Adaptive Bias Discovery for Learning Debiased Classifier

Jun-Hyun Bae[1], Minho Lee[1,2], and Heechul Jung[1]

[1] Kyungpook National University, Daegu, Republic of Korea
{junhyun.bae, heechul}@knu.ac.kr
[2] ALI Co., Ltd., Daegu, Republic of Korea
mholee@gmail.com

Abstract. Training deep neural networks with empirical risk minimization (ERM) often captures dataset biases, hindering generalization to new or unseen data. Previous solutions either require prior knowledge of biases or utilize training intentionally biased models as auxiliaries; however, they still suffer from multiple biases. To address this, we introduce Adaptive Bias Discovery (ABD), a novel learning framework designed to mitigate the impact of multiple unknown biases. ABD trains an auxiliary model to be adapted to biases based on the debiased parameters from the debiasing phase, allowing it to navigate through multiple biases. Then, samples are reweighted based on the discovered biases to update debiased parameters. Extensive evaluations of synthetic experiments and real-world datasets demonstrate that ABD consistently outperforms existing methods, particularly in real-world applications where multiple unknown biases are prevalent.

Keywords: Debiasing · Spurious Correlations · Deep Learning · Classification

1 Introduction

Conventional deep learning methodologies rely on the principle of empirical risk minimization (ERM), approximating the true risk based on empirical risk derived from the training data. However, when the training dataset exhibits bias features, the empirical risk may not accurately reflect the true underlying distribution, potentially giving rise to models that amplify these biases when predicting outcomes on unseen data. Consequently, although the model demonstrates high performance on validation or test data that are independently and identically distributed (i.i.d.) in conformity with the training data distribution—where such biases are chiefly evident—its effectiveness wanes when dealing with data that deviate from the training distribution (*i.e.*, data that represent distributional shifts or out-of-distribution (OoD) scenario). Recent studies indicate that learning methods based on ERM are prone to absorbing spurious signals from the dataset, leading to suboptimal generalization performance in real-world applications [5, 14, 16, 19, 21].

To address the challenges of conventional ERM-based approaches, there have been alternative strategies to better handle biases and distributional shifts in data. One of the promising approaches is distributionally robust optimization (DRO) [8,12]. Instead of relying solely on the empirical distribution of the training data, DRO aims to optimize model performance over a worst-case distribution within a specified uncertainty set, capturing potential variations from the empirical distribution. This approach enables DRO to optimize the model's robustness against uncertain distributional shifts, particularly when biases in the dataset are known a priori. In such cases, uncertainty sets can be crafted by defining groups based on prior bias information, specifically categorizing data samples according to the existence of biases. This strategy leads to a specialized form of DRO known as group DRO [18, 29]. While group DRO can effectively tackle biases when bias annotations are given, a limitation to consider is that the process of annotating data for group membership can be labor-intensive in practical applications, particularly in the absence of pre-existing bias information.

In order to obtain bias information, recent research has considered utilizing intentionally trained biased models as auxiliaries to address the high cost of obtaining information on bias, under the premise that neural networks easily learn biases [26]. Several methods have been proposed for training these biased models, including training models with small capacity [30], using only a subset of the training data [34], or fully training on the training data [4]. While these methods may reduce the necessity for manual bias detection in datasets, their efficacy is limited when addressing multiple coexisting biases, as deep neural networks often exploit the most dominant, simpler biases [20,32]. Consequently, these models may not adequately capture the interplay among multiple biases in complex real-world datasets.

In response to these challenges, we propose a novel strategy for addressing multiple biases in datasets. Our method is rooted in the hypothesis that the diversity of biases in a dataset can be unveiled more effectively by exploring various learning dynamics rather than by relying solely on a singular biased model. Building upon this hypothesis, we introduce a method called Adaptive Bias Discovery (ABD), designed to explore and address the intricate world of multiple biases present within datasets. ABD trains auxiliary models initialized from debiased parameters, thereby capturing different learning dynamics that are more sensitive to previously unrevealed biases in the data. It then reweights samples to allow the model to address the discovered biases. This cyclical process ensures a comprehensive adaptation to the complex bias landscapes.

To demonstrate the efficacy of our proposed method, we evaluate ABD using both synthetic and real-world datasets. In our experiments, ABD consistently outperformed others by effectively addressing multiple biases, a feat the previous methods often fell short of. Moreover, our evaluations confirmed that ABD excels in crafting robust models capable of handling distributional shifts, as evidenced by its performance on real-world datasets including image classification

tasks (e.g., MetaShift, Camelyon17-wilds, FMoW-wilds) and natural language processing tasks (e.g., CivilComments-wilds, MultiNLI).

