Removing Multiple Biases through the Lens of Multi-task Learning

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

Abstract

We consider the problem of training an unbiased and accurate model using a biased dataset with multiple biases. One of the major challenges is to balance improving overall accuracy and ignoring all the biases. To address this, we provide a novel framework connecting the problem to multi-task learning (MTL). To be specific, our framework divides training data into several groups according to their effects on the model bias, and defines each task of MTL as solving the target problem for each group. It in turn trains a single model for all the tasks with a weighted sum of task-wise losses as the training objective, while optimizing the weights as well as the model parameters. At the heart of our method lies the weight adjustment algorithm, which is rooted in a theory of multi-objective optimization and guarantees a Pareto-stationary solution. Our algorithm achieved the state of the art on two datasets with multiple biases, and demonstrated superior performance on conventional single-bias datasets.

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) observed how training a classifier with ERM is prone to *spurious correlations* between the target label and other non-relevant attributes. 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 investigated debiased training algorithms, *i.e.*, algorithms training models to ignore spurious correlations in a dataset (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., 2021; Kim et al., 2022). Their main idea is to balance the performance of models on samples that agree with the spurious correlation, *i.e.*, bias-guiding samples, and the samples that disagree, *i.e.*, bias-conflicting samples. While these algorithms have shown promising results, they are evaluated in an unrealistic setting where only a single type of spurious correlation is present in the training dataset.

We advocate that the debiased training algorithms should be evaluated under a more realistic scenario with multiple biases. This scenario is challenging since the intersection of bias-conflicting samples, *i.e.*, "clean" samples that disagree with all the spurious correlations, are extremely rare. Furthermore, the tasks of mitigating different types of spurious correlation may even conflict with each other. Indeed, we empirically observe that the promising performance of prior work failed to generalize for the multi-bias scenarios.

In this work, we develop an algorithm for debiased training under the presence of multiple spurious correlations. Our novel idea is to realize the removal of different spurious correlations as a multi-task learning (MTL) problem. From the lens of MTL, performance degradation from simultaneously mitigating the spurious correlation can be interpreted as conflicts between different tasks. We develop a new multi-objective optimization (MOO) algorithm to prevent this issue and train a model to reach Pareto-optimal performance. To be specific, we realize our setting as a MTL problem by dividing the entire training set into multiple groups where all data in 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. Unlike the conventional MTL problems, this results in tasks that share the space of target predictions but differ in the distribution of the biased attributes.

Next, our training strategy stems from the existing MOO algorithm (Désidéri, 2012). Namely, we train our model to reach Pareto-optimal performance with respect to the aforementioned tasks. To this end, we devise an algorithm to iteratively adjust task-wise importance weights so that the model parameters converge to a Pareto-stationary point. Our algorithm is also interpreted as an optimization to find a flat minimum of the loss landscape (Li & Gong, 2021),

[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>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

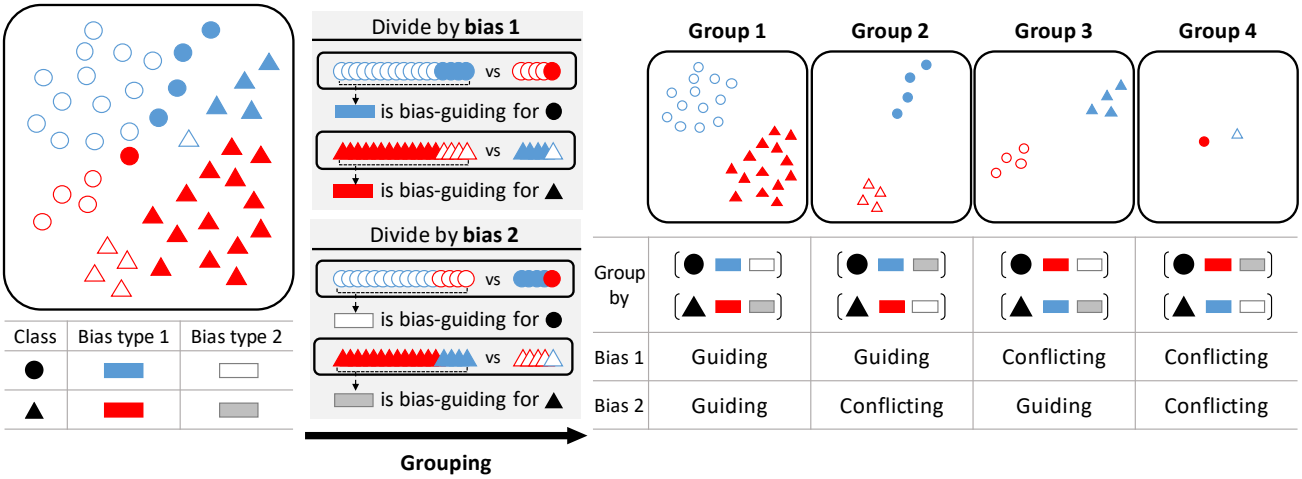


Figure 1. Overview of our grouping strategy. (*left*) For each sample, its shape means its class while its color and pattern indicate its attributes for two different bias types, respectively. (*middle*) For each class and each bias type, we examine which bias attribute is spuriously correlated with the class and induces the model bias in consequence. (*right*) Samples that guide to or conflict with each bias type in the same way are grouped together.

which has shown to improve the generalization of the model performance.

