TOWARDS REAL WORLD DEBIASING: A FINE-GRAINED ANALYSIS ON SPURIOUS CORRELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spurious correlations in training data significantly hinder the generalization capability of machine learning models when faced with distribution shifts in real-world scenarios. To tackle the problem, numerous debiasing approaches have been proposed and benchmarked on datasets intentionally designed with severe biases. However, it remains to be asked: *1. Do existing benchmarks really capture biases in the real world? 2. Can existing debiasing methods handle biases in the real world?* To answer the questions, we revisit biased distributions in existing benchmarks and real-world datasets, and propose a fine-grained framework for analyzing dataset bias by disentangling it into the magnitude and prevalence of bias. We observe and theoretically demonstrate that existing benchmarks poorly represent real-world biases. We further introduce two novel biased distributions to bridge this gap, forming a systematic evaluation framework for real-world debiasing. With the evaluation framework, we focus on the practical setting of debiasing w/o bias supervision and find existing methods incapable of handling real-world biases. Through in-depth analysis, we propose a simple yet effective approach that can be easily applied to existing debiasing methods, named Debias in Destruction (DiD). Empirical results on real-world datasets in both image and language modalities demonstrate the superiority of DiD, improving the performance of existing methods on all types of biases within the proposed evaluation framework.

1 INTRODUCTION

With the rapid development of machine learning, machine learning systems are increasingly deployed in high-stakes applications such as autonomous driving (Janai et al., 2021) and medical diagnosis (Ibrahim & Abdulazeez, 2021), where incorrect decisions may cause severe consequences. As a result, the robustness to distribution shift is crucial in building trustworthy machine learning systems. One of the major reasons why machine learning models fail to generalize to shifted distributions in the real world (Koh et al., 2021; Chu et al., 2023; Xue et al., 2023) is because the existence of spurious correlation in training data (Wiles et al., 2022). Spurious correlation refers to the phenomenon that two distinct concepts are statistically correlated in the training distribution, yet uncorrelated in the test distribution for there is no causal relationship between them (Yao et al., 2021; Chu & Li, 2023). For example, rock wall background may be correlated with the sport climbing in the training data, but they are not causally related and climbing can be indoors or on ice as well (Lee et al., 2021; Chu et al., 2021; 2020). Furthermore, such spurious correlations within the data tend to be captured during training (Nam et al., 2020), resulting in a biased model that fails to generalize to shifted distributions. In this work, we refer to spurious correlation and bias in datasets interchangeably.

To tackle the problem, various debiasing methods (Bahng et al., 2020; Nam et al., 2020; Kim et al., 2021; Lee et al., 2021; Kim et al., 2022; Liu et al., 2021; Lim et al., 2023; Zhao et al., 2023) have been proposed in recent years. And the effectiveness of the methods is benchmarked with synthetic (Reddy et al., 2021; Nam et al., 2020; Liu et al., 2021) and semi-synthetic (Lee et al., 2021; Nam et al., 2020; Lim et al., 2023) (referred as "real-world dataset" in previous works) datasets designed to be severely biased. However, while these benchmarks are indeed biased, they are rough and lack thorough consideration of how data is truly biased in the real world. This raises two questions:

1. *Do existing benchmarks really capture biases in the real world?*
2. *Can existing debiasing methods handle biases in the real world?*

To answer the first question, we revisit the biased distribution in existing benchmarks and real-world datasets, and propose a fine-grained framework for dataset bias analysis. Inspired by the framework proposed by Wiles et al. (2022), which assumes the data is composed of some set of attributes, we further claim that analysis of dataset bias should be conducted on the more fine-grained feature (or value) level rather than attribute level, according to our observation on real-world biases. From the claim, we further propose our fine-grained framework that disentangles dataset bias into the magnitude of bias and the prevalence of bias, where the magnitude of bias generally measures how predictive (or biased) features are on the target task and the prevalence of bias generally measures how many samples in the data contain any biased feature. With our framework, we observe that the magnitude and prevalence of real-world biases are both low, in contrast with high magnitude and high prevalence biases in existing benchmarks. In section 3, we theoretically show that two strong assumptions are implicitly held by existing high bias prevalence benchmarks, which further validates our observation that real-world biases are low in bias prevalence.

As for the second question, *due to the complexity of real-world biases, debiasing methods should be capable of handling various types of biased distributions rather than solely on the narrowed distribution in existing benchmarks*. Thus, we introduce two new biased distributions inspired by real-world applications as a complement to the biased distribution in existing benchmarks, forming a systematic evaluation framework for real-world debiasing. We focus on debiasing methods w/o bias supervision(Nam et al., 2020; Lee et al., 2021; Liu et al., 2021; Kim et al., 2022; Lim et al., 2023; Zhao et al., 2023; Lee et al., 2023), which is more practical as bias feature is expensive to annotate and sometimes even hard to notice (Li & Xu, 2021). Those methods generally involves a biased auxiliary model to capture the bias, along with techniques to learn a debiased model with the captured bias. (See the related work section in Appendix F for details) We refer to such paradigm as debiasing with biased auxiliary model (DBAM). Our empirical results show that existing methods fail to handle real-world biases under various biased distributions.

We further conducted an in-depth analysis of the DBAM paradigm and found that the effectiveness of existing methods is reliant on the high bias prevalence of existing benchmarks and thus fails to handle real-world datasets with low bias prevalence. Finally, based on our analysis, we introduce a simple yet effective enhancement to the DBAM paradigm. Experiments on real-world datasets in both image and language modalities show that our approach significantly boosts the capability of existing methods in handling real-world biases, improving their performances on both high and low bias prevalence datasets. To sum up, this work makes the following contributions:

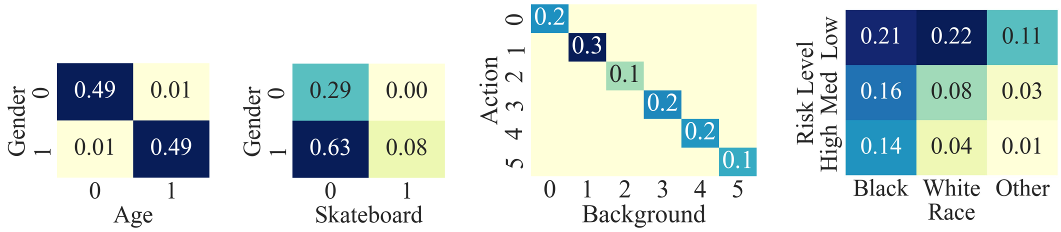
- **Fine-grained analysis and evaluation framework.** We propose a fine-grained framework for analyzing bias in datasets. Based on the framework we further introduce a systematic evaluation framework inspired by real-world applications for real-world debiasing.
- **Theoretical insight.** We derived the hidden strong assumptions held by existing benchmarks, and testify our observation that real-world biases are low in bias prevalence.
- **Principled approach.** A principled approach is proposed based on our analysis, which can be easily applied to not only the DBAM paradigm but bias detection methods as well.
- **Empirical validation.** Extensive empirical results on multiple real-world datasets, distributions, and modalities not only validate our analysis but also demonstrate the superiority of the proposed approach.

2 A Fine-grained Analysis on Bias in Datasets

In this section, we first revisit the biases in existing debiasing benchmarks and biases in the real world. Then, we propose a new framework for assessing dataset bias. Based on the framework we show how existing benchmarks fail to represent bias in real-world conditions.

2.1 Revisiting spurious correlation in datasets

Bias in existing benchmarks. In the area of spurious correlation debiasing, multiple synthetic (Reddy et al., 2021; Nam et al., 2020; Liu et al., 2021) and semi-synthetic datasets (Lee et al., 2021; Nam et al., 2020; Lim et al., 2023) have been adopted to benchmark the effectiveness of the debiasing methods. Generally, those synthetic datasets first select a target attribute as the learning objective (Liu



(a) BFFHQ (b) MS COCO (c) BAR (d) COMPAS

Figure 1: Visualization of the joint distribution for datasets, where the y-axis is the target attribute and the x-axis is the spurious attribute. Figure 1(a) and 1(c) visualise the distribution of existing benchmarks. Figure 1(b) and 1(d) visualize the distribution of real-world datasets. The biased distributions of existing benchmarks and real-world datasets are not alike.

et al., 2021), e.g. object, and another spurious attribute that could potentially cause the learned model to be biased, e.g. background. Then, certain sub-groups jointly defined by the target and spurious attributes, e.g. water birds with water background, are emphasized, i.e. synthesized or sampled from real-world datasets with much higher probability (usually above 95%) than the others in the biased dataset construction process, causing the corresponding spurious feature and target feature to be spuriously correlated, e.g. water background correlated with water bird (Liu et al., 2021). Specifically, one such dominating subgroup is selected for every possible value of the spurious attribute, forming a ”diagonal distribution”, as shown in Figure 1(a) and 1(c). However, it is critical to examine whether this pattern truly aligns with the complexities of real-world biases.

