.7}$	77.5$_{\pm 1.7}$	78.0$_{\pm 1.2}$	77.7$_{\pm 0.6}$
SUBG†	$\boldsymbol{g}$	80.4$_{\pm 2.7}$	78.5$_{\pm 4.3}$	75.9$_{\pm 3.3}$	71.9$_{\pm 2.2}$	67.0$_{\pm 2.3}$	73.3$_{\pm 1.7}$
GroupDRO	$\boldsymbol{b}, t$	83.5$_{\pm 0.7}$	81.2$_{\pm 1.0}$	81.2$_{\pm 1.2}$	76.7$_{\pm 1.5}$	74.6$_{\pm 0.4}$	78.4$_{\pm 0.7}$
GroupDRO†	$\boldsymbol{g}$	85.8$_{\pm 1.5}$	83.1$_{\pm 1.9}$	79.5$_{\pm 2.4}$	80.7$_{\pm 1.2}$	71.8$_{\pm 1.1}$	78.8$_{\pm 1.4}$
Ours	$\boldsymbol{b}$	79.2$_{\pm 0.7}$	79.5$_{\pm 4.6}$	79.8$_{\pm 3.5}$	78.1$_{\pm 2.1}$	77.0$_{\pm 1.6}$	78.6$_{\pm 2.0}$
Ours	$\boldsymbol{b}, t$	78.5$_{\pm 5.5}$	79.4$_{\pm 2.9}$	80.0$_{\pm 2.2}$	79.0$_{\pm 1.9}$	78.5$_{\pm 1.3}$	79.2$_{\pm 1.4}$
Ours	$\boldsymbol{g}$	84.3$_{\pm 0.9}$	82.4$_{\pm 0.9}$	85.1$_{\pm 0.4}$	**81.7**$_{\pm 0.4}$	**82.6**$_{\pm 1.0}$	**82.9**$_{\pm 0.2}$

Table 9: Impact of the update period U of the group-scaling parameter on MultiCelebA in two biases setting.

U	GG	GC	CG	CC	UNBIASED
1	84.2$_{\pm 0.5}$	86.0$_{\pm 0.5}$	80.8$_{\pm 0.5}$	80.8$_{\pm 0.5}$	82.9$_{\pm 0.3}$
5	83.3$_{\pm 0.4}$	85.8$_{\pm 0.7}$	81.2$_{\pm 0.4}$	81.7$_{\pm 0.1}$	83.0$_{\pm 0.1}$
10	82.4$_{\pm 0.6}$	85.1$_{\pm 0.4}$	81.7$_{\pm 0.3}$	82.6$_{\pm 0.9}$	82.9$_{\pm 0.2}$
20	81.9$_{\pm 0.5}$	84.9$_{\pm 0.5}$	81.8$_{\pm 0.5}$	83.0$_{\pm 0.9}$	82.9$_{\pm 0.3}$
30	79.3$_{\pm 1.3}$	84.0$_{\pm 0.2}$	82.6$_{\pm 0.3}$	85.0$_{\pm 0.8}$	82.7$_{\pm 0.2}$

trend for the minority groups (CC (L) and CC (H) in Figure 4). This trend occurs because the minority groups have lower training losses in the early stages of training, leading to lower weights in GroupDRO. As a consequence, it tends to ignore minority groups and exacerbate the bias issue, resulting in inferior performance compared to the upweigthing baseline as shown in Table 1.

Ablation study on the grouping strategy. To verify the contribution of our grouping strategy, we compare ours with two alternatives: grouping samples by the same bias attribute $\boldsymbol{b}$, and grouping those with the same pair of bias attribute $\boldsymbol{b}$ and target class t. Figure 5 demonstrates performance variations by different grouping policies. To ensure a fair comparison, the test data are grouped by **b** and t, which is the same as the conventional grouping strategy. Figure 5(a) shows that the test accuracy gap between groups enlarges as training progresses when using the bias attribute grouping. We conjecture that this problem arises from class imbalance within the groups categorized solely by bias attributes. Specifically, the number of samples belonging to a target class that is spuriously correlated with the bias attribute becomes dominant, leading to an imbalanced representation of target classes within the group. In Figure 5(b), we applied the commonly used strategy: grouping by both target classes and bias attributes. Compared with the conventional grouping, our method demonstrates a smaller performance gap between groups and higher the lowest group accuracy, as shown in Figure 5(c). Finally, we also report the performance metrics in Table 8, which demonstrates that our grouping strategy outperforms the others in all metrics.