Finally, we also present a new multi-bias benchmark, dubbed MultiCelebA, for evaluation of debiased training algorithms.

We extensively evaluate our algorithm on two multi-bias benchmarks including MultiCelebA and two single-bias benchmarks. Our algorithm outperformed all the existing debiased training methods. Our ablation studies verify the importance of each algorithmic component.

The main contribution of this paper is four-fold:

- This work is the first to interpret debiased training as a MTL problem. Based on this notion, we present a novel and effective debiased training algorithm.
- We present a new real-image multi-bias benchmark for evaluating debiased training methods under a realistic and challenging condition.
- We benchmarked existing methods for debiased training in multiple biases settings and demonstrated that they struggle when training data exhibit multiple biases.
- Our algorithm achieved the state of the art on two datasets with multiple biases. Moreover, it also showed superior performance on conventional single-bias datasets.

2. Proposed method

To tackle multiple biases, we propose a novel debiased training algorithm based on a theory of MOO (Désidéri, 2012). Our algorithm divides training data into several groups according to their effects on the model bias, defines each task of MTL as solving the target problem for each group, and trains a single model for all the tasks while optimizing weights of the tasks as well as model parameter. The rest of this section first introduces the MOO theory that motivates our work (Section 2.1) and then describes the proposed algorithm in detail (Section 2.2).

2.1. Preliminary: MTL as MOO

We formulate MTL as a problem to optimize a parameter θ with respect to a collection of task-wise training loss functions $L(\theta) = [\mathcal{L}_1(\theta), \dots, \mathcal{L}_N(\theta)]^\top$. 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 2.1 (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 the training objective of deep neural networks. Instead, one may consider using gradient-based optimization to find a parameter satisfying the Pareto stationarity (Désidéri, 2012), *i.e.*, a state where a convex combination of task-wise gradients equals a zero-vector. Pareto stationarity is a necessary condition for Pareto optimality if the loss functions in $L(\theta)$ are smooth.

Definition 2.2 (Pareto stationarity)**.** A parameter θ^* is Pareto-stationary if there exists a task-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 $\mathbf{1} = [1, \dots, 1]^\top \in \mathbb{R}^N$.

We consider multi-gradient descent algorithm (MGDA) (Désidéri, 2012) to search for a Pareto-stationary

parameter. For U-th step, MGDA finds a task-scaling parameter α which combines the task-wise gradients $\nabla_\theta L$ to be approximately a zero vector based on solving the following optimization problem:

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

After finding $\boldsymbol{\alpha}$, MGDA performs a gradient-based update on the parameter θ with respect to $\boldsymbol{\alpha}^\top L(\theta)$. We also note that (Li & Gong, 2021) interpreted the MGDA algorithm as regularizing the local curvature of loss landscapes.

2.2. Debiased training by MOO

First, we introduce our formulation of debiased training as MOO. To this end, we coin an attribute to be bias-guiding if it is spuriously correlated with the target class and held by a majority of training samples. In contrast, we coin an attribute to be bias-conflicting if it does not exhibit such a correlation and is held by a minority of samples. It is well known that models trained using empirical risk minimization (ERM), *e.g.*, deep neural networks, are prone to over-fitting to bias-guiding attributes and exhibit poor performance under the presence of bias-conflicting attributes.

Therefore, we aim to balance the performance of the model on samples with bias-guiding and bias-conflicting attributes using MOO. Our main idea is to formulate each objective as optimizing over a group of samples constructed using the existence of bias-guiding or bias-conflicting attributes. The remainder of this section elaborates on the grouping strategy and the MOO formulation for debiased training.

2.2.1. Grouping strategy

As illustrated in Figure 1, 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 $\mathcal{D} = \{(x^{(m)}, t^{(m)})\}_{m=1}^M$, where each sample $x^{(m)}$ is associated with a target $t^{(m)}$ and a list of attributes $\boldsymbol{b}^{(m)} = [b_1^{(m)}, \dots, b_D^{(m)}]^\top$. 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 a group label $\boldsymbol{g}^{(m)}$. We note how this grouping policy differs from prior works (Sagawa et al., 2019; Kirichenko et al., 2022; Nam et al., 2022; Sagawa et al., 2020; Zhang et al., 2022) that use the targets and the attributes as the group labels. Each group in our method contains samples from all the classes, while existing ones only keep a group of samples with the same target and the same attributes.

