

A MORE VISUALIZATIONS OF BIASED DISTRIBUTIONS

We plot the biased distributions of more existing benchmarks as follows:

CelebA. CelebA Liu et al. (2015) is a dataset for face recognition where each sample is labeled with 40 attributes, which has been adopted as a benchmark for debiasing methods. Following the experiment configuration suggested by Nam et al. [32], we focus on HeavyMakeup attributes that are spuriously correlated with Gender attributes, i.e., most of the CelebA images with heavy makeup are women. As a result, the biased model suffers from performance degradation when predicting males with heavy makeup and females without heavy makeup. Therefore, we use Heavy_Makeup as the target attribute and Male as a spurious attribute. The joint distribution between the Male and Heavy_Makeup attribute of the CelebA dataset is plotted in Figure 6a. It is clear that the biased distribution of CelebA aligns with that in other existing benchmarks, forming a "diagonal distribution".

WaterBirds. WaterBirds Liu et al. (2021) is a synthetic dataset with the task of classify images of birds as "waterbird" and "landbird", which is adopted as a benchmark for debiasing methods. The label of WaterBirds is spuriously correlated with the image background, i.e. Place attribute, which is either "land" or "water". The joint distribution between the Place and Bird attribute of the WaterBirds dataset is plotted in Figure 6b.

Additional visualization of the biased distribution within real-world datasets is also plotted as follows:

Adult. The Adult Becker & Kohavi (1996) dataset, also known as the "Census Income" dataset, is widely used for tasks such as income prediction and fairness analysis. Each sample is labeled with demographic and income-related attributes. The dataset has been adopted as a benchmark for debiasing methods, particularly focusing on the correlation between race and income. The joint distribution between Race and Income attributes of the Adult dataset is plotted in Figure 6c. It is clear that the biased distribution of Adult does not align with that of other existing benchmarks.

German. The German Hofmann (1994) dataset, also known as the "German Credit" dataset, is commonly used for credit risk analysis and fairness studies. Each sample is labeled with various attributes related to creditworthiness. The dataset serves as a benchmark for debiasing methods, emphasizing the correlation between age and creditworthiness. The joint distribution between Age and Creditworthiness attributes of the German dataset is plotted in Figure 6d. It is clear that the biased distribution of German does not align with that of other existing benchmarks.

B FINE-GRAINED EVALUATION FRAMEWORK

In this section, we elaborate on the proposed evaluation framework by mathematically and visually demonstrating the biased distribution within the biased distribution.

Assume a set of biased features $a_i^s \in B$ whose correlated class in the target attribute is defined by a function $g : y^s \rightarrow y^t$, which is an injection from the spurious to the target attribute. The bias magnitude of each biased feature is controlled by $corr_i = P(y^t = g(a_i^s) | y^s = a_i^s)$. Then, the empirical distribution of the biased train distribution satisfies the following equations.

For samples with biased feature a_i^s within B :

$$P(y^s = a_i^s, y^t = a^t) = \begin{cases} P(y^s = a_i^s) * corr_i & \text{if } g(a_i^s) = a^t, \\ \frac{P(y^s = a_i^s) * (1 - corr_i)}{|y^t| - 1} & \text{otherwise,} \end{cases}$$

For samples without biased features and a set of correlated classes $C = \{g(a_i^s) : a_i^s \in B\}$:

$$P(y^s = a^s, y^t = a^t) = \frac{P(y^t = a^t) - \sum_{a_i^s \in B} P(y^s = a_i^s, y^t = a^t)}{|y^s| - |B|}$$

Following the above equations, we further designed LMLP, HMLP, and HMHP biased distributions with the configurations in Table 4. The visualizations of the distributions when the target is a ten-class attribute are in Figure 7.

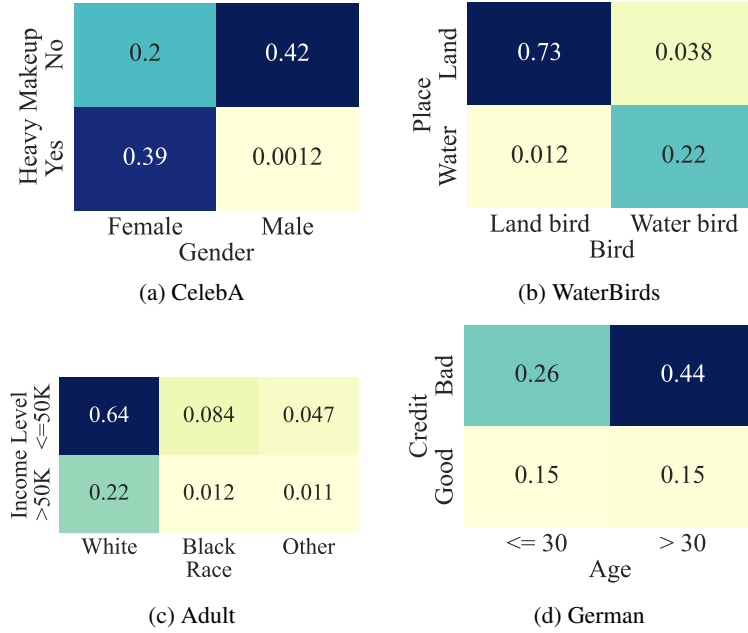


Figure 6: Visualization of the joint distribution for datasets, where the y-axis is the target attribute and the x-axis is the spurious attribute. Figure 6(a) and 6(b) visualize the distribution of existing benchmarks. Figure 6(c) and 6(d) visualize the distribution of real-world datasets. The biased distribution of existing benchmarks and real-world datasets is not alike.