Bias in the real world. We further investigated biases from the real world. COCO (Lin et al., 2015) dataset is a large-scale dataset collected from the internet and widely used in various vision tasks. COCO has been found to contain gender bias in web corpora (Tang et al., 2021), one of which is the spurious correlation between males and skateboards. The joint distribution of gender and Skateboard in COCO is plotted in Figure 1(b). COMPAS (Mattu) dataset consists of the results of a commercial algorithm called COMPAS, used to assess a convicted criminal's likelihood of reoffending. COMPAS dataset is widely known for its bias against African Americans and is widely used in the research of machine learning fairness (Goel et al., 2018; Hort et al., 2021; Li & Liu, 2022; Guo et al., 2023). The joint distribution of Race and Risk Level in the COMPAS dataset is plotted in Figure 1(d). Note that although COMPAS is a tabular dataset, it genuinely reflects the biased distribution in the real world. It is quite obvious that the distribution of biases in existing benchmarks and real-world datasets diverges. Additonally, CelebA (Liu et al., 2015) is another real-world image dataset. CivilComments-WILDS (CCW) (Koh et al., 2021) and MultiNLI (Williams et al., 2018) are also real-world datasets in the NLP domain. *More visualizations of these additional datasets are shown in Appendix A.* In the following subsection, we will further discuss how to measure their differences.

2.2 Previous measures of spurious correlation

We first revisit measures of spurious correlation in previous works, then point out their insufficiency.

Background. We assume a joint distribution of attributes $y^1, y^2, ..., y^K$ with $y^k \in A^k$ where A^k is a finite set. One of these K attributes is the target of learning, denoted as y^t, and a spurious attribute y^s with $t \neq s$. The definition of spurious correlation or the measure of bias magnitude is rather vague or flawed in previous works. We summarize the measures in previous works into three categories.

Target attribute conditioned probability. Previous works (Wang & Russakovsky, 2023; Reddy et al., 2021) measure spurious correlation according to the probability of a biased feature a_s within the correlated class a_t: $Corr_{tcp} = P(y^s = a_s | y^t = a_t)$. A higher value indicates a strong correlation.

Spurious attribute conditioned probability. Some (Tang et al., 2021; Lee et al., 2021; Yenamandra et al., 2023; Hermann et al., 2024) measure spurious correlation according to the probability of the correlated class a_t within samples with biased feature a_s: $Corr_{scp} = P(y^t = a_t | y^s = a_s)$. A higher value of the measure indicates a strong correlation.

Spurious attribute conditioned entropy. Nam et al. (2020) defined an entropy-based measure of bias. They use conditional entropy to measure how skewed the conditioned distribution is: $Corr_{sce} = H(y^t|y^s)$, where H is entropy. Values close to 0 indicate a strong correlation. This is an attribute-level measure, yet it is based on information theory.

We then point out the following requirements a proper measure of spurious correlation should satisfy.

Spurious correlation should be measured at the feature level. As shown in Figure 1(a) and 1(c), the predictivity of every value in the spurious attribute is similar in existing benchmarks. However, this is not the case for real-world datasets, where it is clear that the predictivity of values in the spurious attribute varies greatly, as shown in Figure 1(b) and 1(d). Therefore, to deal with real-world biases, analysis of bias should be conducted on a more fine-grained value level, i.e. feature level, rather than attribute level in previous works (Nam et al., 2020). Note that though $Corr_{tcp}$ and $Corr_{scp}$ are defined at the feature level, it is assumed by previous works (Lee et al., 2021; Reddy et al., 2021; Hermann et al., 2024) that it is consistent cross features in spurious attribute during benchmark construction, i.e. viewed as an attribute level measure.

The spurious attribute rather than the target attribute should be given as a condition. It is well recognized that the spurious attribute should be easier than the target attribute for the model to learn (Nam et al., 2020; Hermann et al., 2024). Thus the spurious attribute should be more available to the model when learning its decision rules (Hermann et al., 2024) and given as a condition when we define spurious correlation.

The marginal distribution of the target attribute should be accounted for. In $Corr_{tcp}$ and $Corr_{scp}$ measure of spurious correlation, the marginal distribution of the target attribute is not taken into account. This is inaccurate for even if the spurious and the target attribute are statistically independent, the value of $Corr_{tcp}$ and $Corr_{scp}$ could be high if the marginal distribution of spurious and target attribute is highly skewed, e.g. long-tail distributed (Zhang et al., 2021; 2023b).

Diverge rather than predictivity should be used. While $Corr_{sce}$ satisfies the above requirements, it measures the entropy difference between the conditional and marginal distribution of the target attribute, i.e. the predictivity difference. This is still inaccurate for when the entropy of the distributions is the same, the conditional distribution could still be highly diverged from the marginal distribution, thus highly correlated with the spurious attribute. However, using divergence of the distributions accurately measures how the given condition affects the distribution shift of the target attribute.

2.3 The proposed analysis framework

Given the above requirements that need to be satisfied when measuring spurious correlations, we first propose the following feature-level measure, i.e. bias magnitude.

Bias Magnitude: spurious attribute conditioned divergence. We propose a feature-level measure of spurious correlation that measures the KL divergence between the conditional and marginal distribution of the target attribute:

$$\rho_a^* = Corr_{scd} = KL(P(y^t), P(y^t|y^s = a)) \tag{1}$$

where a is the biased feature (or value) in the spurious attribute. The proposed measure satisfies all the requirements above. The above measure only describes the bias of a given feature in the dataset, i.e. feature-level bias. To further describe the bias level of a dataset, i.e. dataset or attribute level bias, we further define the prevalence of bias.

Bias Prevalence. Consider a set of biased features whose magnitude of the bias is above a certain threshold θ, i.e. $B = \{a|\rho_a^* > \theta\}$. We define the dataset-level bias by taking not only the number but also the prevalence of the biased features:

$$Prv = \sum_{a \in B} P(y^s = a) \tag{2}$$

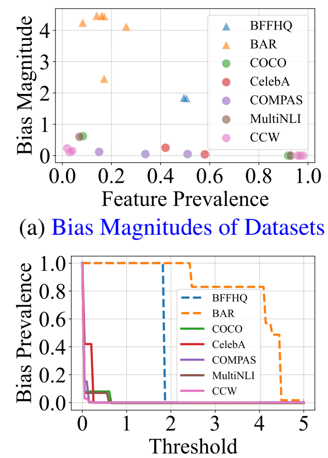
Here, we further claim and define the existence of Bias-Neutral (BN) samples, referring to samples that do not hold any biased feature defined in B. Bias-Neutral sample is a complement to the previous categorization of samples into Bias-Align (BA) and Bias-Conflict (BC) samples, which is only accurate when all samples in the dataset contain a certain biased feature, assumed by existing synthetic benchmarks (Nam et al., 2020; Liu et al., 2021; Lee et al., 2021; Lim et al., 2023). *We elaborate on the categorization of samples in Appendix D.*

2.4 Observation on real-world biases

Given the dataset assessing framework proposed above, we are now able to analyze how are dataset biases in existing benchmarks different from that in the real world.

The magnitude of biases in real-world datasets is low. As shown in Figure 2(a), the magnitude of biases in real-world datasets is significantly lower than that in existing benchmarks, consistent across various modalities. It is surprising to see how low the magnitude of biases in the real-world dataset is, yet still captured by models (Li & Liu, 2022).

The prevalence of bias in real-world datasets is low. As shown in Figure 2(b), the bias prevalence of real-world datasets is also lower than that in existing benchmarks across all thresholds. Considering the bias magnitude of real-world datasets is generally low, it seems fair to set the threshold sufficiently low when calculating the bias prevalence of existing datasets. However, even if we set the threshold to 0.1, the bias prevalence of COCO (Lin et al., 2015) and COMPAS (Mattu) dataset, i.e. 0.08 and 0.15 respectively, are still significantly lower than that of the existing benchmarks, i.e. 1. *In section 3, we further theoretically show that the above observation is not a mere exception but a manifestation of underlying principles with broader implications.*



(a) Bias Magnitudes of Datasets

(b) Bias Prevalence of Datasets

Figure 2: With our analysis framework, we can see that the bias magnitude and prevalence of real-world datasets are significantly smaller than that of existing benchmarks.

2.5 Systematic evaluation framework for real-world debiasing

Based on our analysis, we further introduce two novel bias distributions inspired from real-world applications. Together with high magnitude high prevalence (HMHP) distribution in existing benchmarks (Nam et al., 2020; Liu et al., 2021; Lee et al., 2021; Lim et al., 2023), we form a systematic evaluation framework for real-world debiasing.