We remark that our grouping strategy can be interpreted as an MTL problem where the tasks share the space of targets, but are defined on different groups of samples. Our goal is to train a model capable of accurately classifying samples from all the groups, *i.e.*, its performance is not biased towards a certain group. Similar to MTL, naïvely minimizing a linear combination of loss functions for each group leads to conflicts between bias-guiding groups and bias-conflicting groups.

2.2.2. Training algorithm

Based on the grouping strategy proposed in Section 2.2.1, we propose a framework to optimize over $N = 2^D$ groups while minimizing the conflict between group-wise loss functions. To this end, we $L(\theta) = [\mathcal{L}_1(\theta), \dots, \mathcal{L}_N(\theta)]^\top$ 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 $\boldsymbol{\alpha}^\top \mathbf{1} = 1$. To address between-group conflicts, we propose adjusting the group-scaling parameter $\boldsymbol{\alpha}$ such that the training converges to a Pareto-stationary point with a flat loss landscape.

Our goal is to minimize the training objective $\boldsymbol{\alpha}^\top L(\theta)$ while simultaneously adjusting the group-scaling parameter $\boldsymbol{\alpha}$ to minimize Eq. (2). To this end, we simultaneously optimize the parameters with respect to the following loss function:

$$\hat{L}(\theta) = \boldsymbol{\alpha}^\top L(\theta) + \lambda \left\|\boldsymbol{\alpha}^\top (\nabla L(\theta))_\dagger\right\|_2^2, \tag{3}$$

where $\boldsymbol{\alpha} \geq \mathbf{0}$, $\boldsymbol{\alpha}^\top \mathbf{1} = 1$, $(\cdot)_\dagger$ denotes the stop-gradient operator, and λ is a Lagrangian multiplier for Eq. (2). In practice, we re-parameterize group-scaling parameter using a softmax function, *i.e.*, set $\boldsymbol{\alpha} = \mathrm{SoftMax}(\bar{\boldsymbol{\alpha}})$. This allows optimizing over $\bar{\boldsymbol{\alpha}}$ with gradient-based updates without violating the constraints $\boldsymbol{\alpha} \geq \mathbf{0}$ and $\boldsymbol{\alpha}^\top \mathbf{1} = 1$. We update the group scaling parameter $\boldsymbol{\alpha}$ with gradient descent and the Lagrangian multiplier λ with gradient ascent every U iterations. The learning process of our method is also described in Algorithm 1 in the Appendix.

3. Experiments

3.1. Setup

Datasets. To evaluate our framework, we consider two multi-bias datasets, *i.e.*, MultiCelebA and Multi-Color MNIST, and two single-bias datasets, *i.e.*, Waterbirds and CelebA. We provide details of each dataset including the proposed new benchmark, MultiCelebA, in the Appendix.

Evaluation metrics. For the multi-bias datasets, we evaluate algorithms using the average accuracy for each of the

Table 1. GG, GC, CG, CC, UNBIASED, WORST, and INDIST metrics (%) evaluated on the MultiCelebA dataset. The first element of each of the four combinations {GG, GC, CG, CC} represents the bias type of `gender`, while the second element represents the bias type of `age`. We mark the best and the second-best performance in **bold** and underline, respectively.