2 Related Work

Biases inherent in real datasets can degrade a model's predictive performance in practical applications. While some studies have tackled this challenge using adversarial training [6, 7], data sample reweighting [31], or ensembling biased models [11, 17, 24], these approaches typically depend on prior human expert knowledge of the biases. However, in many practical scenarios, especially with large-scale datasets, this prior knowledge may not be readily available, making these methods less feasible.

To detect unknown biases without relying on domain expertise, recent studies have explored the use of intentionally biased models to identify biased data samples. Using predictions from this biased model, they either reweight data samples [23, 26], combine the predictions of the biased model with those of the main model through a product-of-experts approach [30], or categorize training data based on the biased model's predictions [4]. Most of these approaches first establish a biased model for bias detection. Consequently, as reported by [32], biased models resulting from these methods often capture only superficial biases. In contrast, our method is designed to uncover multiple underlying biases by intertwining the bias discovery and debiasing stages in the learning process.

A recent study demonstrates the limitation of conventional bias discovery strategies in debiasing models [20]. This study employs knowledge distillation from the main classifier to multiple auxiliary classifiers, enabling a focus on biases not addressed by the main classifier. In contrast, our method draws inspiration from meta-learning literature, enabling it to adapt to multiple biases in datasets. By leveraging meta-learning principles, our approach initializes its adaptation stage with parameters from debiased models, enhancing its ability to adjust to previously unaddressed biases.

3 Method

3.1 Preliminaries

Given a training dataset drawn from the distribution $\mathcal{P}$, we consider a classification task under the assumption that dataset biases are present. Let f_θ be a classifier parameterized by θ, and let $\mathbb{E}_{\mathcal{P}}\left[l(f_\theta;(x,y))\right]$ represent the expected loss of the classifier f_θ under the distribution $\mathcal{P}$, where $l(\cdot)$ is a loss function, x and y are input and label, respectively. Empirical Risk Minimization (ERM) seeks to identify parameters θ that minimize the expected loss under the empirical distribution $\hat{\mathcal{P}}$:

$$\theta^*_{\text{ERM}} := \arg\min_\theta \mathbb{E}_{(x,y)\sim\hat{\mathcal{P}}}\left[l(f_\theta;(x,y))\right]. \tag{1}$$

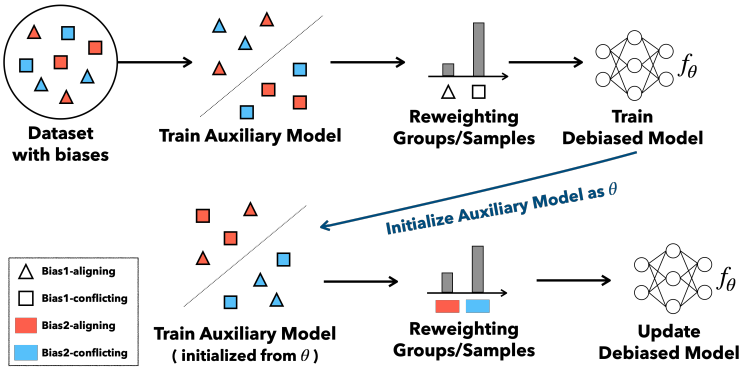


Fig. 1: Schematic of our adaptive bias discovery (ABD) framework. For simplicity, this illustration considers only a two-bias case, Bias1 and Bias2, along with two training steps. In this scenario, Bias1 represents simpler features that neural networks can learn more readily than the more complex features of Bias2. Our proposed method initializes the auxiliary model with parameters derived from debiased models, enabling it to focus on previously unidentified biases.

In the group distributionally robust optimization (group DRO) setting [18, 29], the distribution $\mathcal{P}$ is conceptualized as a combination of m groups, where each group is denoted as $\mathcal{P}_g$ and $g \in \mathcal{G} = \{1, 2, \ldots, m\}$. It is important to note that group DRO operates on the premise that prior knowledge of biases, which inform the grouping, is available during the training phase. Then, by optimizing for the worst-case group, the model can become more robust against these biases. The optimization goal for group DRO, therefore, is to optimize parameters θ to minimize the expected loss of the most challenging group for each learning step:

$$\theta^*_{\text{DRO}} := \arg\min_{\theta} \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{\mathcal{P}}_g} \left[l(f_\theta; (x, y)) \right]. \tag{2}$$

Proposition 2 from [1] shows that group DRO is equivalent to minimizing a weighted average of training group errors. As a result, the algorithm aims to identify combinations of group distributions that eliminate spurious correlations between biases and labels. However, when biases are not pre-identified, and groups are formed either randomly or solely based on domain/environment knowledge, simply combining training groups may not effectively capture the desired distribution.