Low Magnitude Low Prevalence (LMLP) Bias. Inspired by the distribution of the COMPAS (Mattu) dataset shown in Figure 2(a), bias in the real world might be low in both magnitude and prevalence. To take it even further, we should not even assume the dataset is biased at all when applying debiasing methods, because we usually lack such information in practice. Thus, unbiased data distribution can be considered as a special case of the distribution.

High Magnitude Low Prevalence (HMLP) Bias. As shown in Figure 2, the COCO (Lin et al., 2015) dataset may contain features with relatively high bias magnitude, yet low bias prevalence in the dataset due to the sparsity of the biased feature, i.e. low feature prevalence.

Note that datasets with low magnitude high prevalence (LMHP) bias do not exist due to the fact that high bias magnitude is the premise of high bias prevalence. Also, the bias distributions proposed are applicable to existing synthetic and semi-synthetic datasets mentioned in section 2. *More details on how biased distributions are defined and how to synthesize datasets with given distributions can be found in Appendix B.* In section 5, We further examine existing methods and our proposed method under real-world biases with the evaluation framework.

3 Theoretical Analysis

In this section, we theoretically show that the high bias prevalence (HP) distribution requires two strong assumptions implicitly held by existing benchmarks. Furthermore, the invalidity of the assumptions in real-world scenarios results in low bias prevalence (LP) distributions.

Data distribution. 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 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. $\tau_+ = P(y^t = +1|y^s = +1)$. 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.

Definition 1 (Simplified Magnitude of Bias)**.** *For the simplicity of theoretical analysis, we propose a simplified version of bias magnitude defined in section 1. Instead of using KL divergence as the measure of distance, we use total variation distance as a proxy for the sake of simplicity:*

$$\rho_+ = \tau_+ - p^t_+, \quad \rho_- = \tau_- - p^t_- \tag{3}$$

The simplification is consistent for it satisfies all the conditions proposed in section 2.

Definition 2 (Biased Feature)**.** *We consider a feature $y^s = a$ biased if the ratio of its bias magnitude ρ_a to its theoretical maximum $\rho_a^{max} = 1 - p^t_a$ is above certain threshold $0 \leq \theta \leq 1$:*

$$\phi_a = \frac{\rho_a}{\rho_a^{max}} > \theta$$

Definition 3 (High Bias Prevalence Distribution)**.** *We consider distribution as a high bias prevalence distribution only if both of the features in the spurious attribute are biased, i.e. $\phi_+ > \theta, \phi_- > \theta$.*

Note that the definitions above are adjusted and different from those defined in section 2.3 for the simplicity of the analysis. We then propose the two assumptions implied by high prevalence distributions, whose *proof can be found in Appendix C.*

Proposition 1 (High bias prevalence distribution assumes matched marginal distributions)**.** *Assume feature $y^s = +1$ is biased. The 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. $lim_{\theta \to 1} p^s_+ = p^t_+$.*

Proposition 2 (High bias prevalence distribution further assumes uniform marginal distributions even if they are matched)**.** *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 $\lim_{p \to 0^+} \rho^*_+ = -log(1 - \phi_+)$.The bias magnitude of the dense feature, i.e. ρ^*_-, is monotone increasing at p, with $\lim_{p \to 0^+} \rho^*_- = 0$.*

Remark 1. Proposition 2 reveals the fact that as the distribution of attributes becomes increasingly skewed, i.e. p approaches 0, the magnitude of bias for features diverges, the magnitude of sparse increases while the magnitude of dense bias approaches 0. This results in biased sparse features and unbiased dense features, resulting in LP distributions.

4 Methodology

In this section, we dive into how real-world biases would raise a challenge to existing debiasing methods. Specifically, we first set up the debiasing with biased auxiliary model (DBAM) paradigm that existing methods generally adopt in section 4.1. Then we claim the insufficiency of existing DBAM paradigm when facing real-world biases in section 4.2. Finally, aiming at the insufficiency, we proposed our approach in section 4.3.

4.1 Debiasing with biased auxiliary model

In recent years, research in the field of debiasing has been more focused on the practical setting of debiasing w/o bias supervision (Nam et al., 2020; Lee et al., 2021; Kim et al., 2022; Lim et al., 2023; Zhao et al., 2023; Lee et al., 2023). Though different in technical details, they generally adopt a biased auxiliary model to capture the bias, followed by techniques to learn a debiased model with the captured bias. The bias capture process is based on the assumption that the spurious attributes are easier and learned more preferentially than the target attribute, thus a variant version of cross entropy (CE) that emphasizes easier samples has been widely adopted, i.e. generalized cross entropy (GCE) (Zhang & Sabuncu, 2018), to train a biased auxiliary model M_b. To utilize M_b for debiasing, sample

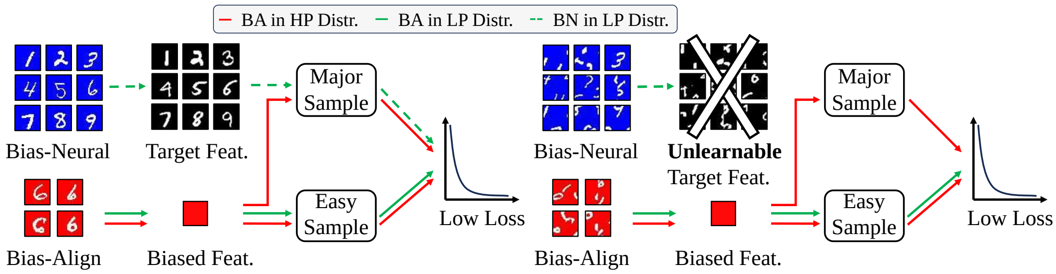


Figure 3: The bias capture process of biased models on LP and HP datasets. Assuming the red background is spuriously correlated with digit 6, and only the major learning of the biased models is illustrated with arrows. DiD eliminates the undesired learning of BN samples on the LP dataset in Figure 3(a) by destroying the target feature, as shown in Figure 3(b).

re-weighing schemes have been a major approach to train the debiased model M_d, one common implementation of which is loss-based re-weighing $W(x)$ (Nam et al., 2020):

$$W(x) = \frac{CE(M_b(x), y)}{CE(M_d(x), y) + CE(M_b(x), y)} \tag{4}$$

where (x, y) are samples from the training data and $CE(\cdot, \cdot)$ is the cross entropy loss. We note that while the above design is widely adopted in many debiasing methods, we do not limit the scope of DBAM to only methods that adopt the design of GCE and $W(x)$, but rather more generally as we name it: Debiasing methods that adopt a biased auxiliary model for debiasing, including other variants such as JTT (Liu et al., 2021) and B2T (Kim et al., 2024).

4.2 Reliance of DBAM methods on high bias prevalence

We claim that the bias capture module of the DBAM paradigm relies on the high bias magnitude of existing benchmarks, which causes failure in real-world debiasing. It is assumed by the DBAM paradigm that the biased model M_b predicts according to the bias within the training data, giving high loss to BC samples and low loss to BA samples (Nam et al., 2020; Lee et al., 2021; Kim et al., 2022; Lim et al., 2023; Zhao et al., 2023; Lee et al., 2023). Existing works attribute this loss difference to the fact that spurious attributes are easier (Nam et al., 2020; Lim et al., 2023), i.e. learned more preferentially by models, making BA the *easy sample*. While such a claim is true, we claim that the dominance of BA samples in the HP datasets is another vital causing factor of the loss difference, for *dominant/major samples* are learned more frequently than others, as shown in Figure 3(a).

However, on LP datasets, while BA samples are still easier to learn due to the biased feature, the dominant/major samples in the training data are no longer BA samples, but rather BN samples. This not only results in the loss difference between BA and BC samples decreasing but also causes low loss on BN samples, as shown in Figure 3(a). According to sample weighing scheme 4, such low loss on BN samples further leads to low weights for BN samples when training the debiased model, which is unintended because BN samples carry an abundant amount of knowledge concerning the target attribute without the interference of the spurious features. *The overlooking of BN samples results in server utility degradation when DBAM methods are applied to LP datasets*. We further empirically testify our claim in section 5.

4.3 Bias capture with feature destruction

Based on our analysis in section 4.2, we introduce a minor yet effective enhancement to the bias capture module in the existing DBAM framework. We name the refined framework as Debias in Destruction (DiD). As shown in Figure 3(b), the problem with the existing bias capture method comes from the side branch learning on BN samples of the biased auxiliary model, which not only captures the bias but also learns the target feature. This is undesired for this further causes the overlooking of BN samples when training the debiased model, as discussed in section 4.2.