Applying our grouping strategy to GroupDRO and SUBG. We also compare our method with GroupDRO and SUBG using the proposed grouping strategy. Results in Table 8 suggest that applying our grouping strategy alone to existing debiased training methods failed to achieve performance comparable to ours. This highlights the contribution of both our debiased training algorithm and grouping strategy to performance improvement.

Impact of the update period U. We conducted experiments to examine how hyperparameter U affects the performance of our method. Table 9 reports the performance in GG, GC, CG, CC and UNBIASED metrics on MultiCelebA using five different values of U. To disregard the influence of the learning rate η_2, we adjusted the learning rate η_2 inversely proportional to the increase in the value of U. We found that the UNBIASED remained consistent across all U values we examined, which suggests that our algorithm is not sensitive to U.

6. Conclusion

We have presented a novel debiased training algorithm that addresses the challenges posed by multiple biases in training data, inspired by multi-objective optimization (MOO). In addition, we have introduced a new real-image multi-bias benchmark, dubbed MultCelebA. Our method surpassed existing algorithms for debaised training in both multi-bias and single-bias settings on six benchmarks in total.

Acknowledgement

We express our gratitude to the reviewers for their thoughtful comments, which help improve our paper. This research was supported by the NRF grant and IITP grants funded by Ministry of Science and ICT (2021R1A2C3012728, RS-2022-II220290, RS-2019-II191906, IITP-2021-0-00739) and the NRF grant funded by the Ministry of Education, Korea (2022R1A6A1A03052954).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmed, F., Bengio, Y., van Seijen, H., and Courville, A. Systematic generalisation with group invariant predictions. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 794–803. PMLR, 2018.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 2189–2200. PMLR, 2021.

Darlow, L., Jastrzebski, S., and Storkey, A. Latent adversarial debiasing: Mitigating collider bias in deep neural networks. *arXiv preprint arXiv:2011.11486*, 2020.

Désidéri, J.-A. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Dhar, P., Gleason, J., Roy, A., Castillo, C. D., and Chellappa, R. Pass: Protected attribute suppression system for mitigating bias in face recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 15087–15096, 2021.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Fernando, H. D., Shen, H., Liu, M., Chaudhury, S., Murugesan, K., and Chen, T. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gong, S., Liu, X., and Jain, A. K. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 330–347. Springer, 2020.

Hwang, I., Lee, S., Kwak, Y., Oh, S. J., Teney, D., Kim, J.-H., and Zhang, B.-T. Selecmix: Debiased learning by contradicting-pair sampling. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

Javaloy, A. and Valera, I. Rotograd: Gradient homogenization in multitask learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2022.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

Jung, S., Chun, S., and Moon, T. Learning fair classifiers with partially annotated group labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Kim, E., Lee, J., and Choo, J. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 14992–15001, 2021.

Kim, N., Hwang, S., Ahn, S., Park, J., and Kwak, S. Learning debiased classifier with biased committee. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips,

R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proc. International Conference on Machine Learning (ICML)*, pp. 5637–5664. PMLR, 2021.

Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W. T., Isola, P., Globerson, A., Irani, M., et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 693–702, 2021.

Lee, J., Kim, E., Lee, J., Lee, J., and Choo, J. Learning debiased representation via disentangled feature augmentation. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Li, X. and Gong, H. Robust optimization for multilingual translation with imbalanced data. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Li, Y. and Vasconcelos, N. Repair: Removing representation bias by dataset resampling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9572–9581, 2019.

Li, Z., Hoogs, A., and Xu, C. Discover and mitigate unknown biases with debiasing alternate networks. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 270–288. Springer, 2022.

Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C. C., Xu, C., and Ibrahim, M. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20071–20082, 2023.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *Proc. International Conference on Machine Learning (ICML)*, pp. 6781–6792. PMLR, 2021a.

Liu, L., Li, Y., Kuang, Z., Xue, J., Chen, Y., Yang, W., Liao, Q., and Zhang, W. Towards impartial multi-task learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2021b.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: Training debiased classifier from biased classifier. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *Proc. International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=_F9xpOrqyX9.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *Proc. International Conference on Machine Learning (ICML)*, pp. 8346–8356. PMLR, 2020.

Scimeca, L., Oh, S. J., Chun, S., Poli, M., and Yun, S. Which shortcut cues will dnns choose? a study from the parameter-space perspective. In *Proc. International Conference on Learning Representations (ICLR)*, 2022.