Table 4: Configurations for biased distributions within the proposed evaluation framework

Distribution	$ y^t $	$ B $	$corr_i$
LMLP	10	10	0.5
HMLP	10	1	0.98
HMHP	10	10	0.98
Unbiased	10	0	0.1

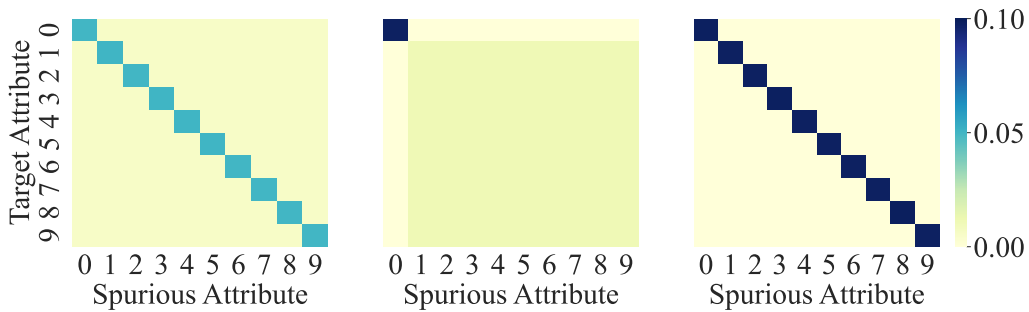


Figure 7: Visualization of biased distributions within the proposed evaluation framework under ten-class classification task. The left, middle, and right plots are visualizations for LMLP, HMLP, and HMHP distribution respectively.

C THEORETICAL PROOFS

C.1 PRELIMINARY

Consider a classification task on binary target attribute $y^t \sim \{-1, +1\}$ and a binary spurious attribute $y^s \sim \{-1, +1\}$. Let the marginal distribution of the target and spurious attribute to be $p_+^t = P(y^t = +1)$ and $p_+^s = P(y^s = +1)$. Then the joint distribution between y^t and y^s can be defined according to the conditional distribution of y^t given $y^s = +1$, i.e. $\tau_+ = P(y^t = +1|y^s = +1)$. Specifically, we can derive the probability of each subgroup in the distribution:

$$P(y^s = +1, y^t = +1) = p_+^s \cdot \tau_+, \quad (5)$$

$$P(y^s = +1, y^t = -1) = p_+^s (1 - \tau_+), \quad (6)$$

$$P(y^s = +1, y^t = -1) = p_+^t - p_+^s \cdot \tau_+, \quad (7)$$

$$P(y^s = -1, y^t = -1) = 1 - p_+^t - p_+^s (1 - \tau_+) \quad (8)$$

We assume that feature $y^s = +1$ and $y^s = -1$ is correlated $y^t = +1$ and $y^t = -1$ respectively, i.e. $\tau_+ > p_+^t$, in the following analysis.

C.2 PROOF OF PROPOSITION 1

Proposition 1 shows that high bias prevalence distribution assumes matched marginal distributions.

Proposition 1. *Assume feature $y^s = +1$ is biased. Then high bias prevalence distribution, i.e. feature $y^s = -1$ is biased as well, implying that the marginal distribution of y^t and y^s is matched, i.e. $p_+^t = p_+^s$. Specifically, as θ approaches to 1, the marginal distribution of y^s approaches to that of y^t , i.e. $\lim_{\theta \rightarrow 1} p_+^s = p_+^t$.*

Proof. We first derive the upper and lower bound of the p_+^s , and then we can prove the proposition with the squeeze theorem Stewart (2012).

According to the condition that both features in the spurious attribute are biased and the definition of biased feature in ref, we can have the following inequalities:

$$\rho_+ > \theta \cdot \rho_{max}^+ = \theta \cdot (1 - p_+^t), \quad (9)$$

$$\rho_- > \theta \cdot \rho_{max}^- = \theta \cdot p_+^t \quad (10)$$

where $0 < \theta \leq 1$ is the threshold.