To prune the side branch learning of the target features, it is intuitive to destroy the target feature and make them unlearnable when training the biased model, as shown in Figure 3(b). Such action is practical because the target features we intend to learn are usually clear, and no information on the biased feature is required. Specifically, we can achieve this by applying target feature destructive data transformation when training the biased model:

$$Loss_b = GCE(M_b(T_{fd}(x)), y)$$

where $T_{fd}(\cdot)$ is the feature destruction transformation. As an example, in visual recognition tasks, the shape of objects is a basic element of human visual perception (Geirhos et al., 2019). Therefore, the patch-shuffle destruction of shape (Lee et al., 2024) when capturing bias from visual recognition datasets is a feasible approach. As for NLP tasks, we adopt a word shuffling approach which we will elaborate on in Appendix D.5.

5 Experiments

In this section, with the systematic evaluation framework proposed in section 2.5, we design our experiments to answer the following questions: 1) How do existing DBAM methods and DiD perform on real-world biases? 2) Does DiD really emphasize BN samples as we intend? 3) How do debiasing methods perform on unbiased datasets? 4) How do the magnitude and prevalence of bias in datasets affect debiasing? 5) How sensitive is DiD to the hyper-parameters? (see Appendix E.2) 6) Is DiD effective on bias detection tasks as well? (see Appendix E.3) More analytical results in Appendix E.

5.1 Experimental settings

Metrics. Following previous works (Nam et al., 2020; Lee et al., 2021; 2023), we adopt the accuracy of BC samples (**BC**), the average accuracy on the balanced test set (**Avg**), and the worst group accuracy (**Worst Acc.**) as the evaluation metrics. We note that DiD is not a self-contained method but rather a plug-in module for existing debiasing methods to improve from their original performances. Thus, the effectiveness of DiD should be measured as the performance gain from the base method for debiasing.

Datasets. We adopt 8 datasets in various modalities for evaluation. Specifically, we adopt the basic setting of **Colored MNIST** (Reddy et al., 2021) and **Corrupted CIFAR10** (Nam et al., 2020) to implement the distributions within the proposed systematic evaluation framework. We also evaluated our method on more existing synthetic benchmarks who is more complex in terms of the target and spurious feature: **BAR** (Nam et al., 2020), **NICO** (Kim et al., 2022), and **WaterBirds** (Sagawa* et al., 2020). We also adopt 2 real-world NLP datasets **MultiNLI** (Williams et al., 2018) and **CivilComments-WILDS** (Koh et al., 2021), and 1 real-world image dataset **CelebA** (Liu et al., 2015).

Baselines. We adopt 7 baselines, covering classic and recently proposed DBAM methods. **ERM** directly applies standard training on the biased datasets. **LfF** (Nam et al., 2020) is the first work in the DBAM paradigm. **DisEnt** (Lee et al., 2021) disentangles bias and intrinsic features and applies feature augmentation when training the debiased model. **BE LfF** and **BE DisEnt** were recently proposed by Lee et al. (2023), and is based on LfF and DisEnt, respectively. **JTT** (Liu et al., 2021) is a classic DBAM method adapted to both the image and NLP domain. **Group DRO** (Sagawa* et al., 2020) is a classic debiasing method with bias supervision used as an upper bound. Detailed settings can be found in Appendix D.

5.2 Main results

Evaluation on various bias distributions. As shown in Table 1, while performing decently on HMHP distributed datasets, existing methods (Nam et al., 2020; Lee et al., 2021; 2023) fail to handle both LMLP and HMLP biases. Generally, for existing methods, the accuracy of BC samples and the average accuracy on the balanced test set is lower than the ERM baseline. This indicates the failure of exiting DBAM paradigm on the task of debiasing with low bias prevalence datasets, sacrificing utility, i.e. average accuracy, without improving worst group performance, i.e. BC accuracy. We further tested the effectiveness of DiD by combining DiD with existing DBAM methods (Nam et al., 2020; Lee et al., 2021; 2023). As shown in Table 1, when combined with DiD, the BC and average accuracy both improve on existing HMHP benchmarks. On LMLP and HMLP datasets, the superiority of DiD is even more prominent, where BC and average accuracy both improve significantly, achieving an average of +16.6 and +11.7 for LfF and DisEnt respectively.

Table 1: The performance of our approach is presented in *absolute accuracy increase* of existing methods. Results show that existing DBAM methods perform poorly on LP distributions, yet our method effectively boosts the performance of existing methods across all types of biases.

	Colored MNIST						Corrupted CIFAR10					
	LMLP		HMLP		HMHP		LMLP		HMLP		HMHP	
Algorithm	BC	Avg	BC	Avg	BC	Avg	BC	Avg	BC	Avg	BC	Avg
ERM	91.1	91.7	85.2	89.8	48.5	53.4	62.5	64.3	55.9	65.1	29.4	35.4
LfF	68.4	69.7	58.0	63.3	65.6	64.6	55.0	55.4	47.7	54.1	35.3	39.0
+ Ours	**+22.6**	**+21.4**	**+32.6**	**+25.8**	**+1.3**	**+3.4**	**+7.0**	**+7.3**	**+7.1**	**+8.9**	**+1.8**	**+2.5**
DisEnt	73.9	74.9	66.5	72.2	68.3	67.4	55.5	56.1	52.5	54.5	36.0	39.5
+ Ours	**+17.2**	**+16.5**	**+22.0**	**+16.8**	**+0.8**	**+3.1**	**+5.4**	**+5.9**	**+2.8**	**+7.1**	**+3.0**	**+3.3**
BE LfF	83.6	83.5	80.0	82.3	66.9	67.6	52.1	54.0	51.0	54.0	31.5	36.6
+ Ours	**+5.7**	**+6.1**	**+9.1**	**+4.9**	-0.5	**+0.7**	**+1.1**	**+0.2**	-0.8	**+0.1**	**+1.4**	**+0.8**
BE DisEnt	81.1	81.0	77.6	80.2	67.5	68.5	56.6	57.2	49.1	56.3	34.2	38.6
+ Ours	**+8.7**	**+9.0**	**+11.7**	**+5.5**	**+2.0**	**+2.5**	**+4.3**	**+4.2**	**+4.9**	**+5.1**	**+3.5**	**+3.2**

Table 2: Our approach consistently demonstrated the effectiveness on real-world datasets in both image and language modality. Group DRO is a supervised debiasing method, acting as an upper bound for worst-group accuracy.

	Bias supervision?	MultiNLI		CivilComments-WILDS		CelebA	
		Avg	Worst Acc.	Avg	Worst Acc.	Avg	Worst Acc.
ERM	No	80.1	76.41	92.06	50.87	95.75	45.56
JTT	No	80.51	73.02	91.25	59.49	80.49	73.13
+Ours	No	**+1.06**	**+2.71**	**+0.38**	**+6.41**	**+6.43**	**+8.50**
Group DRO	Yes	82.11	78.67	83.92	80.20	91.96	91.49

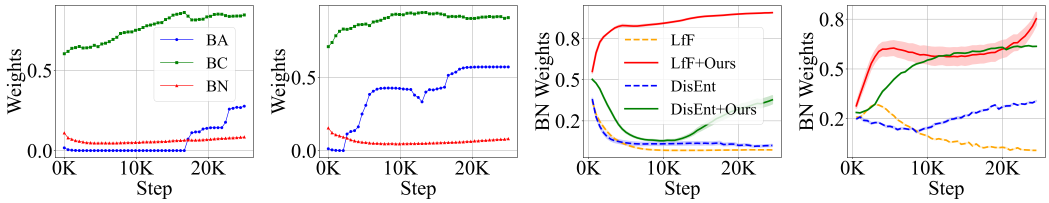
Evaluation on complex visual features. As shown in Table 3, our approach is not merely effective under the setting of Colored MNIST and Corrupted CIFAR10, but rather consistently demonstrating its superiority on datasets with more complex sets of target and spurious features. This shows the adaptability of DiD to more sophisticated visual data. Refer to Appendix D.2 for the metrics used.

Evaluation on real-world datasets in various modalities. We choose JTT as the baseline for this part of the experiment for it is a classic method adopted to both the image and NLP domain. As shown in Table 2, our approach is consistently effective on real-world datasets in various modalities, further demonstrating its generalizability.

5.3 Analysis

Emphasis on BN samples. We further validate our method by tracking the weights of samples to see if they match the purpose of our design. Figure 4(a) and 4(b) plots the average weights of all kinds of samples on HMLP distributed Colored MNIST dataset, which shows that the failure of existing methods is indeed caused by the overlooking of BN samples when training the debiased model M_d, as claimed in section 4.2. Figure 4(c) and 4(d) track the sample weight of BN samples when training LfF on HMLP distribution. As we can see, our proposed approach significantly raise the weights of the BN sample, which further demonstrates the effectiveness of our design.

Debias on unbiased datasets. As we do not know how biased or is the training data biased at all in real-world scenarios, it is important to evaluate the performance of debiasing methods on unbiased



(a) DisEnt (b) LfF (c) Colored MNIST (d) Corrupted CIFAR10

Figure 4: Figure 4(a) and 4(b) support our claim in section 4 that existing DBAM methods tend to overlook BN samples when training on LP distributions. Figure 4(a) and 4(b) show that our approach effectively emphasizes BN samples by raising its weights.