Method	Bias label	GG	GC	CG	CC	UNBIASED	WORST	INDIST
ERM	✗	$90.97_{\pm 4.10}$	$\underline{84.43}_{\pm 1.55}$	$50.06_{\pm 3.52}$	$30.29_{\pm 4.09}$	$63.94_{\pm 1.46}$	$23.88_{\pm 6.48}$	$89.18_{\pm 2.32}$
LfF (Nam et al., 2020)	✗	$79.82_{\pm 2.58}$	$71.66_{\pm 2.18}$	$80.20_{\pm 1.68}$	$71.52_{\pm 3.32}$	$75.80_{\pm 0.47}$	$66.83_{\pm 1.20}$	$81.85_{\pm 3.11}$
JTT (Liu et al., 2021)	✗	$73.35_{\pm 2.68}$	$58.67_{\pm 5.10}$	$60.22_{\pm 9.35}$	$52.08_{\pm 2.17}$	$62.89_{\pm 1.17}$	$49.64_{\pm 2.58}$	$75.75_{\pm 5.45}$
DebiAN (Li et al., 2022)	✗	$64.39_{\pm 30.4}$	$63.60_{\pm 22.2}$	$49.80_{\pm 7.58}$	$45.50_{\pm 13.2}$	$55.82_{\pm 11.7}$	$25.72_{\pm 6.00}$	$66.82_{\pm 34.1}$
Upsampling	✓	$79.79_{\pm 1.45}$	$80.97_{\pm 1.30}$	$76.68_{\pm 1.09}$	$75.56_{\pm 1.16}$	$78.25_{\pm 0.75}$	$71.54_{\pm 1.96}$	$82.59_{\pm 0.79}$
Upweighting	✓	$78.96_{\pm 4.05}$	$79.20_{\pm 6.02}$	$\underline{80.83}_{\pm 0.04}$	$\underline{78.65}_{\pm 3.64}$	$\underline{81.16}_{\pm 3.87}$	$\underline{73.47}_{\pm 4.23}$	$83.40_{\pm 5.92}$
GroupDRO (Sagawa et al., 2019)	✓	$81.17_{\pm 0.98}$	$81.15_{\pm 1.24}$	$76.74_{\pm 1.48}$	$74.62_{\pm 0.40}$	$78.43_{\pm 0.68}$	$71.58_{\pm 1.07}$	$83.46_{\pm 0.65}$
SUBG (Sagawa et al., 2020)	✓	$77.09_{\pm 1.03}$	$78.37_{\pm 0.70}$	$77.46_{\pm 1.66}$	$77.95_{\pm 1.24}$	$77.72_{\pm 0.60}$	$69.57_{\pm 0.74}$	$80.31_{\pm 1.11}$
LISA (Yao et al., 2022)	✓	$82.84_{\pm 1.29}$	$83.19_{\pm 0.50}$	$79.84_{\pm 0.80}$	$77.56_{\pm 2.56}$	$80.86_{\pm 0.16}$	$72.79_{\pm 1.54}$	$84.47_{\pm 1.69}$
DFR^{tr}_{tr} (Kirichenko et al., 2022)	✓	$91.25_{\pm 3.49}$	$83.55_{\pm 4.02}$	$46.70_{\pm 3.75}$	$28.53_{\pm 4.59}$	$62.51_{\pm 0.55}$	$12.31_{\pm 8.48}$	$85.51_{\pm 6.15}$
Ours	✓	$82.43_{\pm 0.58}$	$\mathbf{85.12}_{\pm 0.43}$	$\mathbf{81.65}_{\pm 0.35}$	$\mathbf{82.58}_{\pm 0.92}$	$\mathbf{82.94}_{\pm 0.23}$	$\mathbf{77.90}_{\pm 0.18}$	$84.29_{\pm 0.92}$

Table 2. GG, GC, CG, CC, and UNBIASED metrics (%) for the Multi-Color MNIST dataset. The first element of each of the four combinations {GG, GC, CG, CC} represents the bias type of `left-color`, while the second element represents the bias type of `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	✗	$100.0_{\pm 0.0}$	$\underline{96.5}_{\pm 1.2}$	$79.5_{\pm 2.5}$	$20.8_{\pm 1.1}$	$74.2_{\pm 1.1}$
LfF (Nam et al., 2020)	✗	$99.6_{\pm 0.5}$	$4.7_{\pm 0.5}$	$\mathbf{98.6}_{\pm 0.4}$	$5.1_{\pm 0.4}$	$52.0_{\pm 0.1}$
EIIL (Creager et al., 2021)	✗	$100.0_{\pm 0.0}$	$\mathbf{97.2}_{\pm 1.5}$	$70.8_{\pm 4.9}$	$10.9_{\pm 0.8}$	$69.7_{\pm 1.0}$
PGI (Ahmed et al., 2021)	✗	$98.6_{\pm 2.3}$	$82.6_{\pm 19.6}$	$26.6_{\pm 5.5}$	$9.5_{\pm 3.2}$	$54.3_{\pm 4.0}$
DebiAN (Li et al., 2022)	✗	$100.0_{\pm 0.0}$	$95.6_{\pm 0.8}$	$76.5_{\pm 0.7}$	$16.0_{\pm 1.8}$	$72.0_{\pm 0.8}$
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	✓	$100.0_{\pm 0.0}$	$90.0_{\pm 2.5}$	$\underline{83.4}_{\pm 2.1}$	$37.1_{\pm 2.8}$	$77.6_{\pm 1.0}$
GroupDRO (Sagawa et al., 2019)	✓	$98.0_{\pm 0.0}$	$87.2_{\pm 4.3}$	$77.3_{\pm 7.5}$	$\mathbf{52.3}_{\pm 2.6}$	$\underline{78.7}_{\pm 2.7}$
Ours	✓	$99.7_{\pm 0.6}$	$90.4_{\pm 3.4}$	$81.8_{\pm 4.0}$	$\underline{48.1}_{\pm 0.3}$	$\mathbf{80.0}_{\pm 2.0}$

Table 3. 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.