Sener, O. and Koltun, V. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Tartaglione, E., Barbano, C. A., and Grangetto, M. End: Entangling and disentangling deep representations for bias correction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13508–13517, 2021.

Teney, D., Abbasnejad, E., and van den Hengel, A. Unshuffling data for improved generalization in visual question answering. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1417–1427, 2021.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Wang, H., He, Z., Lipton, Z. C., and Xing, E. P. Learning robust representations by projecting superficial statistics out. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Wang, H., He, Z., Lipton, Z. C., and Xing, E. P. Learning robust representations by projecting superficial statistics out. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Wu, S., Yuksekgonul, M., Zhang, L., and Zou, J. Discover and cure: Concept-aware mitigation of spurious correlation. In *Proc. International Conference on Machine Learning (ICML)*, 2023.

Yang, Y., Nushi, B., Palangi, H., and Mirzasoleiman, B. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proc. International Conference on Machine Learning (ICML)*, 2023.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 25407–25437. PMLR, 2022.

Zhang, M. and Ré, C. Contrastive adapters for foundation model group robustness. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *Proc. International Conference on Machine Learning (ICML)*, 2022.

Zhu, W., Zheng, H., Liao, H., Li, W., and Luo, J. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 15002–15012, October 2021.

Appendices

A.1. Explaining spurious correlation in machine learning

Spurious correlation refers to a relationship between variables that appears to be statistically significant but is actually caused by some other factor. For example, if most samples of class a have an attribute i and most samples of class b have an attribute j, where $a \neq b$ and $i \neq j$, and neither attribute i nor j is not the actual cause of the target classes, then a trained model can rely on the bias attributes to classify most training samples. In this case, the bias attributes i and j can be considered as spuriously correlated with the target class, and each bias attribute is the bias-guiding attribute for its respective class.

A.2. The construction process of MultiCelebA

In this section, we explain the construction of the two-bias setting of MultiCelebA, including the selection of the target class and bias types (Section A.2.1). We then describe two additional evaluation settings: the three-bias setting and the four-bias setting (Section A.2.2).

Figure A1: Unbiased accuracy for predicting each attribute

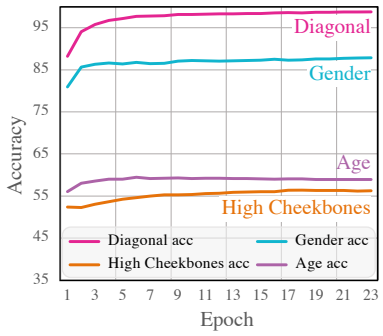


Table A1: Configuration of MultiCelebA in two biases setting

Group	{Target class, Bias type 1, Bias type 2}	# of training samples
GG	{`High Cheekbones`, `Female`, `Young`}	44582
	{`Low Cheekbones`, `Male`, `Old`}	16220
GC	{`High Cheekbones`, `Female`, `Old`}	2200
	{`Low Cheekbones`, `Male`, `Young`}	800
CG	{`High Cheekbones`, `Male`, `Young`}	2200
	{`Low Cheekbones`, `Female`, `Old`}	800
CC	{`High Cheekbones`, `Male`, `Old`}	110
	{`Low Cheekbones`, `Female`, `Young`}	40

A.2.1. Two-bias setting of MultiCelebA

We first selected `gender` and `age` as bias types among the 40 attributes of CelebA. We then chose `high-cheekbones` as the target class and verified whether the target class and both bias types exhibit spurious correlations and invoke shortcut learning for ERM.

Scimeca et al. (2022) examined how deep neural networks exhibit a preference for attributes based on their ease of learning. Following Scimeca et al. (2022), we assessed the preference of ResNet18 for the target class (`high-cheekbones`) and biases (`gender` and `age`) by evaluating a model trained on diagonal set (GG group in the main paper), where all samples are spuriously correlated with all biases, as shown in Figure A1. Each line on Figure A1 represents unbiased accuracy of a testing attribute, which we used to evaluate the model's ability to predict each testing attribute. ResNet18 exhibited higher unbiased accuracy for `gender` and `age` compared to that of `high-cheekbones`, indicating that the model tends to exploit `gender` and `age` as shortcuts when learning `high-cheekbones` classification task on MultiCelebA. With `high-cheekbones`, `gender`, and `age` labels, we subsampled the training set of CelebA to simulate challenging scenarios where training data are extremely biased. The configurations of MultiCelebA in the two-bias setting are shown in Table A1.