We can also derive the simplified bias magnitude of feature $y^s = -1$ based on the conditional distribution, and find its relationship with ρ_+ :

$$\rho_- = \tau_- - p_-^t \quad (11)$$

$$= \frac{1 - p_+^t - p_+^s (1 - \tau_+)}{1 - p_+^s} - (1 - p_+^t) \quad (12)$$

$$= \frac{p_+^s (\tau_+ - p_+^t)}{1 - p_+^s} \quad (13)$$

$$= \frac{p_+^s}{1 - p_+^s} \rho_+ \quad (14)$$

We can then derive the lower bound of p_+^s with the above equation and inequalities:

$$\frac{p_+^s}{1 - p_+^s} (1 - p_+^t) \geq \frac{p_+^s}{1 - p_+^s} \rho_+ = \rho_- \geq \theta \cdot p_+^t \quad (15)$$

$$p_+^s \geq \frac{\theta \cdot p_+^t}{1 - p_+^t + \theta \cdot p_+^t} \geq \theta \cdot p_+^t = LB(\theta) \quad (16)$$

We can also derive the following equation and inequalities of τ_+ according to its definition.

$$\tau_+ = \frac{p_+^s \cdot P(y^s = +1|y^t = +1)}{p_+^s} \leq \frac{p_+^t}{p_+^s} \quad (17)$$

$$\tau_+ = p_+^t + \rho_+ \geq \theta(1 - p_+^t) + p_+^t \quad (18)$$

Then we can derive the upper bound of p_+^s :

$$\theta(1 - p_+^t) + p_+^t \leq \tau_+ \leq \frac{p_+^t}{p_+^s} \quad (19)$$

$$p_+^s \leq \frac{p_+^t}{\theta(1 - p_+^t) + p_+^t} = UB(\theta) \quad (20)$$

We then demonstrate the convergence of the $LB(\theta)$ and $UB(\theta)$ as $\theta \rightarrow 1$:

$$\lim_{\theta \rightarrow 1} LB(\theta) = \lim_{\theta \rightarrow 1} \theta \cdot p_+^t = p_+^t \quad (21)$$

$$\lim_{\theta \rightarrow 1} UB(\theta) = \lim_{\theta \rightarrow 1} \frac{p_+^t}{\theta(1 - p_+^t) + p_+^t} = p_+^t \quad (22)$$

Finally, we can prove the proposition according to the squeeze theorem Stewart (2012):

$$LB(\theta) \leq p_+^s \leq UB(\theta) \quad (23)$$

$$\lim_{\theta \rightarrow 1} p_+^s = \lim_{\theta \rightarrow 1} LB(\theta) = \lim_{\theta \rightarrow 1} UB(\theta) = p_+^t \quad (24)$$

C.3 PROOF OF PROPOSITION 2

Proposition 2 shows that high bias prevalence distribution implies uniform marginal distributions.

Proposition 2. *Given that the marginal distribution of y^s and y^t are matched and not uniform, i.e. $p = p_+^s = p_+^t < 0.5$. The bias magnitude of sparse feature, i.e. ρ_+^* , is monotone decreasing at p , with $\lim_{p \rightarrow 0+} \rho_+^* = -\log(1 - \phi_+)$. The bias magnitude of the dense feature, i.e. ρ_-^* , is monotone increasing at p , with $\lim_{p \rightarrow 0+} \rho_-^* = 0$.*

Proof. Given the distribution proposed in section C.1 and the condition $p = p_+^s = p_+^t < 0.5$, we further use $\phi_+ = \frac{\rho_+^*}{\rho_{max}^*}$ to express τ :

$$\tau_+ = p + \phi_+(1 - p) \quad (25)$$

$$\tau_- = 1 - p + \phi_+ \cdot p \quad (26)$$

We can then derive the bias magnitude of the sparse feature $y^s = +1$, given $p = p_+^s = p_+^t < 0.5$, and warp it with a function $t(p)$.

$$\rho_+^* = KL(P(y^t), P(y^t|y^s = +1)) \quad (27)$$

$$= p \cdot \log\left(\frac{p}{\tau_+}\right) + (1 - p) \cdot \log\left(\frac{1 - p}{1 - \tau_+}\right) \quad (28)$$

$$= p \cdot \log\left(\frac{p}{p + \phi_+(1 - p)}\right) + (1 - p) \cdot \log\left(\frac{1 - p}{1 - p - \phi_+(1 - p)}\right) \quad (29)$$

$$= p \cdot \log\left(\frac{p}{p + \phi_+(1 - p)}\right) + (1 - p) \cdot \log\left(\frac{1}{1 - \phi_+}\right) \quad (30)$$

$$= p \cdot \log\left(\frac{p(1 - \phi_+)}{p + \phi_+(1 - p)}\right) + \log\left(\frac{1}{1 - \phi_+}\right) = t(p) \quad (31)$$

We further derive the partial derivative of ρ_+^* on p as follows:

$$\frac{\partial t(p)}{\partial p} = p \cdot \log\left(\frac{p(1 - \phi_+)}{p + \phi_+(1 - p)}\right) + 1 - \frac{p(1 - \phi_+)}{p + \phi_+(1 - p)} \quad (32)$$

Here we apply substitution method to replace $\frac{p(1-\phi_+)}{p+\phi_+(1-p)}$ with x :

$$\frac{\partial t(p)}{\partial p} = f(x) = \log x - (x - 1) \quad (33)$$

$$0 < x = \frac{p(1-\phi_+)}{p+\phi_+(1-p)} \leq 1 \quad (34)$$

We then show that $f(x)$ is monotone increasing in the interval $0 < x \leq 1$ and the critical point is at $x = 1$.