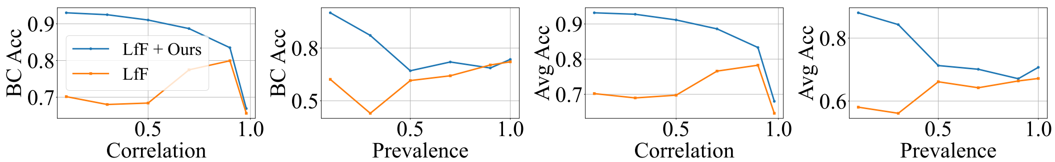


Figure 5: The performance of debiasing methods under various bias magnitudes and prevalence.

training data to ensure that they do not cause severe performance degradation, if not improving the performance. As shown in Table 4, existing DBAM methods perform poorly on unbiased training data, causing severe performance degradation, yet our approach greatly boosts their performances.

Effect of bias magnitude and prevalence in debiasing. As shown in Figure 5, we use the correlation $Corr_{scp}$ defined in section 2 as a proxy for the bias magnitude and vary it from low to high. With the increase of the bias magnitude, the performance of LfF first increases as the data become biased, and then decreases as the bias magnitude becomes extremely high, while DiD consistently improves the performance. As shown in 5, we vary the prevalence of bias by controlling the number of biased features, which can also be viewed as an interpolation between HMLP and HMHP distribution. With the increase of the bias prevalence, the performance of LfF generally keeps increasing for its reliance on high prevalence as discussed in section 4, while DiD consistently improves the performance. Those experiments are conducted on Colored MNIST dataset.

Table 3: Results on 3 datasets with more complex and realistic sets of features further show the effectiveness of our approach.

Algorithm	BAR	NICO	WaterBirds
ERM	35.32 ± 0.27	42.61 ± 0.33	56.53 ± 0.27
LfF	37.73 ± 1.00	51.69 ± 3.06	50.02 ± 0.00
+ Ours	**+3.34** ± 1.69	**+2.80** ± 3.10	**+4.47** ± 0.13
DisEnt	59.11 ± 1.75	39.73 ± 0.58	56.75 ± 4.19
+ Ours	**+3.92** ± 0.62	**+16.55** ± 1.29	**+11.29** ± 0.69
BE LfF	38.40 ± 0.65	44.09 ± 1.95	52.98 ± 0.28
+ Ours	**+1.08** ± 1.67	**+8.23** ± 1.49	**+1.92** ± 0.12
BE DisEnt	62.74 ± 1.23	39.58 ± 0.91	53.85 ± 2.14
+ Ours	**+0.70** ± 1.20	**+13.50** ± 1.62	**+1.99** ± 3.65

Table 4: Existing DBAM methods perform poorly on unbiased training data, while DiD greatly boosts the performance.

Algorithm	Colored MNIST	Corrupted CIFAR10
ERM	94.14 ± 0.21	67.91 ± 0.13
LfF	70.19 ± 1.50	52.04 ± 2.14
+ Ours	93.18 ± 0.26	57.29 ± 0.22
DisEnt	75.24 ± 3.40	58.50 ± 0.20
+ Ours	92.24 ± 0.44	64.58 ± 0.02
BE LfF	84.14 ± 0.61	55.64 ± 0.66
+ Ours	90.02 ± 0.54	56.28 ± 0.45
BE DisEnt	80.66 ± 0.90	58.57 ± 0.12
+ Ours	89.10 ± 1.28	62.97 ± 0.16

6 Conclusions and Discussion

In this work, we emphasize the importance of debias within the real world. To tackle real-world biases, we first proposed a fine-grained analysis framework to analyze dataset biases, based on which we further proposed a systematic evaluation framework for benchmarking debiasing methods under real-world biases. According to our result, we identified the insufficiency of existing methods and proposed a new approach to resolve it. Our experiments demonstrate the effectiveness of our approach. In Appendix G, we further discuss the limitations and future directions of this work.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Zhixuan Chu and Sheng Li. Causal effect estimation: Recent progress, challenges, and opportunities. *Machine Learning for Causal Inference*, pp. 79–100, 2023.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. Matching in selective and balanced representation space for treatment effects estimation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 205–214, 2020.

Zhixuan Chu, Stephen L Rathbun, and Sheng Li. Graph infomax adversarial learning for treatment effect estimation with networked observational data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 176–184, 2021.

Zhixuan Chu, Ruopeng Li, Stephen Rathbun, and Sheng Li. Continual causal inference with incremental observational data. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 3430–3439. IEEE, 2023.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z.

Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-Discriminatory Machine Learning Through Convex Fairness Criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. doi: 10.1609/aaai.v32i1.11662.

Dongliang Guo, Zhixuan Chu, and Sheng Li. Fair attribute completion on graph with missing attributes. *arXiv preprint arXiv:2302.12977*, 2023.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. arXiv, December 2015.

Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Hendrik Heuer, Christof Monz, and Arnold W. M. Smeulders. Generating captions without looking beyond objects, 2016.

Hans Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994.

Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairea: A Model Behaviour Mutation Approach to Benchmarking Bias Mitigation Methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, pp. 994–1006, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8562-6. doi: 10.1145/3468264.3468565.

Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. SelecMix: Debiased learning by contradicting-pair sampling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Ibrahim Ibrahim and Adnan Abdulazeez. The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal of Applied Science and Technology Trends*, 2(01):10–19, March 2021. doi: 10.38094/jastt20179.

Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art, 2021.

Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. BiaSwap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14992–15001, October 2021.

Nayeong Kim, SEHYUN HWANG, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18403–18415. Curran Associates, Inc., 2022.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11082–11092, June 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-Wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25123–25133. Curran Associates, Inc., 2021.

Jungsoo Lee, Jeonghoon Park, Daeyoung Kim, Juyoung Lee, Edward Choi, and Jaegul Choo. Revisiting the Importance of Amplifying Bias for Debiasing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14974–14981, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i12.26748.

Peizhao Li and Hongfu Liu. Achieving Fairness at No Utility Cost via Data Reweighing with Influence. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12917–12930. PMLR, July 2022.

Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14970–14979, October 2021.

Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. BiasAdv: Bias-Adversarial Augmentation for Model Debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3832–3841, June 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context, 2015.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Lauren Kirchner Surya Julia Angwin Mattu, Jeff Larson. Machine Bias. *ProPublica*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *Acm Computing Surveys*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684. Curran Associates, Inc., 2020.

Jeonghoon Park, Chaeyeon Chung, Juyoung Lee, and Jaegul Choo. Enhancing intrinsic features for debiasing via investigating class-discerning common attributes in bias-contrastive pair, 2024.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Nips '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabanian, and Sina Honari. Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

James Stewart. *Calculus : Early Transcendentals*. Brooks/Cole, Cengage Learning, Belmont, Cal., 2012. ISBN 0-538-49790-4 978-0-538-49790-9 0-8400-5885-3 978-0-8400-5885-0 0-538-49871-4 978-0-538-49871-5 0-538-49887-0 978-0-538-49887-6 0-8400-4825-4 978-0-8400-4825-7.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 633–645, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8312-7. doi: 10.1145/3442381.3449950.

Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3957–3968, October 2023.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3091–3100, October 2021.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101.

Siqiao Xue, Yan Wang, Zhixuan Chu, Xiaoming Shi, Caigao Jiang, Hongyan Hao, Gangwei Jiang, Xiaoyun Feng, James Zhang, and Jun Zhou. Prompt-augmented temporal point process for streaming event sequence. *Advances in Neural Information Processing Systems*, 36:18885–18905, 2023.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. FACTS: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4794–4804, October 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2361–2370, 2021.

Yi-Kai Zhang, Qi-Wei Wang, De-Chuan Zhan, and Han-Jia Ye. Learning Debiased Representations via Conditional Attribute Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7599–7608, June 2023a.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 8792–8802, Red Hook, NY, USA, 2018. Curran Associates Inc.

Bowen Zhao, Chen Chen, Qian-Wei Wang, Anfeng He, and Shu-Tao Xia. Combating Unknown Bias with Effective Bias-Conflicting Scoring and Gradient Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3561–3569, June 2023. doi: 10.1609/aaai.v37i3.25466.

A MORE VISUALIZATIONS OF BIASED DISTRIBUTIONS

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

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 6a.

Additional visualization of the biased distribution within real-world datasets is also plotted 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 6b. It is clear that the biased distribution of CelebbA aligns with that in other existing benchmarks, forming a ”diagonal distribution”.

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.

The description of MultiNLI and CivilComments-WILDS can be found in Appendix E.

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.