	Bias	Waterbirds		CelebA	
Method	label	WORST	INDIST	WORST	INDIST
ERM	✗	$63.7_{\pm 1.9}$	$97.0_{\pm 0.2}$	$47.8_{\pm 3.7}$	$94.9_{\pm 0.2}$
LfF (Nam et al., 2020)	✗	78.0	91.2	70.6	86.0
EIIL (Creager et al., 2021)	✗	$77.2_{\pm 1.0}$	$96.5_{\pm 0.2}$	$81.7_{\pm 0.8}$	$85.7_{\pm 0.1}$
JTT (Liu et al., 2021)	✗	$83.8_{\pm 1.2}$	$89.3_{\pm 0.7}$	$81.5_{\pm 1.7}$	$88.1_{\pm 0.3}$
LWBC (Kim et al., 2022)	✗	-	-	$85.5_{\pm 1.4}$	$88.9_{\pm 1.6}$
CNC (Zhang et al., 2022)	✗	$88.5_{\pm 0.3}$	$90.9_{\pm 0.1}$	$88.8_{\pm 0.9}$	$89.9_{\pm 0.5}$
Upweighting	✓	$88.0_{\pm 1.3}$	$95.1_{\pm 0.3}$	$83.3_{\pm 2.8}$	$92.9_{\pm 0.2}$
GroupDRO (Sagawa et al., 2019)	✓	$\underline{89.9}_{\pm 0.6}$	$92.0_{\pm 0.6}$	$88.9_{\pm 1.3}$	$93.9_{\pm 0.1}$
SUBG (Sagawa et al., 2020)	✓	$89.1_{\pm 1.1}$	-	$85.6_{\pm 2.3}$	-
SSA (Nam et al., 2022)	✓	$89.0_{\pm 0.6}$	$92.2_{\pm 0.9}$	$\mathbf{89.8}_{\pm 1.3}$	$92.8_{\pm 0.1}$
LISA (Yao et al., 2022)	✓	$89.2_{\pm 0.6}$	$91.8_{\pm 0.3}$	$\underline{89.3}_{\pm 1.1}$	$92.4_{\pm 0.4}$
DFR^{tr}_{tr} (Kirichenko et al., 2022)	✓	$90.2_{\pm 0.8}$	$97.0_{\pm 0.3}$	$80.7_{\pm 2.4}$	$90.6_{\pm 0.7}$
Ours	✓	$\mathbf{91.8}_{\pm 0.3}$	$95.6_{\pm 0.3}$	$\mathbf{89.8}_{\pm 1.3}$	$91.4_{\pm 1.2}$

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. We also report the average of these four metrics, denoted as UNBIASED.

3.2. Quantitative result

MultiCelebA. In Table 1, we present the results of our experiments evaluating the performance of various baselines and existing debiased training methods on the MultiCelebA dataset. One can observe how our method outperforms the baselines by a significant margin in terms of UNBIASED, GC, CG, CC, and WORST metrics. Our algorithm even achieves a moderate accuracy for the GG metric. This highlights how our algorithm successfully prevents performance degradation by simultaneously removing multiple spurious correlations. Interestingly, we observe that existing debiasing algorithms struggle with conflicts when attempting to remove different spurious correlations. We also observe that DFR (Kirichenko et al., 2022) achieves lower CC and CG metrics than ERM, suggesting that an ERM-based feature representation alone is insufficient to debias the training.

Multi-Color MNIST. In Table 2, we report the evaluation results for Multi-Color MNIST. Overall, our proposed method demonstrates the best performance along with GroupDRO. In particular, our algorithm exhibits the highest UNBIASED accuracy and the second-best CC accuracy.

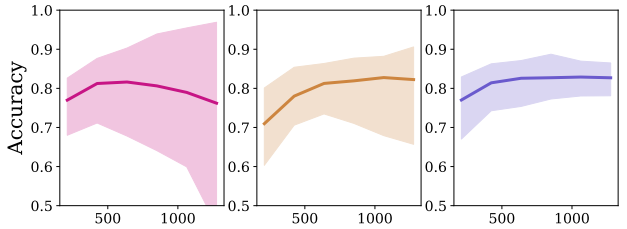
Single-bias datasets. Surprisingly, as shown in Table 3, our method achieves the best WORST accuracy on Waterbirds and CelebA, indicating that our method is effective not only for multi-bias settings but also for single-bias settings.

3.3. Ablation study

Comparison of strategies to choose $\boldsymbol{\alpha}$. We first conduct an ablation study to verify our strategy to adjust the group-scaling parameter $\boldsymbol{\alpha}$. In Table 4, we compare our strategy to choose $\boldsymbol{\alpha}$ with two alternatives: (i) using a fixed uniform group-scaling parameter, *i.e.*, $\boldsymbol{\alpha} = \frac{1}{N}\mathbf{1}$, (ii) minimizing

Table 4. Ablation studies for stratigies to choose group-scaling parameter $\boldsymbol{\alpha}$ on MultiCelebA: (i) fixing by $\frac{1}{N}\mathbf{1}$, (ii) minimizing $\boldsymbol{\alpha}^\top L(\theta)$, and (iii) minimizing $\hat{L}(\theta)$, *i.e.*, ours.