Table A2: Performance in CCCC and UNBIASED (%) on MultiCelebA in four biases for evaluation two biases for training setting.

Method	CCCC	UNBIASED
ERM	$6.0_{\pm 0.9}$	$59.0_{\pm 0.7}$
GroupDRO	$29.6_{\pm 3.7}$	$62.3_{\pm 0.9}$
Ours	$43.1_{\pm 1.5}$	$65.8_{\pm 0.2}$

A.2.2. three-bias and four-bias settings of MultiCelebA

Next, we extend our MultiCelebA dataset with two bias types by adding extra bias types for further evaluation settings. To this end, we explain which attributes can be considered as bias types. For evaluating debiased training algorithms, the training set should be designed in such a way that ERM exploits undesirable shortcuts stemming from spurious correlations between the target labels and predefined bias types (*i.e.*, the ERM solution trained on the set is biased). To this end, the selected bias types have to hold the two conditions below:

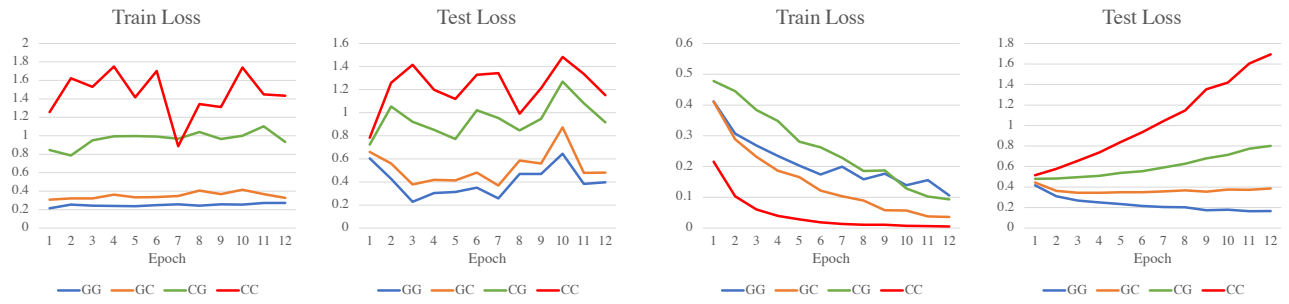
- The bias types are all spuriously correlated to the target class.
- Not every sample in the training set has the same labels for a pair of the bias types.

The first condition is trivial, and the second is required to reject bias-type candidates redundant to those previously chosen.

Among the 37 attributes of CelebA, we identified `mouth slightly open` as a third bias type, and then we chose `smiling` as a fourth bias type, both of which satisfy the two conditions. We demonstrated the superior results on the three-bias setting and the four-bias setting, as shown in Table 2 and Table A2, respectively.

A.2.3. Considerations on exceeding four bias types

We empirically found that it is difficult to identify more than five attributes (*i.e.*, bias types) that satisfy the two conditions at once in existing natural image datasets, even for CelebA with 40 attributes. Furthermore, some groups often become empty sets when the number of bias types increases; this is a critical issue particularly for test data as some groups with zero cardinality are never used for evaluation.



(a) Train with ERM

(b) Train with the averaged group losses

Figure A2: Group losses of (a) model with ERM (b) model with the averaged group losses.

A.3. Empirical analysis of the objective for optimizing the group-scaling parameter

When training a model by ERM, the training loss for the small group is larger than that for the large group, and a similar trend is observed in the test loss, as shown in Figure A2(a). Thus, increasing the weight of groups with a larger training loss can be beneficial in giving more weight to the minority group.

However, when we compute the objective by averaging group-wise losses, the gap between training loss and test loss for each group varies depending on group size, as shown in Figure A2(b). This phenomenon arises because smaller groups are more susceptible to memorization effects.

To mitigate the gap between training loss and test loss resulting from the memorization effect, Sagawa et al. (2019) proposed the use of strong regularization on model parameters and an increase in the weight of group with large training loss. This approach of increasing the weight of groups or samples with large train loss has evolved into a standard practice within debiased training methods.

However, in scenarios involving multiple biases, a trained model easily overfits minor groups (*e.g.*, the CC group) that have a small number of samples, leading to decreased loss scales for such groups and consequently neglecting them in training, resulting in a biased model. Applying a strong ℓ_2 regularizer to model parameters was successful in the single bias settings,

but we empirically found that it does not work as desired in multi-bias settings. This is particularly due to minor groups, *e.g.*, the CC group in the two-bias setting, having an extremely small number of samples. That is the underlying cause of the inferior performance of GroupDRO in multiple biases settings compared to Upweighting.