$$f'(x) = \frac{1}{x} - 1 \geq 0 \quad (35)$$

$$f(1) = 0 \quad (36)$$

Thus, we have $f(x) < 0$ in the interval $0 < x \leq 1$, proving $\rho_+^* = t(p)$ to be monotone decreasing at p .

$$\frac{\partial \rho_+^*}{\partial p} = \frac{\partial t(p)}{\partial p} < 0 \quad (37)$$

Similarly, we can derive the bias magnitude of the dense feature $y^s = -1$, and see that it is just $t(1-p)$

$$\rho_-^* = KL(P(y^t), P(y^t|y^s = -1)) \quad (38)$$

$$= (1-p) \cdot \log\left(\frac{(1-p)(1-\phi_+)}{1-p+\phi_+ \cdot p}\right) + \log\left(\frac{1}{1-\phi_+}\right) \quad (39)$$

$$= t(1-p) \quad (40)$$

As a result, we can prove the monotonicity of ρ_-^* with the chain rule.

$$\frac{\partial \rho_-^*}{\partial p} = \frac{\partial t(1-p)}{\partial p} \quad (41)$$

$$= \frac{\partial t(1-p)}{\partial(1-p)} \cdot \frac{\partial(1-p)}{\partial p} \quad (42)$$

$$= -\frac{\partial t(1-p)}{\partial(1-p)} \quad (43)$$

$$= -\frac{\partial t(p)}{\partial p} > 0 \quad (44)$$

We can then derive the convergence of sparse feature bias magnitude ρ_+^* when p approaches 0 with L'Hôpital's Rule Stewart (2012).

$$\lim_{p \rightarrow 0^+} \rho_+^* = \lim_{p \rightarrow 0^+} t(p) \quad (45)$$

$$= \lim_{p \rightarrow 0^+} \left(p \cdot \log\left(\frac{p(1-\phi_+)}{p+\phi_+(1-p)}\right) + \log\left(\frac{1}{1-\phi_+}\right) \right) \quad (46)$$

$$= \lim_{p \rightarrow 0^+} (p \cdot \log(p)) + \lim_{p \rightarrow 0^+} \left(p \cdot \log\left(\frac{1-\phi_+}{p+\phi_+(1-p)}\right) + \log\left(\frac{1}{1-\phi_+}\right) \right) \quad (47)$$

$$= \lim_{p \rightarrow 0^+} \frac{\log(p)}{\frac{1}{p}} + \log\left(\frac{1}{1-\phi_+}\right) \quad (48)$$

$$= \lim_{p \rightarrow 0^+} \frac{(\log(p))'}{(\frac{1}{p})'} + \log\left(\frac{1}{1-\phi_+}\right) \quad (49)$$

$$= \lim_{p \rightarrow 0^+} \frac{\frac{1}{p}}{-\frac{1}{p^2}} + \log\left(\frac{1}{1-\phi_+}\right) \quad (50)$$

$$= \log\left(\frac{1}{1-\phi_+}\right) \quad (51)$$

Similarly, we can derive the convergence of dense feature bias magnitude ρ_-^* when p approaches to 0.

$$\lim_{p \rightarrow 0^+} \rho_-^* = \lim_{p \rightarrow 0^+} t(1-p) \quad (52)$$

$$= \lim_{p \rightarrow 1^-} (p \cdot \log(\frac{p(1-\phi_+)}{p+\phi_+(1-p)}) + \log(\frac{1}{1-\phi_+})) \quad (53)$$

$$= \log(1-\phi_+) + \log(\frac{1}{1-\phi_+}) \quad (54)$$

$$= 0 \quad (55)$$

D EXPERIMENT DETAILS

D.1 EVALUATION METRICS

Following previous works Nam et al. (2020); Lee et al. (2021); Kim et al. (2022); Lim et al. (2023); Zhao et al. (2023); Lee et al. (2023), we use the accuracy of BC samples and the average accuracy on balanced test set as our main metrics. As a complement, we also present the accuracy of BN and BA samples when analyzing the performance of methods. Formally, we categorize samples according to the attributes (y^s, y^t) and a function $g: y^s \rightarrow y^t$ that maps the biased features to its correlated class.

$$BA = \{i | y^s[i] \in B, y^t[i] = g(y^s[i])\} \quad (56)$$

$$BC = \{i | y^s[i] \in B, y^t[i] \neq g(y^s[i])\} \quad (57)$$

$$BN = \{i | y^s[i] \notin B\} \quad (58)$$

where $y^s[i]$ and $y^t[i]$ the attribute value of sample i , and $B = \{a | \rho_a^* > \theta\}$ is the set of biased features.

D.2 DATASETS

Colored MNIST Reddy et al. (2021). We construct the Colored MNIST dataset based on the MNIST Lecun et al. (1998) dataset and set the background color as the bias attribute. Different from Colored MNIST used in previous work that simply correlates each of the 10 digits with a distinct color, where the strength of the correlation is controlled by setting the number of bias-aligned samples to $\{0.95\%, 0.98\%, 0.99\%, 0.995\%\}$, we proposed a more fine-grained generation process that is capable of various biased distributions, including LMLP, HMLP, HMHP. See Appendix B for more details.