	GG	GC	CG	CC	UNBIASED
(i)	76.6	79.0	76.8	79.7	78.0
(ii)	75.1	78.1	75.9	79.0	77.0
(iii)	**82.3**	**84.7**	**81.9**	**82.7**	**82.9**



(i) Group by $\boldsymbol{b}$ (ii) Group by $\boldsymbol{b}, t$ (iii) Group by $\boldsymbol{g}$

Figure 2. Group-wise test accuracy of each grouping strategy: (i) group by bias attribute, (ii) group by both bias attribute and target class, and (iii) group by list of binary group labels. The line represents the UNBIASED performance, and the lower and upper bounds of the shaded regions indicate the minimum accuracy (*i.e.*, WORST) and maximum accuracy among each group.

$\boldsymbol{\alpha}^\top L(\theta)$, and (iii) our method to minimize $\hat{L}(\theta)$. One can observe how our strategy consistently outperforms other strategies, supporting the use of our strategy.

Intriguingly, one can also observe how learning $\boldsymbol{\alpha}$ as in (ii) leads to a worse performance compared to (i) that uses a fixed value of $\boldsymbol{\alpha}$. Our analysis is that with a learnable group scaling parameter based solely on the weight sum of group-wise losses led to worse performance in all five metrics compared to training without it. We found out such a performance degradation to happen is because optimizing $\boldsymbol{\alpha}$ solely on the weight sum of group-wise losses increase in the weight of a group that has a small training loss, while decreasing the weight of a group that has difficulty learning. This runs counter to the goal of debiased training and thus negatively impacts the overall performance of the model.

Comparison of grouping strategies. We next conduct an ablation study for verifying our grouping strategy. We compare ours with two other possible strategies: grouping samples with (i) the same bias attribute $\boldsymbol{b}$ and (ii) the same pair of bias attribute $\boldsymbol{b}$ and target class t. We report the group-wise test accuracies of each model are in Figure 2. Here, we represent the UNBIASED metric as a solid line and the range between the minimum and maximum accuracies of the test groups by shaded regions.

Figure 2 (i) shows that the test accuracy gap between groups widens as training progresses when using the bias attribute grouping. We hypothesize that this is due to class imbalance within the groups caused by training with biased dataset samples grouped 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 2 (ii), we applied the commonly used grouping by both target classes and bias attributes. Compared to the conventional grouping, our approach demonstrates a smaller performance gap between groups and higher worst group accuracy, as shown in Figure 2 (iii). Finally, we also report the performances of the grouping strategies in Appendix.

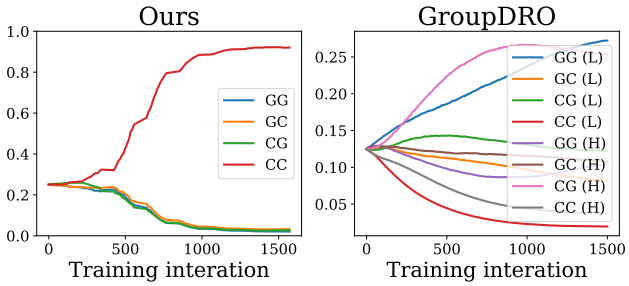


Figure 3. Learning curve of group scaling parameter $\boldsymbol{\alpha}$ of our method and GroupDRO on MultiCelebA. In case of GroupDRO, (H) and (L) denote `High-cheekbones` and `Low-cheekbones`, respectively.

Learning curve for group-scaling parameter $\boldsymbol{\alpha}$. We compare learning curve of group scaling parameter $\boldsymbol{\alpha}$ of our method with that of GroupDRO (Sagawa et al., 2019) on MultiCelebA, as illustrated in Figure 3. 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 groups, but later focuses more on the minority group. On the other hand, GroupDRO exhibits a decreasing weight trend for the minority groups (CC (L) and CC (H) in Figure 3). This trend occurs because the minority groups have lower training losses in the early stages of training, leading to lower weights in GroupDRO method. As a result, it tends to ignore minority groups and exacerbate the bias issue, and resulting in inferior performance compared to the upweigthing.

4. Conclusion

We have presented a novel debiased training method that addresses the challenges posed by multiple biases in training data, inspired by multi-task learning (MTL). We evaluated our approach on multiple benchmarks both in multi-bias and single-bias settings. The empirical results demonstrated that our proposed method achieves state-of-the-art performance in all benchmarks, surpassing existing debiased training methods and baselines. Since we propose a debiased training method inspired by MTL and highlight its potential, we hope to inspire more future works that use the advancements in MTL to benefit debiased training.

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.

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.

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

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.

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 ICML 2022: Workshop on Spurious Correlations, Invariance, and Stability*, 2022.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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, 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.

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, 2021.

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.

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.

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., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

A. MultiCelebA

To evaluate debiased training algorithms in multi-bias settings, prior work (Li et al., 2022) proposed Multi-Color MNIST benchmark that injects two biased attributes, *i.e.*, color of left background and color of right background, to the vanilla MNIST dataset (LeCun et al., 1998). Unfortunately, such a synthetic dataset is underwhelming as it is not sufficient to reflect the performance of the algorithms in the real world. To resolve this issue, we propose a new, natural image dataset, coined MultiCelebA, for evaluating debiased training algorithms under the presence of multiple biases.