In contrast, our algorithm optimizes the group-scaling parameter based on MOO. In a nutshell, it increases the weights of groups with low loss magnitudes. Since minor groups usually exhibit low loss scales as a model easily overfits them due to their small cardinality, our algorithm emphasizes the impact of minor groups in training.

A.4. Implementation details

A.4.1. Datasets

To evaluate our framework, we consider three multi-bias datasets, *i.e.*, MultiCelebA, Multi-Color MNIST, and UrbanCars and three single-bias datasets, *i.e.*, Waterbirds, CelebA, and BFFHQ. In what follows, we provide details of each dataset.

MultiCelebA. First, we mainly consider MultiCelebA in two biases setting as the dataset to evaluate debiased training algorithms. As introduced in Section 4, this dataset requires training a model to predict whether if a given face image has `high-cheekbones` or not. Each image is additionally annotated with `gender` and `age` attributes which are spuriously correlated with the target `high-cheekbones`. For MultiCelebA in three biases setting, each image is annotated with `gender`, `age`, and `mouth slightly open` attributes which are spuriously correlated with the target `high-cheekbones`.

UrbanCars. UrbanCars (Li et al., 2023) is a dataset created by synthesizing `background`, `co-occurring object`, and `car` to generate multi-biased images. Its task involves classifying whether an image contains `urbancars` or not.

Multi-Color MNIST. We consider Multi-Color MNIST dataset proposed by Li et al. (2022). Its task is to predict the digit number from an image. The digit numbers are spuriously correlated with left and right background colors, coined `left-color` and `right-color`, respectively. As proposed by Li et al., we set the proportion of bias-guiding attributes to be 99% and 95% for `left-color` and `right-color`, respectively.

Waterbirds. Waterbirds (Sagawa et al., 2019) is a single-bias dataset consisting of bird images. Given an image, the target is `bird-type`, *i.e.*, whether if the bird is "landbird" or a "waterbird." The biased attribute is `background-type`, *i.e.*, whether if the image contains "land" or "water." The proportion of biased attribute is set to 95%.

CelebA. CelebA (Liu et al., 2015) is a face recognition dataset where each sample is labeled with 40 attributes. Following the previous settings (Sagawa et al., 2019; Yao et al., 2022), we use `HairColor` as the target and `gender` as the bias attribute.

BFFHQ. BFFHQ (Lee et al., 2021) is a real-world face image dataset curated from FFHQ. Its task is to predict the age from an image. the age is spuriously correlated with gender attributes. The proportion of bias-guiding attributes is 99.5%.

A.4.2. Baselines

We extensively compare our algorithm against the existing debiased training algorithms. In particular, one can categorize a baseline by whether it explicitly uses the supervision on biased attributes, *i.e.*, bias labels, or not. To this end, compare our method with nine training algorithms, consisting of five that do not use the bias label and six that do. Algorithms that do not require using the bias label are as follows: (1) training with vanilla ERM, (2) LfF (Nam et al., 2020) employs a reweighting scheme where samples that are more likely to be misclassified by a biased model are assigned higher weights, (3) JTT (Liu et al., 2021a) retrains a model using different weights for each group, where the groups are categorized as either bias-guiding or bias-conflicting based on an ERM model, (4) EIIL (Creager et al., 2021) conducts domain-invariant learning, (5) PGI (Ahmed et al., 2021) matches the class-conditional distribution of groups by introducing predictive group invariance, and (6) DebiAN (Li et al., 2022) utilizes a pair of alternate networks to discover and mitigate unknown biases sequentially. We consider debiased training methods using bias attribute labels as follows: (1) Upsampling assigns higher sampling probability to minority groups, (2) Upweighting assigns scales the sample-wise loss to be higher for minority groups; group weight $=$ (# of training samples)$/$(# of group samples), (3) GroupDRO (Sagawa et al., 2019) computes group-scaling weights using group-wise training loss to upweight the worst-case group samples. (4) SUBG (Sagawa et al., 2020) proposes a group-balanced sampling scheme by undersampling the majority groups. (5) LISA (Yao et al., 2022) performs group mixing (mixup) augmentation to learn from both intra- and inter-group information. (6) DFR (Kirichenko

et al., 2023) retrains the last layer of an ERM model using a balanced set obtained through undersampling.

Table A3: The search spaces of hyperparameters.