Corrupted CIFAR10 Nam et al. (2020). We construct the Corrupted CIFAR10 dataset based on the CIFAR10 Krizhevsky (2009) dataset and set the corruption as the bias attribute. Different from Corrupted CIFAR10 used in previous work that simply correlates each of the 10 objects with a distinct corruption, where the strength of the correlation is controlled by setting the number of bias-aligned samples to $\{0.95\%, 0.98\%, 0.99\%, 0.995\%\}$, we proposed a more fine-grained generation process that is capable of various biased distributions, including LMLP, HMLP, HMHP. See Appendix B for more details.

BAR Nam et al. (2020). Biased Action Recognition (BAR) is a real-world dataset that contains spurious correlations between six human action classes and six place attributes. Following Nam et al. (2020), the ratio of bias-conflicting samples in the training set was set to 5%, and the test set consisted of only bias-conflicting samples.

NICO Kim et al. (2022) NICO is a real-world dataset for simulating out-of-distribution image classification scenarios. Following the setting used by Wang et al. (2021), we use an animal subset of NICO, which is labeled with 10 object and 10 context classes for evaluating the debiasing methods. The training set consists of 7 context classes per object class and they are long-tailed distributed (e.g., dog images are more frequently coupled with the ‘on grass’ context than any of the other 6 contexts). The validation and test sets consist of 7 seen context classes and 3 unseen context classes per object class. We verify the ability of debiasing a model from object-context correlations through evaluation on NICO.

WaterBirds Sagawa* et al. (2020). The task is to classify images of birds as “waterbird” or “landbird”, and the label is spuriously correlated with the image background, which is either “land” or “water”.

D.3 BASELINES

LfF. Learning from Failure (LfF) Nam et al. (2020) is a debiasing technique that addresses the issue of models learning from spurious correlations present in biased datasets. The method involves training two neural networks: one biased network that amplifies the bias by focusing on easily learnable spurious correlations, and one debiased network that emphasizes samples the biased network misclassifies. This dual-training scheme enables the debiased network to focus on more meaningful features that generalize better across various datasets.

DisEnt. The DisEnt Lee et al. (2021) method enhances debiasing by using disentangled feature augmentation. It identifies intrinsic and spurious attributes within data and generates new samples by swapping these attributes among the training data. This approach significantly diversifies the training set with bias-conflicting samples, which are crucial for effective debiasing. By training models with these augmented samples, DisEnt achieves better generalization and robustness against biases in various datasets.

BE. BiasEnsemble (BE) Lee et al. (2023) is a recent advancement in debiasing techniques that emphasizes the importance of amplifying biases to improve the training of debiased models. BE involves pretraining multiple biased models with different initializations to capture diverse visual attributes associated with biases. By filtering out bias-conflicting samples using these pre-trained models, BE constructs a refined bias-amplified dataset for training the biased network. This method ensures the biased model is highly focused on bias attributes, thereby enhancing the overall debiasing performance of the subsequent debiased model.

D.4 IMPLEMENTATION DETAILS

Reproducibility. To ensure the statistical robustness and reproducibility of the result in this work, we repeat each experiment within this work 3 times with consistent random seeds [0, 1, 2]. All results are the average of the three independent runs.

Architecture. Following Nam et al. (2020); Lee et al. (2021), we use a multi-layer perceptron (MLP) which consists of three hidden layers for Colored MNIST. For the Corrupted CIFAR10 dataset, we train ResNet18 He et al. (2016) with random initialization.

Training hyper-parameters. We set the learning rate as 0.001, batch size as 256, momentum as 0.9, and number of steps as 25000. We used the default values of hyper-parameters reported in the original papers for the baseline models.

Data augmentation. The image sizes are 28×28 for Colored MNIST and 224×224 for the rest of the datasets. For Colored MNIST, we do not apply additional data augmentation techniques. For Corrupted CIFAR10, we apply random crop and horizontal flip transformations. Also, images are normalized along each channel (3, H, W) with the mean of (0.4914, 0.4822, 0.4465) and standard deviation of (0.2023, 0.1994, 0.2010).

Training device. We conducted all experiments on a workstation with an Intel(R) Xeon(R) Gold 5220R CPU at 2.20GHz, 256 G memory, and 4 NVIDIA GeForce RTX 3090 GPUs. Note that only a single GPU is used for a single task.

D.5 APPLYING DiD TO DBAM METHODS

As aforementioned in the main paper, when applying our method to the existing DBAM methods Nam et al. (2020); Lee et al. (2021; 2023), we do not modify the training procedure of the debiased model M_d . For both methods, we train the biased model M_b with target feature destroyed data.

This is done by simply adding a feature destructive data transformation during data processing, with minimal computational overhead.

Note, for BE Lee et al. (2023), such feature destructive data transformation is not applied when training the bias-conflicting detectors.