MultiCelebA is curated using images of the CelebA dataset, which is a large-scale collection of face images each with 40 attribute annotations. Among the attributes, we chose `high-cheekbones` as the target attribute to predict from a face image. We further inspect the dataset and identified two bias attributes, `gender` and `age`, that hinder the training through spurious correlations with `high-cheekbones`. In particular, we observed that a deep neural network trained by ERM to predict `high-cheekbones` relies on the spurious correlation from the `gender` attribute and then the `age` attribute. To simulate challenging scenarios where training data are extremely biased, we set the ratio of bias-guiding samples for each bias type (either `gender` or `age`) to 95.3% so that only 0.22% of training samples are free from such spurious correlations. Example images and the number of samples for each attribute are presented in Fig. 4.

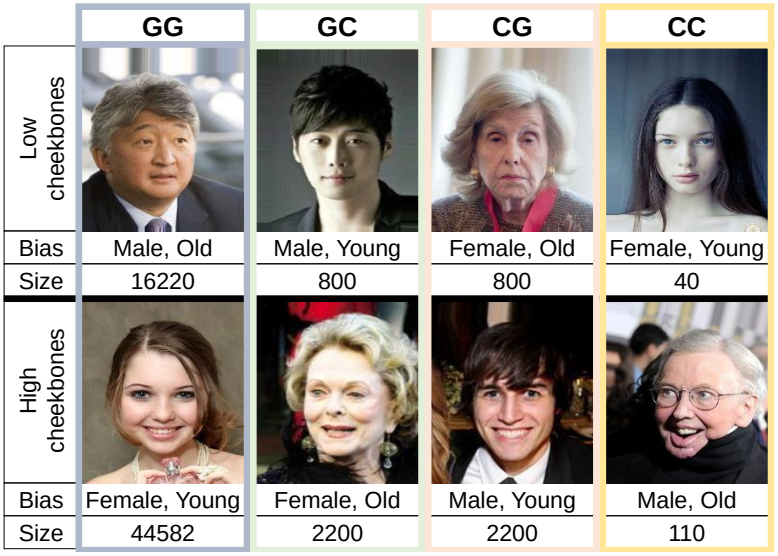


Figure 4. Configuration of MultiCelebA training set. The target class is `High-cheekbones`, and there are two bias types: `gender` and `age`. We provide an example of each attribute along with the number of samples for each attribute. The samples are grouped by colored borders, with each color representing a group for MTL training. The name of each group, from GG to CC, denotes whether if the sample in the group has a guiding attribute (G) or a conflicting attribute (C) for `gender` and `age`, in respective order.

B. Algorithm

Algorithm 1 Debiased training by MOO

while not converged **do**
  Let $\boldsymbol{\alpha} = \mathrm{SoftMax}(\bar{\boldsymbol{\alpha}})$.
  for $u \leftarrow 1$ to U **do**
    Update $\theta \leftarrow \theta - \eta_1 \boldsymbol{\alpha}^\top \nabla_\theta L(\theta)$.
  end for
  Let $\hat{L}(\theta) = \boldsymbol{\alpha}^\top L(\theta) + \lambda \left\| \boldsymbol{\alpha}^\top \nabla_\theta L(\theta) \right\|_2^2$.
  Update $\bar{\boldsymbol{\alpha}} \leftarrow \bar{\boldsymbol{\alpha}} - \eta_2 \nabla_{\bar{\boldsymbol{\alpha}}} \hat{L}(\theta)$.
  Update $\lambda \leftarrow \lambda + \eta_2 \nabla_\lambda \hat{L}(\theta)$.
end while

C. Experiments

C.1. Datasets

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

First, we mainly consider MultiCelebA as the dataset to evaluate debiased training algorithms. As introduced in Section A, 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`.

As the secondary multi-bias dataset, we consider Multi-Color MNIST dataset proposed by Li et al. (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.

Next, 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%.

Finally, we consider CelebA (Liu et al., 2015) as the single-bias dataset. It 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.

C.2. 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`. 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. We also report the average of these four metrics, denoted as UNBIASED.

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. (Sagawa et al., 2019). We also report WORST and INDIST metrics on MultiCelebA.

C.3. 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 four 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., 2021) 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, and (4) 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, (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., 2022) retrains the last layer of an ERM model using a balanced set obtained through undersampling.

C.4. Implementation details.