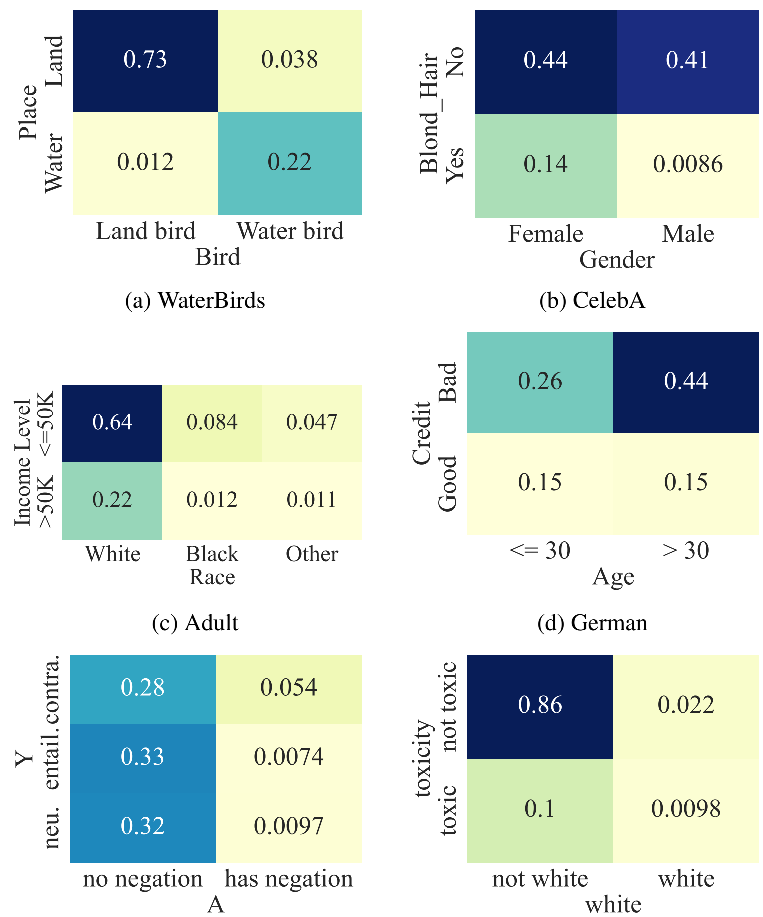
Assume a set of biased features $a_i^s \in B$ whose correlated class in the target attribute is defined by a function $g : y^s \to y^t$, which is an injection from the spurious to the target attribute. The bias magnitude of each biased feature is controlled by $corr_i = P(y^t = g(a_i^s)|y^s = a_i^s)$.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

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



(a) WaterBirds (b) CelebA (c) Adult (d) German (e) MultiNLI (f) CivilComments-WILDS

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) visualize the distribution of existing benchmarks. Figure 6(b), 6(c), 6(d), 6(e), and 6(f) visualize the distribution of real-world datasets. The biased distribution of existing benchmarks and real-world datasets is not alike.

Table 5: Configurations for biased distributions within the proposed evaluation framework

Distribution	$\lvert y^t \rvert$	$\lvert B \rvert$	$corr_i$
LMLP	10	10	0.5
LMLP'	10	5	0.5
HMLP	10	1	0.98
HMHP	10	10	0.98
Unbiased	10	0	0.1

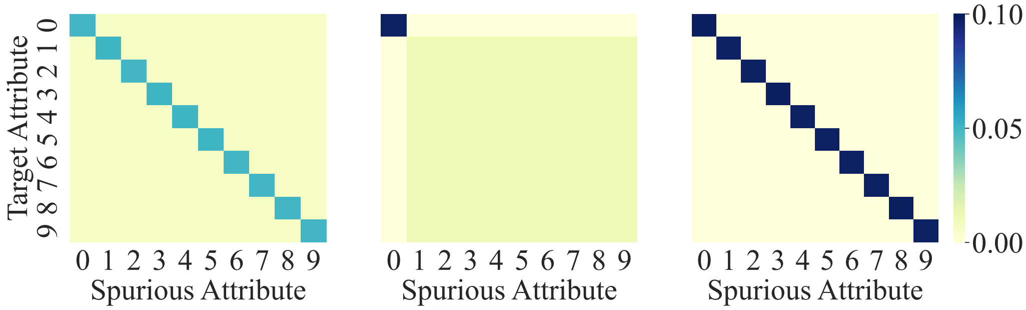


Figure 7: Visualization of biased distributions within the proposed evaluation framework under ten-class classification task. The left, middle, and right plots are visualizations for LMLP, HMLP, and HMHP distribution respectively.

Following the above equations, we further designed LMLP, HMLP, and HMHP biased distributions with the configurations in Table 5. The visualizations of the distributions when the target is a ten-class attribute are in Figure 7.

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:

$$P(y^s = +1, y^t = +1) = p^s_+ \cdot \tau_+, \tag{5}$$

$$P(y^s = +1, y^t = -1) = p^s_+(1 - \tau_+), \tag{6}$$

$$P(y^s = +1, y^t = -1) = p^t_+ - p^s_+ \cdot \tau_+, \tag{7}$$

$$P(y^s = -1, y^t = -1) = 1 - p^t_+ - p^s_+(1 - \tau_+) \tag{8}$$

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. $lim_{\theta \to 1} p^s_+ = p^t_+$.*

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:

$$\rho_+ > \theta \cdot \rho_{max}^+ = \theta \cdot (1 - p_+^t), \tag{9}$$
$$\rho_- > \theta \cdot \rho_{max}^- = \theta \cdot p_+^t \tag{10}$$

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_+:

$$\rho_- = \tau_- - p_-^t \tag{11}$$
$$= \frac{1 - p_+^t - p_+^s(1 - \tau_+)}{1 - p_+^s} - (1 - p_+^t) \tag{12}$$
$$= \frac{p_+^s(\tau_+ - p_+^t)}{1 - p_+^s} \tag{13}$$
$$= \frac{p_+^s}{1 - p_+^s}\rho_+ \tag{14}$$

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

$$\frac{p_+^s}{1 - p_+^s}(1 - p_+^t) \geq \frac{p_+^s}{1 - p_+^s}\rho_+ = \rho_- \geq \theta \cdot p_+^t \tag{15}$$
$$p_+^s \geq \frac{\theta \cdot p_+^t}{1 - p_+^t + \theta \cdot p_+^t} \geq \theta \cdot p_+^t = LB(\theta) \tag{16}$$

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

$$\tau_+ = \frac{p_+^s \cdot P(y^s = +1 | y^t = +1)}{p_+^s} \leq \frac{p_+^t}{p_+^s} \tag{17}$$
$$\tau_+ = p_+^t + \rho_+ \geq \theta(1 - p_+^t) + p_+^t \tag{18}$$

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

$$\theta(1 - p_+^t) + p_+^t \leq \tau_+ \leq \frac{p_+^t}{p_+^s} \tag{19}$$
$$p_+^s \leq \frac{p_+^t}{\theta(1 - p_+^t) + p_+^t} = UB(\theta) \tag{20}$$

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

$$\lim_{\theta \to 1} LB(\theta) = \lim_{\theta \to 1} \theta \cdot p_+^t = p_+^t \tag{21}$$
$$\lim_{\theta \to 1} UB(\theta) = \lim_{\theta \to 1} \frac{p_+^t}{\theta(1 - p_+^t) + p_+^t} = p_+^t \tag{22}$$

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

$$LB(\theta) \leq p_+^s \leq UB(\theta) \tag{23}$$
$$\lim_{\theta \to 1} p_+^s = \lim_{\theta \to 1} LB(\theta) = \lim_{\theta \to 1} UB(\theta) = p_+^t \tag{24}$$

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 $\lim_{p\to 0^+} \rho^*_+ = -log(1-\phi_+)$.The bias magnitude of the dense feature, i.e. ρ^*_-, is monotone increasing at p, with $\lim_{p\to 0^+} \rho^*_- = 0$.*

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:

$$\tau_+ = p + \phi_+(1-p) \tag{25}$$
$$\tau_- = 1 - p + \phi_+ \cdot p \tag{26}$$

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 $t(p)$.

$$\begin{aligned}
\rho^*_+ &= KL(P(y^t), P(y^t|y^s = +1)) && (27)\\
&= p \cdot log(\frac{p}{\tau_+}) + (1-p) \cdot log(\frac{1-p}{1-\tau_+}) && (28)\\
&= p \cdot log(\frac{p}{p+\phi_+(1-p)}) + (1-p) \cdot log(\frac{1-p}{1-p-\phi_+(1-p)}) && (29)\\
&= p \cdot log(\frac{p}{p+\phi_+(1-p)}) + (1-p) \cdot log(\frac{1}{1-\phi_+}) && (30)\\
&= p \cdot log(\frac{p(1-\phi_+)}{p+\phi_+(1-p)}) + log(\frac{1}{1-\phi_+}) = t(p) && (31)
\end{aligned}$$