E ADDITIONAL EMPIRICAL RESULTS

E.1 DETAILED RESULTS AND EXPLANATIONS OF THE MAIN EXPERIMENTS

The main results in the main paper are presented in the form of performance gain and only contain results of BC accuracy and average accuracy on the unbiased test set, here we present the results in their original form, together with error bars, detailed results of accuracies for BA and BN samples of each dataset as well. Results on the Colored MNIST and Corrupted CIFAR10 datasets can be found in Table 5 and Table 6, respectively. It shows that combining DiD not only boosts the performance of existing DBAM methods but also achieves the best performances.

The performance generally varies between different datasets, different types of biased distribution, and algorithms with and without BiasEnsemble, e.g. between LfF and BE LfF. Firstly, the inconsistency between datasets is likely to depend on how thoroughly the target feature is destroyed within the dataset. The target features of Colored MNIST, i.e. digits, are destroyed more completely by patch shuffling, for shape is the only feature within digits. In comparison, the target feature of Corrupted CIFAR10 is more complicated (including shape, texture, color, etc.), and thus can not be thoroughly destroyed by patch shuffling, causing relatively lower performance gain. Secondly, the performance inconsistency between different biased distributions is due to the reliance of existing DBAM methods on the high bias prevalence assumption for bias capturing as discussed in section 4.2. Specifically, as the bias prevalence of the training distribution becomes higher, better bias capture can be achieved by existing DBAM even without our method, thus making our improvement on the performance less significant. This conclusion is supported by our experimental results shown in Figure 5. As for the performance inconsistency between algorithms with and without BiasEnsemble, it is due to the fact that BiasEnsemble is also a method targeted to enhance the bias capture procedure of the debiasing framework. As we can see that BiasEnsemble is much more robust to the change in the bias magnitude and prevalence from Table 1. In other words, certain overlap between the goals of BiasEnsemble and our method resulted in smaller improvement of our method on BiasEnsemble-based baselines.

E.2 HYPER-PARAMETER SENSITIVITY

As shown in Table 7, we examine three feature destruction methods: pixel-shuffling, patch-shuffling, and center occlusion, to destroy object shapes. We observed that patch-shuffle with patch-size 8 exhibits the best performance on Corrupted CIFAR10 which is of size 32x32.

F RELATED WORKS

Model Bias. The tendency of machine learning models to learn and predict according to spurious Arjovsky et al. (2020) or shortcut Geirhos et al. (2020) features instead of intrinsic features, i.e. model bias, is found in a variety of domains Heuer et al. (2016); Tang et al. (2021); Gururangan et al. (2018); McCoy et al. (2019); Sagawa* et al. (2020) and is of interest from both a scientific and practical perspective. For example, visual recognition models may overly rely on the background of the picture rather than the targeted foreground object during prediction. One subtopic of model bias is model fairness, which generally refers to the issue that social biases are captured by models Hort et al. (2021), where the spurious features are usually human-related and annotated, such as gender, race, and age Mattu et al. (2016); Hofmann (1994;?).

Data Bias: spurious correlation. Generally, spurious correlation refers to the phenomenon that two distinct concepts are statistically correlated within the training distribution, though there is no causal relationship between them, e.g. background and foreground object Chu et al. (2024). The spurious correlation is a vital aspect of understanding how machine learning models learn and generalize Arjovsky et al. (2020). Specifically, studies on distribution shift Wiles et al. (2022) claim

Table 5: Results on Colored MNIST dataset show that combining DiD not only boosts the performance of existing DBAM methods but also achieves the best performances. The accuracy of BN samples is marked as '-' in LMLP and HMHP distribution for there is no BN sample within the dataset according to our evaluation setting in Appendix D.