We conduct experiments using the following neural network architectures: a three-layered MLP for Multi-Color MNIST, ResNet18 for MultiCelebA, and ResNet50 for single-bias datasets. For the implementation of our method, we set the batch size to $\{512, 512, 128, 128\}$, learning rate η_1 to $\{2\text{e}{-4}, 2\text{e}{-2}, 1\text{e}{-3}, 2\text{e}{-3}\}$ and η_2 to $\{1\text{e}{-2}, 2\text{e}{-3}, 1\text{e}{-3}, 1\text{e}{-4}\}$, weight decay to $\{1, 1\text{e}{-4}, 1\text{e}{-1}, 1\text{e}{-5}\}$, the update frequency U to $\{10, 50, 5, 1\}$ with optimizer $\{$SGD, Adam, SGD, Adam$\}$, respectively for $\{$MultiCelebA, Multi-Color MNIST, Waterbirds, CelebA$\}$. The group scaling parameter $\boldsymbol{\alpha}$ is initialized to $1/N$ 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 tasks. We report the average and standard deviation of each metric calculated from three runs with different random seeds.

Training existing methods on multi-bias setting. When training a model using SUBG (Sagawa et al., 2020), group-DRO (Sagawa et al., 2019) and DFR (Kirichenko et al., 2022), 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., 2021), 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.

We conducted 'Upsampling' method in Table 1 and 2 by uniformly sampling from groups, which is upsampling the minority groups with replacement.

For the 'Upweighting' method, we calculate the weight of each group as the ratio between the total number of training samples and the number of samples in that group, as follows:

$$\text{group weight} = \frac{(\#\text{ of training samples})}{(\#\text{ of group samples})}$$

Hyperparameters. We tune all hyperparameters, including early stopping, for both our method and existing methods, based on highest WORST for MultiCelebA, Waterbirds and CelebA on validation set. For Multi-Color MNIST, we tune hyperparemters based on highest UNBIASED on test set, following the previous work (Li et al., 2022). We use a single GPU (RTX 3090) for training. The hyperparameter search spaces used in all experiments conducted in this paper are summarized in Table 5. Furthermore, the search space for the upweight value λ_{up} in JTT is 5, 10, 20, 30, 40, 50, 100. JTT (Liu et al., 2021) and DFR (Kirichenko et al., 2022) 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.

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}.

Table 5. The search spaces of hyperparameters.

Method	Search space
Learning rate η_1, η_2	$\{5\text{e}{-4}, 2\text{e}{-4}, 1\text{e}{-4}, 5\text{e}{-3}, 2\text{e}{-3}, 1\text{e}{-3}, 5\text{e}{-2}, 2\text{e}{-2}, 1\text{e}{-2}\}$
Weight decay	$\{0, 1\text{e}{-4}, 1\text{e}{-2}, 1\text{e}{-1}, 1\}$
Updating iteration U	$\{1, 5, 10, 50\}$

C.5. Additional 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 towards `gender` attributes if there is a significant difference between the G* and C* combinations, whereas a significant difference between the *G and *C combinations indicates bias towards `age` attributes.

Intriguingly, we observe that algorithms like JTT, DebiAN, DFR exhibit UNBIASED metric even lower than the vanilla ERM algorithm. Our hypothesis is that this performance degradation stems from conflicts between removal of different spurious correlations. To be specific, JTT (Liu et al., 2021) 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., 2022) achieves lower CC and CG metrics than ERM, suggesting that an ERM-based feature representation alone is insufficient to debias the training.

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 slightly biased towards the `gender` attribute. Surprisingly, Upweighting achieved the second-best performance in CC, UNBIASED, and WORST on MultiCelebA, surpassing all the existing debiased training methods.

C.6. Ablation study

Comparison of grouping strategies. We report the performances of the grouping strategies in Table 6. Here one can observe that our grouping strategy outperforms the others in all five metrics.

Table 6. Ablation studies on the grouping strategy on MultiCelebA: group by bias attribute $\boldsymbol{b}$, bias attribute and target class $(\boldsymbol{b}, t)$, and list of binary group labels $\boldsymbol{g}$.

Group by	GG	GC	CG	CC	UNBIASED
$\boldsymbol{b}$	75.3	78.2	75.7	78.8	77.0
$\boldsymbol{b}, t$	76.5	79.0	76.8	79.8	78.0
$\boldsymbol{g}$	**83.1**	**85.0**	**81.3**	**81.6**	**82.7**

Updating frequency U. We conducted experiments to examine how hyperparameter U affects the performance of our method. Table 7 presents the GG, GC, CG, CC and UNBIASED accuracies on the MultiCelebA dataset using five different values of U. To exclude 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. Additionally, the GC, CC, and UNBIASED of our proposed method outperformed existing methods and baselines regardless the value of U.

U	GG	GC	CG	CC	UNBIASED
1	83.7	85.7	81.2	81.1	82.9
5	83.0	85.3	81.5	81.8	82.9
10	83.1	85.0	81.3	81.6	82.7
20	82.7	85.4	80.9	81.8	82.7
30	80.8	84.3	82.4	84.4	83.0

Table 7. Ablation studies on the updating frequency U of the group scaling parameter $\boldsymbol{\alpha}$ on the MultiCelebA dataset.