Hyperparameter	Search space
Learning rate η_1, η_2	$\{5e{-}4, 2e{-}4, 1e{-}4, 5e{-}3, 2e{-}3, 1e{-}3, 5e{-}2, 2e{-}2, 1e{-}2\}$
Weight decay	$\{0, 1e{-}4, 1e{-}2, 1e{-}1, 1\}$
Update frequency U	$\{1, 5, 10, 50\}$

Table A4: Hyperparameters of our method.

	MultiCelebA	Multi-Color MNIST	UrbanCars	Waterbirds	CelebA	BFFHQ
Batch size	512	512	128	128	128	64
Learning rate η_1	$2e{-}4$	$2e{-}2$	$1e{-}2$	$1e{-}3$	$2e{-}3$	$2e{-}3$
Learning rate η_2	$1e{-}2$	$2e{-}3$	$1e{-}3$	$1e{-}3$	$1e{-}4$	$5e{-}4$
Update frequency U	10	50	10	5	1	1
Optimizer	SGD	Adam	SGD	SGD	Adam	Adam

A.4.3. Hyperparameters

We tune all hyperparameters, as well as early stopping, based on highest WORST for MultiCelebA, UrbanCars, Waterbirds, and CelebA on validation set, except for ERM. For Multi-Color MNIST and BFFHQ, we tune hyperparemters based on highest UNBIASED on test set, following the previous work (Li et al., 2022; Lee et al., 2021). We use a single GPU (RTX 3090) for training. Following the previous work (Lee et al., 2021; Hwang et al., 2022), we conduct experiments on BFFHQ using ResNet18 with random initialization as the neural network architecture. For remaining datasets, we initialized the model with parameters pretrained on ImageNet. The hyperparameter search spaces used in all experiments conducted in this paper are summarized in Table A3. The selected hyperparameters for our method are represented in Table A4. Furthermore, the search space for the upweight value λ_{up} in JTT is $\{5, 10, 20, 30, 40, 50, 100\}$. JTT (Liu et al., 2021a) and DFR (Kirichenko et al., 2023) utilize the ERM model as a pseudo labeler and frozen backbone network, respectively. We used the ERM model as reported in the literature for our implementation of these methods. We did not use a learning rate scheduler in any of the experiments.

Given that the proportion of samples from minority groups can impact the performance of debiased training, we trained DFR exclusively on the training set to ensure a fair comparison, which is denoted as DFR^{tr}_{tr}.

A.4.4. Training existing methods on multi-bias setting

When training a model using SUBG (Sagawa et al., 2020), GroupDRO (Sagawa et al., 2019) and DFR (Kirichenko et al., 2023), we grouped the training set based on the same pair of bias attribute $\boldsymbol{b}$ and target class t and followed the approach outlined in the original paper.

LISA (Yao et al., 2022) adopts the two kinds of selective augmentation strategies, Intra-label LISA and Intra-domain LISA. In the multi-bias setting, Intra-label LISA (LISA-L) interpolates samples with the same target label but different all bias labels ($t^{(m)} = t^{(m')}$, $b_d^{(m)} \neq b_d^{(m')}\ \forall d$). Intra-domain LISA (LISA-D) interpolates samples with the same bias labels but different target label ($t^{(m)} \neq t^{(m')}$, $\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m')}$).

When training a model using biased training methods that do not require bias labels, such as LfF (Nam et al., 2020), JTT (Liu et al., 2021a), and DebiAN (Li et al., 2022), we followed the approach outlined in the original paper without modification, regardless of the number of bias types presented in the dataset.

A.4.5. Evaluation metrics

We consider various metrics to evaluate whether if the trained model is biased towards a certain group in the dataset. We remark that no metric is universally preferred over others, *e.g.*, worst-group and average-group accuracy reflects different aspects of a debiased training algorithm.

For the multi-bias datasets, we evaluate algorithms using the average accuracy for each of the four groups categorized by the guiding or conflicting nature of the biases: {GG, GC, CG, CC}. Here, G and C describes whether a group contains bias-guiding or bias-conflicting samples for each bias type, respectively. For example, GC group for MultiCelebA is an intersection of bias-guiding samples with respect to the first bias type, *i.e.*, `gender`, and bias-conflicting samples with respect to the second bias type, *i.e.*, `age`. We also report the average of these four metrics, denoted as UNBIASED. Conceptually, the GG metric can be high regardless of whether a model is biased or not. However, the CC metric can be high only when a model is debiased from all spurious correlations. Similarly, the GC accuracy can be high only when a model is debiased from the spurious correlation of the second bias type. Meanwhile, the InDist accuracy measures the average accuracy on biased data. A biased model will achieve high scores in InDist and GG metrics, but low scores in Unbiased and CC metrics. This also means that huge performance variations among GG, GC, CG, and CC groups suggest model bias to spurious correlations (as shown in the scores of ERM). The effectiveness of debiased training algorithms can be assessed by examining the Unbiased accuracy first (higher is better) and in turn the CC/Worst accuracy (higher is better); the GC, CG, and GG accuracies also have to be sufficiently high compared to the CC accuracy.