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

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

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

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

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

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

$$f'(x) = \frac{1}{x} - 1 \geq 0 \tag{35}$$
$$f(1) = 0 \tag{36}$$

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

$$\frac{\partial \rho^*_+}{\partial p} = \frac{\partial t(p)}{\partial p} < 0 \tag{37}$$

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

$$\begin{aligned}
\rho^*_- &= KL(P(y^t), P(y^t|y^s = -1)) && (38)\\
&= (1-p) \cdot log(\frac{(1-p)(1-\phi_+)}{1-p+\phi_+ \cdot p}) + log(\frac{1}{1-\phi_+}) && (39)\\
&= t(1-p) && (40)
\end{aligned}$$

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

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

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

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

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

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

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

$$= \lim_{p\to 0^+} (p \cdot log(\frac{p(1-\phi_+)}{p+\phi_+(1-p)})) + log(\frac{1}{1-\phi_+}) \tag{46}$$

$$= \lim_{p\to 0^+} (p \cdot log(p)) + \lim_{p\to 0^+} (p \cdot log(\frac{1-\phi_+}{p+\phi_+(1-p)})) + log(\frac{1}{1-\phi_+}) \tag{47}$$

$$= \lim_{p\to 0^+} \frac{log(p)}{\frac{1}{p}} + log(\frac{1}{1-\phi_+}) \tag{48}$$

$$= \lim_{p\to 0^+} \frac{(log(p))'}{(\frac{1}{p})'} + log(\frac{1}{1-\phi_+}) \tag{49}$$

$$= \lim_{p\to 0^+} \frac{\frac{1}{p}}{-\frac{1}{p^2}} + log(\frac{1}{1-\phi_+}) \tag{50}$$

$$= log(\frac{1}{1-\phi_+}) \tag{51}$$

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

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

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

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

$$= 0 \tag{55}$$

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 $g : y^s \to y^t$ that maps the biased features to its correlated class.

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

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

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

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 semi-synthetic dataset deliberately curated to contain 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. We report the accuracy of bias-conflicting samples following Nam et al. (2020).

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 a curated animal subset of NICO that exhibits strong biases (thus still semi-synthetic), 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. We report the average accuracy on the test set following Kim et al. (2022).

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". We report the worst group accuracy following Liu et al. (2021).

MultiNLI (Williams et al., 2018). Given a pair of sentences, the task is to classify whether the second sentence is entailed by, neutral with, or contradicts the first sentence. We use the spurious attribute from Sagawa* et al. (2020), which is the presence of negation words in the second sentence; due to the artifacts from the data collection process, contradiction examples often include negation words.

CivilComments-WILDS (Koh et al., 2021). The task is to classify whether an online comment is toxic or non-toxic, and the label is spuriously correlated with mentions of certain demographic identities (male, female, White, Black, LGBTQ, Muslim, Christian, and other religion). We use the evaluation metric from Koh et al. (2021), which defines 16 overlapping groups (a, toxic) and (a, non-toxic) for each of the above 8 demographic identities a, and report the worst-group performance over these groups.

D.3 Baselines

LfF (Nam et al., 2020). Learning from Failure (LfF) 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 (Lee et al., 2021) . The DisEnt 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 (Lee et al., 2023). BiasEnsemble (BE) 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.

JTT (Liu et al., 2021). JTT is a classic DBAM method that has been applied to both the image and NLP domains. JTT identifies challenging examples by training an initial model using standard empirical risk minimization (ERM) and collecting misclassified examples into an error set. The second stage involves re-training the model while upweighting the error set to prioritize examples that the first-stage model struggled with. This approach aims to address performance disparities caused by spurious correlations, leading to better generalization across groups with minimal additional annotation costs.

Group DRO (Sagawa* et al., 2020). Group DRO is a supervised debiasing method aiming to improve the worst group accuracy. It is commonly used as an upper bound in the worst group accuracy for unsupervised methods.

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, BAR, NICO, WaterBirds dataset, we train ResNet18 He et al. (2015) with random initialization. For CelebA dataset, we train ResNet50 with random initialization, following Liu et al. (2021). For MultiNLI and CivilComments-WILDS datasets, we use Bert for training, following Liu et al. (2021).

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 Design of feature destructing methods

For in visual recognition tasks, the shape of objects is a basic element of human visual perception (Geirhos et al., 2019). Therefore, the patch-shuffle destruction of shape (Lee et al., 2024) when capturing bias from visual recognition datasets is a feasible approach. We adopt the patch-shuffle approach for all the visual dataset within the paper except for CelebA. We apply a gray-scale transformation for CelebA as its recognition task is hair color. Anyhow, the feature destruction method could be highly flexible for different tasks.

For NLP tasks, we first introduce the common biases within the NLP domain followed by a simple design of feature destruction method in the NLP domain. The commonly used NLP datasets for debiasing are MultiNLI and CivilComments-WILDS dataset. Specifically, the bias within the MultiNLI dataset is the correlation between the negation words and the entailment task and the bias within the CivilComments-WILDS dataset is the correlation between words implying demographic identities and the toxicity task. The target features of both datasets are semantic information of the sentences where the position of words matters, and the spurious features are the individual words which is insensitive to positions. Furthermore, such position sensitivity difference between target and spurious features within NLP biases is not limited to these two datasets but rather quite common. For example, CLIP has also been found with the "bag of words" phenomenon (Yuksekgonul et al., 2023), which ignores the semantic meaning of the inputs and relies on words individually for prediction. As a result, a straightforward approach for feature destruction is to shuffle the words within the sentences.

D.6 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 6 and Table 7, 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.

Table 6: 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 ± 0.09	91.13 ± 0.17	-	91.73 ± 0.16
	LfF	80.25 ± 4.86	68.41 ± 2.01	-	69.74 ± 2.41
	+ DiD	92.16 ± 0.35	91.03 ± 0.15	-	91.15 ± 0.17
	BE LfF	82.95 ± 1.68	83.60 ± 0.85	-	83.53 ± 0.75
	+ DiD	93.49 ± 0.81	89.25 ± 0.64	-	89.67 ± 0.54
	DisEnt	84.45 ± 1.72	73.87 ± 2.52	-	74.93 ± 2.44
	+ DiD	**94.03 ± 0.66**	**91.09 ± 0.24**	-	**91.38 ± 0.28**
	BE DisEnt	80.18 ± 1.94	81.07 ± 2.50	-	80.98 ± 2.29
	+ DiD	91.89 ± 0.26	89.80 ± 0.97	-	90.01 ± 0.89
HMLP	ERM	99.32 ± 0.34	85.25 ± 1.62	90.30 ± 0.56	89.82 ± 0.70
	LfF	87.76 ± 4.12	57.98 ± 3.58	63.72 ± 3.22	63.35 ± 3.02
	+ DiD	82.99 ± 5.08	**90.54 ± 0.74**	**89.04 ± 0.84**	89.12 ± 0.77
	BE LfF	57.65 ± 32.14	80.02 ± 1.10	82.84 ± 1.68	82.33 ± 1.93
	+ DiD	63.95 ± 15.64	89.11 ± 1.29	87.28 ± 1.54	87.22 ± 1.58
	DisEnt	77.55 ± 7.93	66.52 ± 8.75	72.69 ± 5.91	72.18 ± 6.05
	+ DiD	**88.78 ± 7.24**	88.52 ± 1.47	89.04 ± 1.13	**88.99 ± 1.16**
	BE DisEnt	41.84 ± 6.21	77.59 ± 0.69	80.87 ± 1.78	80.19 ± 1.71
	+ DiD	31.97 ± 7.08	89.33 ± 1.07	85.88 ± 0.86	85.66 ± 0.89
HMHP	ERM	99.57 ± 0.07	48.54 ± 1.22	-	53.38 ± 1.10
	LfF	57.16 ± 8.27	65.62 ± 2.87	-	64.59 ± 3.31
	+ DiD	77.84 ± 2.49	66.91 ± 1.73	-	68.00 ± 1.80
	BE LfF	73.61 ± 1.03	66.90 ± 0.43	-	67.57 ± 0.47
	+ DiD	**85.65 ± 2.53**	66.37 ± 2.54	-	68.30 ± 2.50
	DisEnt	59.89 ± 4.19	68.29 ± 1.43	-	67.45 ± 1.28
	+ DiD	83.65 ± 0.13	69.05 ± 0.38	-	70.51 ± 0.33
	BE DisEnt	77.74 ± 2.51	67.51 ± 1.33	-	68.53 ± 1.45
	+ DiD	84.62 ± 1.16	**69.50 ± 1.23**	-	**71.01 ± 1.08**

E.2 Hyper-parameter sensitivity

As shown in Table 9, 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.