Distr.	Algorithm	Accuracy			
		BA acc	BC acc	BN acc	Avg acc
LMLP	ERM	97.73 \pm 0.09	91.13 \pm 0.17	-	91.73 \pm 0.16
	LfF	80.25 \pm 4.86	68.41 \pm 2.01	-	69.74 \pm 2.41
	+ DiD	92.16 \pm 0.35	91.03 \pm 0.15	-	91.15 \pm 0.17
	BE LfF	82.95 \pm 1.68	83.60 \pm 0.85	-	83.53 \pm 0.75
	+ DiD	93.49 \pm 0.81	89.25 \pm 0.64	-	89.67 \pm 0.54
	DisEnt	84.45 \pm 1.72	73.87 \pm 2.52	-	74.93 \pm 2.44
	+ DiD	94.03 \pm 0.66	91.09 \pm 0.24	-	91.38 \pm 0.28
	BE DisEnt	80.18 \pm 1.94	81.07 \pm 2.50	-	80.98 \pm 2.29
	+ DiD	91.89 \pm 0.26	89.80 \pm 0.97	-	90.01 \pm 0.89
HMLP	ERM	99.32 \pm 0.34	85.25 \pm 1.62	90.30 \pm 0.56	89.82 \pm 0.70
	LfF	87.76 \pm 4.12	57.98 \pm 3.58	63.72 \pm 3.22	63.35 \pm 3.02
	+ DiD	82.99 \pm 5.08	90.54 \pm 0.74	89.04 \pm 0.84	89.12 \pm 0.77
	BE LfF	57.65 \pm 32.14	80.02 \pm 1.10	82.84 \pm 1.68	82.33 \pm 1.93
	+ DiD	63.95 \pm 15.64	89.11 \pm 1.29	87.28 \pm 1.54	87.22 \pm 1.58
	DisEnt	77.55 \pm 7.93	66.52 \pm 8.75	72.69 \pm 5.91	72.18 \pm 6.05
	+ DiD	88.78 \pm 7.24	88.52 \pm 1.47	89.04 \pm 1.13	88.99 \pm 1.16
	BE DisEnt	41.84 \pm 6.21	77.59 \pm 0.69	80.87 \pm 1.78	80.19 \pm 1.71
	+ DiD	31.97 \pm 7.08	89.33 \pm 1.07	85.88 \pm 0.86	85.66 \pm 0.89
HMHP	ERM	99.57 \pm 0.07	48.54 \pm 1.22	-	53.38 \pm 1.10
	LfF	57.16 \pm 8.27	65.62 \pm 2.87	-	64.59 \pm 3.31
	+ DiD	77.84 \pm 2.49	66.91 \pm 1.73	-	68.00 \pm 1.80
	BE LfF	73.61 \pm 1.03	66.90 \pm 0.43	-	67.57 \pm 0.47
	+ DiD	85.65 \pm 2.53	66.37 \pm 2.54	-	68.30 \pm 2.50
	DisEnt	59.89 \pm 4.19	68.29 \pm 1.43	-	67.45 \pm 1.28
	+ DiD	83.65 \pm 0.13	69.05 \pm 0.38	-	70.51 \pm 0.33
	BE DisEnt	77.74 \pm 2.51	67.51 \pm 1.33	-	68.53 \pm 1.45
	+ DiD	84.62 \pm 1.16	69.50 \pm 1.23	-	71.01 \pm 1.08

that spurious correlation is one of the major types of distribution shift in the real world, and thus an important distribution shift that a reliable model should be robust to. Furthermore, studies on fairness and bias Mehrabi et al. (2021) have demonstrated the pernicious impact of spurious correlation in classification Geirhos et al. (2019), conversation Beery et al. (2020), and image captioning Tang et al. (2021). However, despite its broad impact, spurious correlation is generally used as a vague concept in previous works and lacks a proper definition and deeper understanding of it. This is also the major motivation of this work.

Debiasing without bias supervision. In this work, we focus only on debiasing methods that do not require bias information, i.e. without annotation on the spurious attribute, for it is more practical. Existing work Nam et al. (2020); Lee et al. (2021); Kim et al. (2022); Hwang et al. (2022); Lim et al. (2023); Zhao et al. (2023); Lee et al. (2023); Park et al. (2024) in the area generally involve a biased auxiliary model to capture biases within the training data, according to which the debiased is trained with various techniques. We call such paradigm debiasing with biased auxiliary model (DBAM). Specifically, Nam et al. (2020) is the first work that follows the DBAM paradigm, proposing to use

Table 6: Results on Corrupted CIFAR10 dataset show that combining DiD not only boosts the performance of existing DBAM methods but also achieves the best performances. The accuracy of BN samples is marked as '-' in LMLP and HMHP distribution for there is no BN sample within the dataset according to our evaluation setting in Appendix D.

Distr.	Algorithm	Accuracy			
		BA acc	BC acc	BN acc	Avg acc
LMLP	ERM	80.40 \pm 0.81	62.50 \pm 0.15	-	64.29 \pm 0.06
	LfF	59.13 \pm 0.68	55.03 \pm 0.04	-	55.44 \pm 0.09
	+ DiD	69.47 \pm 0.96	62.04 \pm 0.21	-	62.78 \pm 0.10
	BE LfF	70.87 \pm 1.30	52.10 \pm 0.30	-	53.98 \pm 0.40
	+ DiD	63.23 \pm 2.10	53.21 \pm 0.20	-	54.21 \pm 0.38
	DisEnt	61.58 \pm 0.57	55.45 \pm 0.23	-	56.06 \pm 0.17
	+ DiD	72.23 \pm 0.74	60.84 \pm 0.40	-	61.98 \pm 0.30
	BE DisEnt	62.73 \pm 0.61	56.59 \pm 0.08	-	57.20 \pm 0.13
	+ DiD	65.98 \pm 0.40	60.92 \pm 0.20	-	61.42 \pm 0.21
	ERM	84.67 \pm 0.64	55.85 \pm 0.17	65.75 \pm 0.00	65.05 \pm 0.13
	LfF	73.33 \pm 1.67	47.70 \pm 0.58	54.58 \pm 0.49	54.15 \pm 0.41
	+ DiD	78.67 \pm 2.14	54.81 \pm 2.26	63.71 \pm 2.69	63.06 \pm 2.63
HMLP	BE LfF	70.33 \pm 2.19	50.96 \pm 2.35	54.14 \pm 0.25	54.02 \pm 0.36
	+ DiD	68.80 \pm 0.88	50.20 \pm 0.79	54.39 \pm 0.18	54.15 \pm 0.15
	DisEnt	61.67 \pm 1.67	52.48 \pm 0.56	54.65 \pm 0.56	54.53 \pm 0.49
	+ DiD	73.67 \pm 2.64	55.26 \pm 0.93	62.11 \pm 0.17	61.61 \pm 0.13
	BE DisEnt	75.33 \pm 5.21	49.15 \pm 1.54	56.86 \pm 0.30	56.35 \pm 0.35
	+ DiD	78.40 \pm 1.00	54.09 \pm 1.07	62.05 \pm 0.34	61.50 \pm 0.38
	ERM	89.97 \pm 0.34	29.37 \pm 0.30	-	35.43 \pm 0.24
	LfF	72.70 \pm 0.81	35.30 \pm 0.33	-	39.04 \pm 0.33
HMHP	+ DiD	82.07 \pm 1.09	37.05 \pm 0.31	-	41.55 \pm 0.19
	BE LfF	82.73 \pm 0.92	31.48 \pm 0.82	-	36.61 \pm 0.65
	+ DiD	78.30 \pm 0.47	32.90 \pm 1.79	-	37.44 \pm 1.61
	DisEnt	70.77 \pm 2.27	36.04 \pm 0.62	-	39.51 \pm 0.36
	+ DiD	76.60 \pm 0.70	39.05 \pm 0.35	-	42.80 \pm 0.25
	BE DisEnt	78.60 \pm 1.56	34.20 \pm 0.43	-	38.64 \pm 0.38
	+ DiD	78.70 \pm 1.47	37.72 \pm 0.96	-	41.82 \pm 0.91