Next, for the single-bias datasets, the minimum group average accuracy is reported as WORST, and the weighted average accuracy with weights corresponding to the relative proportion of each group in the training set as INDIST (in-distribution) following Sagawa et al. (2019).

In calculating the GG, GC, CG, CC accuracies on the MultiCelebA dataset, we excluded the impact of class imbalance within each group by first computing the mean accuracy for each class within the group, and then taking the average of the class accuracies to obtain the group accuracy.

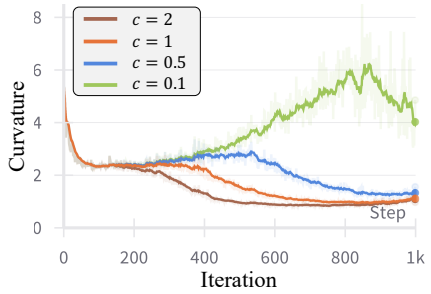
A.4.6. Interpretation of the results on MultiCelebA

In Table 1, we analyzed whether a model is biased toward the two bias types based on the difference between GG, GC, CG, CC, while also evaluating the UNBIASED accuracy. Let G* denote the combination of GG and GC, and similarly for C* and others. A model is biased toward `gender` attribute if there is a significant difference between the G* and C* combinations, whereas a significant difference between the *G and *C combinations indicates bias toward `age` attribute.

Table A5: The number of groups for training on Multi-Color MNIST, UrbanCars, MultiCelebA, Waterbirds, CelebA, and BFFHQ.

Benchmark	GroupDRO	Ours
Multi-Color MNIST	40	4
UrbanCars	8	4
MultiCelebA	8	4
Waterbirds	4	2
CelebA	4	2
BFFHQ	4	2

Figure A3: Local loss curvature of the loss landscape of model parameter.



A.5. Impact of the loss function on local curvature

According to Li & Gong (2021), the second term in Eq. (4), $\left\|\sigma(\boldsymbol{\alpha})^\top(\nabla_\theta L(\theta))\right\|_2^2$, serves as an approximation for the local curvature of the loss landscape associated with the model parameter θ. Although this term is minimized by updating α, the local curvature of loss landscape of model parameter is reduced. To verify this, we conducted an ablation study by adjusting the relative weight of the second term in Eq. (4) using constant c. The objective formula for this experiment is presented as:

$$\hat{L}_{\boldsymbol{\alpha}} = \sigma(\boldsymbol{\alpha})^\top L(\theta) + c\lambda \left\|\sigma(\boldsymbol{\alpha})^\top(\nabla_\theta L(\theta))_\dagger\right\|_2^2. \tag{5}$$

We updated $\boldsymbol{\alpha}$ and λ by minimizing $\hat{L}_{\boldsymbol{\alpha}}$ and θ by minimizing Eq. (3). Figure A3 demonstrates how the loss curvature evolves over training iterations. We observed that as the value of c decreases, there is a corresponding increase in loss curvature. Hence, minimizing the second term in Eq. (4) contributes to improving model generalization.

A.6. Comparison of the number of groups in previous methods and our method

We compared the number of groups for training with GroupDRO on all the benchmarks we used. For all experiments in our paper, the same annotations have been used across ours and the other methods using bias labels, *i.e.*, upsampling, upweighting, SUBG, GroupDRO, LISA, and DFR. For example, on multiple bias settings with two bias types, all the aforementioned methods including ours utilize labels for two bias types for group division. As shown in Table A5, our method defines a smaller number of groups compared with GroupDRO for debiased training. The number of groups of GroupDRO, denoted as N_{GroupDRO}, and that of ours, denoted as N_{Ours}, are calculated as follows:

$$N_{\text{GroupDRO}} = C \times (\#\text{of attributes in bias type 1}) \times \ldots \times (\#\text{of attributes in bias type D}) \geq 2^D \times C, \tag{6}$$

$$N_{\text{Ours}} = 2 \times 2 \times \ldots \times 2 = 2^D, \tag{7}$$

where C is the number of classes, D is the number of bias types, and $2^D \times C$ is the lower bound of N_{GroupDRO}. Since the number of classes C is greater than 1, N_{Ours} is always smaller than N_{GroupDRO}. The number of groups for each dataset is presented in Table A5.