E.3 Application of DiD on the bias detection task

Some recent work (Yenamandra et al., 2023; Kim et al., 2024) has focused on the task of detecting biases rather than debiasing directly. Such methods also involve a biased auxiliary model for the detection. To test the effectiveness of DiD on bias detection tasks, we apply DiD to the recently proposed B2T (Kim et al., 2024) method. Specifically, B2T detects bias keywords by calculating their CLIP score, whose calculation involves a biased auxiliary model to define an error dataset, similar to

Table 7: 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 ± 0.81	62.50 ± 0.15	-	64.29 ± 0.06
	LfF	59.13 ± 0.68	55.03 ± 0.04	-	55.44 ± 0.09
	+ DiD	69.47 ± 0.96	**62.04 ± 0.21**	-	**62.78 ± 0.10**
	BE LfF	70.87 ± 1.30	52.10 ± 0.30	-	53.98 ± 0.40
	+ DiD	63.23 ± 2.10	53.21 ± 0.20	-	54.21 ± 0.38
	DisEnt	61.58 ± 0.57	55.45 ± 0.23	-	56.06 ± 0.17
	+ DiD	**72.23 ± 0.74**	60.84 ± 0.40	-	61.98 ± 0.30
	BE DisEnt	62.73 ± 0.61	56.59 ± 0.08	-	57.20 ± 0.13
	+ DiD	65.98 ± 0.40	60.92 ± 0.20	-	61.42 ± 0.21
HMLP	ERM	84.67 ± 0.64	55.85 ± 0.17	65.75 ± 0.00	65.05 ± 0.13
	LfF	73.33 ± 1.67	47.70 ± 0.58	54.58 ± 0.49	54.15 ± 0.41
	+ DiD	**78.67 ± 2.14**	**54.81 ± 2.26**	**63.71 ± 2.69**	**63.06 ± 2.63**
	BE LfF	70.33 ± 2.19	50.96 ± 2.35	54.14 ± 0.25	54.02 ± 0.36
	+ DiD	68.80 ± 0.88	50.20 ± 0.79	54.39 ± 0.18	54.15 ± 0.15
	DisEnt	61.67 ± 1.67	52.48 ± 0.56	54.65 ± 0.56	54.53 ± 0.49
	+ DiD	73.67 ± 2.64	55.26 ± 0.93	62.11 ± 0.17	61.61 ± 0.13
	BE DisEnt	75.33 ± 5.21	49.15 ± 1.54	56.86 ± 0.30	56.35 ± 0.35
	+ DiD	78.40 ± 1.00	54.09 ± 1.07	62.05 ± 0.34	61.50 ± 0.38
HMHP	ERM	89.97 ± 0.34	29.37 ± 0.30	-	35.43 ± 0.24
	LfF	72.70 ± 0.81	35.30 ± 0.33	-	39.04 ± 0.33
	+ DiD	82.07 ± 1.09	37.05 ± 0.31	-	41.55 ± 0.19
	BE LfF	**82.73 ± 0.92**	31.48 ± 0.82	-	36.61 ± 0.65
	+ DiD	78.30 ± 0.47	32.90 ± 1.79	-	37.44 ± 1.61
	DisEnt	70.77 ± 2.27	36.04 ± 0.62	-	39.51 ± 0.36
	+ DiD	76.60 ± 0.70	**39.05 ± 0.35**	-	**42.80 ± 0.25**
	BE DisEnt	78.60 ± 1.56	34.20 ± 0.43	-	38.64 ± 0.38
	+ DiD	78.70 ± 1.47	37.72 ± 0.96	-	41.82 ± 0.91

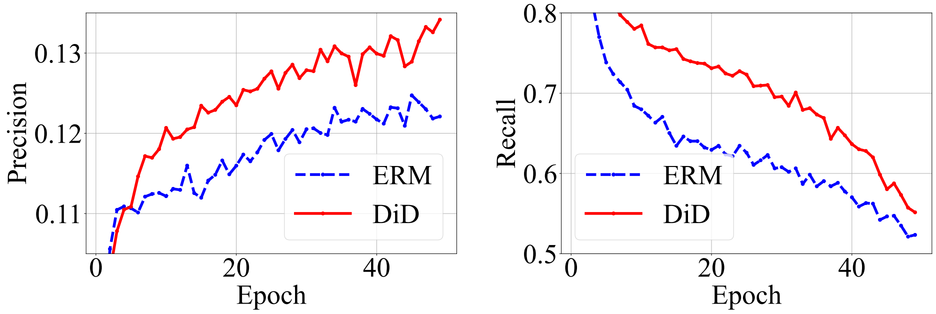
JTT. A keyword is identified as biased if it has a higher CLIP score and the subgroup defined by it should have lower accuracy.

Following Kim et al. (2024), we use CelebA as the dataset for bias detection, where the keyword "Actor" (a proxy for Male) is considered ground truth for class Blond, and the keyword "Actress" (a proxy for Female) is considered ground truth for the class not Blond. As we can see in Table **??**, by applying DiD to the training of the auxiliary model, we effectively improve both metrics CLIP score and subgroup accuracy, enhancing B2T's bias detection ability.

To further validate the effectiveness of DiD in improving the quality of the error dataset, we adopt the worst group precision and recall metrics proposed by Liu et al. (2021) for evaluation. Specifically, the worst group precision and recall indicate how accurately the error dataset represents the worst group samples. As shown in Figure 8, DiD improves both worst group precision and recall, demonstrating better bias identification ability.

Table 8: DiD effectively improves the bias identification ability of B2T, improving both CLIP Score and Subgroup Accuracy on the ground truth bias keywords of CelebA dataset.

	Blond: Actor		Not Blond: Actress	
	CLIP Score↑	Subgroup Acc.↓	CLIP Score↑	Subgroup Acc.↓
B2T	0.125	86.71	2.188	97.11
B2T + DiD	0.188	85.29	2.297	95.81

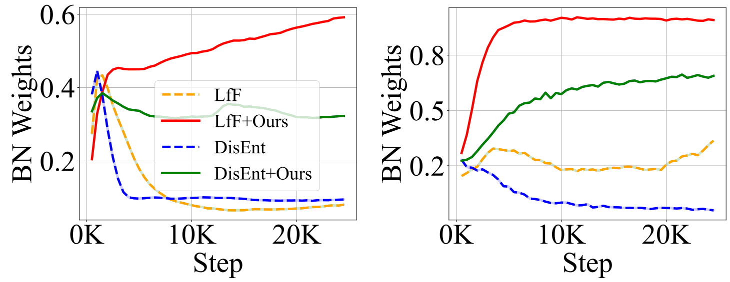


(a) Worst group precision (b) Worst group recall

Figure 8: DiD consistently improves the worst group precision and recall in the error dataset across the epochs.

Table 9: 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



(a) Colored MNIST (b) Corrupted CIFAR10

Figure 9: DiD consistently emphasizes BN samples in LMLP distributions across datasets and algorithms. Our approach is marked with solid lines.

Table 10: Results demonstrate that DiD is consistently effective regardless of different experimental settings of WaterBirds. The results are based on the ResNet50 architecture.

	Bias supervision	WaterBirds	
		Avg Acc.	Worst-group Acc.
ERM	No	78.82	31
JTT	No	90.99	65.26
+DiD	No	**+3.45**	**+17.45**
Group DRO	Yes	92.89	83.49

E.4 Results of BN samples under LMLP settings

To further examine the correctness of our analysis and the effectiveness of our design, we show the weights of BN samples under the LMLP settings. As the LMLP distribution defined in the main paper contains biased features with similar levels of bias magnitude, the choice of threshold for identifying BN samples becomes not so intuitive. Thus a threshold of 0 is selected for the categorization in the main paper, defining all samples either BA or BC samples. Consequently, we define another version of LMLP distribution named LMLP' where the magnitude of bias for each feature is low but at the same time distinguishable from each other. (Please refer to Appendix B for details) Based on LMLP' we are able to confidently define BN samples for the BN weights analysis. As shown in Figure 9, DiD consistently emphasizes BN samples in the LMLP distribution across datasets and debiasing algorithms.

E.5 Addtional results on the WaterBirds dataset

As mentioned in Appendix D, the evaluations on the WaterBirds dataset are based on the ResNet18 architecture, which is the architecture widely adopted by many previous works (Nam et al., 2020; Lee et al., 2021; 2023). However, there are also some other works (Liu et al., 2021) that evaluate the WaterBirds dataset based on the ResNet50 architecture with better baseline performances. To demonstrate that our approach is consistently effective regardless of the experimental settings, we further test our approach with the exact same setting in Liu et al. (2021). As shown in Table 10, DiD is consistently effective regardless of different experimental settings of WaterBirds.

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; 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 **?**. 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 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 GCE for bias capture, and the loss-based sample re-weighing 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.

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: 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 Boarder 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.