GCE for bias capture, and the loss-based sample re-weighting scheme to train the debiased model. Lee et al. (2021) further proposed a feature augmentation technique to further utilize the captured bias, enhancing the BC samples. Hwang et al. (2022) proposed to augment biased data identified according to the biased auxiliary model by applying mixup Zhang et al. (2018) to contradicting pairs. Lim et al. (2023) proposed to conduct adversarial attacks on the biased auxiliary model to augment BC samples aiming to increase the diversity of BC samples. Lee et al. (2023) proposed to first filter out BC samples before training the biased auxiliary model aiming to enhance the bias capture process of the biased model. Liu et al. (2021) regard the samples misclassified by the biased auxiliary model as BC samples and emphasize them during training of the debiased model. Recently, Park et al. (2024) proposed to provide models with explicit spatial guidance that indicates the region of intrinsic features according to a biased auxiliary model. Kim et al. (2021) create images without bias attributes using an image-to-image translation model Park et al. (2020) built upon a biased auxiliary model. A recent pair-wise debiasing method χ^2 model Zhang et al. (2023a) based on biased auxiliary models encourages the debiased model to retain intra-class compactness using samples generated via feature-level interpolation between BC and BA samples.

Table 7: We experiment with three feature destruction methods with various hyper-parameters on HMLP distributed dataset with LfF.

T_{fd}	param	BC	Avg
N/A	N/A	47.70 ± 3.58	54.15 ± 3.02
pixel-shuffle	1	51.44 ± 1.01	55.43 ± 0.20
patch-shuffle	2	51.07 ± 0.48	55.29 ± 0.27
	4	49.41 ± 0.26	55.40 ± 0.26
	8	54.81 ± 0.74	63.06 ± 0.77
	16	49.74 ± 1.10	53.69 ± 0.31
center-occlusion	8	45.19 ± 1.41	51.61 ± 1.31
	16	47.26 ± 0.54	50.94 ± 0.59
	24	49.00 ± 0.80	52.60 ± 0.55
	32	52.44 ± 0.87	55.76 ± 0.16

G LIMITATIONS AND FUTURE WORK

We uncover the insufficiency of existing debiasing benchmarks theoretically and empirically, highlighting the importance of debiasing on real-world biases. We further proposed a feature-destruction-based method that focuses on DBAM methods. However, there are still a few limitations of this work:

- While DBAM methods are the predominant works in debiasing, there are also other lines of work such as data generation methods. Thus, one limitation is that We have not evaluated such methods with our proposed evaluation framework which might also yield some interesting insights on debiasing.
- Another limitation is that, though we have already seen the potential of target feature destruction, whether it can be applied to other lines of work remains to be studied.
- As shown in section E, while our proposed approach effectively improves the performance of existing DBAM methods on all biased distributions from the real world, the performance is still far from satisfactory, which remains to be further improved in future works.

We see potential within those limitations and leave them for future research.

H BORDER IMPACT

From a technical standpoint, our research provides a comprehensive framework for analyzing and mitigating biases in datasets. The proposed fine-grained analysis framework and evaluation benchmarks offer a new perspective on how biases manifest in real-world data and how existing debiasing methods can be improved. Our approach, which involves the destruction of target features during bias capture, demonstrates significant improvements in handling real-world biases, as evidenced by our extensive experimental results.

By advancing the understanding of dataset biases and improving the performance of debiasing methods, our research contributes to the development of more robust and generalizable AI models. This is particularly relevant in an era where AI systems are increasingly deployed in dynamic and diverse environments, necessitating models that can adapt and maintain high performance across different contexts and populations.