A.7. Analysis of the limitation of CivilComments as a multi-bias setting

CivilComments has been used to benchmark debiased training algorithms in a single spurious correlation setting (Borkan et al., 2019; Koh et al., 2021). Its target task is to classify an online comment into toxic or non-toxic, and the class label is spuriously correlated with certain demographic identities (e.g., male, female, White, Black, LGBTQ, Muslim, Christian, and other religions) mentioned in the comment.

To investigate whether CivilComments involves multiple spurious correlations, we first categorized the demographic labels into 3 types: Gender: {male, female, LGBTQ}, Race: {White, Black}, Religions: {Muslim, Christian, and other religions}. We then examined if these 3 types are spuriously correlated with the target class (*i.e.*, toxic). If most samples of class a have a bias attribute i and most samples of class b have a bias attribute j ($a \neq b$ and $i \neq j$), then a trained model can rely on the bias attributes to classify most training samples, and the bias attributes i and j can be considered as spuriously correlated with the target class in this case. Based on this notion, we investigated the type-wise data population of the dataset as shown in Table A6, and found that none of the three bias types are spuriously correlated with the toxic class.

Table A6: Training data population of non-toxic and toxic comments based on identity presence across gender, race, and religion.

	Gender		Race		Religion	
	no identities	has identities	no identities	has identities	no identities	has identities
non-toxic	188585 (70%)	49938 (19%)	202071 (75%)	36452 (14%)	222348 (83%)	16175 (6%)
toxic	21207 (8%)	9308 (3%)	24852 (9%)	5663 (2%)	24000 (9%)	6515 (2%)

A.8. Computational complexity

To demonstrate the scalability of our algorithm, we show that the overall computational complexity of our algorithm grows slower than the number of groups 2^D by the proof below.

Preliminary:

There are six factors contributing to the total computational complexity of our debiased training algorithm, as enumerated below:

- D: the number of bias types (the number of groups is then 2^D)
- U: the update period of $\boldsymbol{\alpha}$ (e.g., $U = 10$ means $\boldsymbol{\alpha}$ is updated every 10 iterations.)
- a: the computational complexity for forward process per epoch
- b: the computational complexity for backward process per epoch
- c: the computational complexity for model parameter update per epoch
- d: the computational complexity for $\boldsymbol{\alpha}$ update per epoch
- e: the computational complexity for λ update per epoch

The overall computation complexity of our algorithm with D bias types for each epoch is then denoted and defined by $C_D := a + b + c + 2^D \times b/U + d/U + e/U$.

Proposition: The overall complexity (C_D) grows slower than the number of groups (2^D).

Proof by Contradiction: We first assume a negation of the proposition: "The overall complexity increases at least linearly with the number of groups."

Under this assumption, $C_{D+1} \geq 2 \cdot C_D$, where $D \geq 1$.

Then, regarding $C_D := a + b + c + 2^D \times b/U + d/U + e/U$,

$C_{D+1} \geq 2 \cdot C_D$

$\Leftrightarrow a + b + c + 2^{D+1} \times b/U + d/U + e/U \geq 2a + 2b + 2c + 2^{D+1} \times b/U + 2d/U + 2e/U$

$\Leftrightarrow 0 \geq a + b + c + d/U + e/U$,

which is a contradiction since the right-hand side is always greater than 0. Therefore, the assumption is false and the proposition holds.

The proposition suggests that *the total time complexity of training in our algorithm grows slower than the number of groups* (*i.e.*, 2^D); a simple analysis reveals that $O(C_D) = 2^D$ **when** D **goes to the infinity**.

To empirically verify this conclusion, we further increased the number of bias types of MultiCelebA up to four and estimated the wall-clock time of our training algorithm versus the number of bias types. As demonstrated in Table A7, the wall-clock time does not increase exponentially, even the number of groups N is exponentially increased in the number of bias types D.

Table A7: Comparison of time complexity according to the number of bias types

# of bias types (D)	# of groups (N)	Training time per 1 epoch	Relative training time (compare to D=2)	Relative training time (compare to D=3)
2	4	67s	-	-
3	8	82s	1.22 times	-
4	16	113s	1.69 times	1.37 times