3.2 Proposed Method

Our proposed method, Adaptive Bias Discovery (ABD), operates in two distinct stages, each tailored for a specific purpose: (1) The initial stage focuses on detecting biases by adapting to the training data and forming groups based on the predictions of a bias-adapted model. (2) The subsequent stage then mitigates the

identified biases by minimizing the worst-case group loss for updating debiased parameters.

Adaptive Bias Discovery In situations where bias information is not readily available, the first stage of our method aims to autonomously discover these latent biases using the predictions from an intentionally biased model. Drawing inspiration from existing automated bias identification techniques, we simply train a model so that the model is sensitive to the surface patterns inherent in the training data [2, 34, 37]. Given a model f_θ, we determine the parameters ϕ that are tailored to these superficial biases by employing the following gradient descent step:

$$\phi = \theta - \alpha \nabla_\theta \mathcal{L}(f_\theta). \tag{3}$$

Here, α denotes the step size, and $\mathcal{L}(f_\theta)$ indicates the average loss computed for f_θ over the training data. During the learning process, the model parameters θ are adjusted based on the training data, aiming to ensure that the newly adapted parameters, ϕ, capture superficial biases present within the data. Initially, the parameters θ are randomly initialized; the outcomes from the subsequent debiasing stage of ABD are then employed as the foundation for the next iteration of bias discovery. While Equation (3) illustrates a single gradient update for simplicity, multi-step gradient updates can also be employed to obtain ϕ.

With bias-adapted parameters ϕ, we categorize data samples into two groups based on predictions of the model f_ϕ: $G^{\odot}$, which contains samples correctly predicted, and $G^{\otimes}$, for those incorrectly predicted. When f_ϕ leverages biased features, samples in $G^{\odot}$ are those that align with these biases, while those in $G^{\otimes}$ counter them.

Reweighting Data Samples based on Discovered Biases After identifying biases and forming the corresponding groups, our next objective is to minimize the influence of the spurious correlations between bias features and labels. To achieve this, we minimize the worst-case risk among the obtained groups to update the debiased parameters θ. We employ the online version of group DRO [29], which involves reweighting groups during the learning process. This approach characterizes the worst-case group through convex combinations of groups, rather than directly selecting the worst-case group as the minimization objective for updating θ.

At every learning iteration, $J(\theta)$, the minimization objective of ABD is expressed as:

$$J(\theta) = a\,\mathcal{L}_{G^{\odot}}(f_\phi) + b\,\mathcal{L}_{G^{\otimes}}(f_\phi). \tag{4}$$

Here, $\mathcal{L}_g(f_\phi)$ represents the average loss of group g for f_ϕ, a and b are group weights. In our method, the weights a and b are derived from the softmax function applied to group losses with temperature coefficient, τ. Concretely, weights a and

b are determined by following equations:

$$a = \frac{e^{(\mathcal{L}_{G^{\odot}}(f_\phi)/\tau)}}{e^{(\mathcal{L}_{G^{\odot}}(f_\phi)/\tau)} + e^{(\mathcal{L}_{G^{\otimes}}(f_\phi)/\tau)}}, b = 1 - a. \tag{5}$$

This coefficient, τ, modulates the output distribution's sharpness: a higher τ results in a more uniform distribution, whereas a reduced τ sharpens it. Therefore, opting for a large τ ensures that a is approximately equivalent to b, indicating that $J(\theta)$ aligns with the average training loss typical in ERM while opting for a small temperature, $\tau \ll 1$, represents an explicit selection of the worst-group, i.e. unbiased samples, for optimization.

The model parameters θ are updated to minimize the worst-case loss with gradient descent as the following update rule:

$$\theta \leftarrow \theta - \beta \nabla_\theta J(\theta), \tag{6}$$

where β represents the step size in the gradient descent. The updated parameters θ then become the initial point for the next learning step of bias discovery. To improve efficiency of computing $\nabla_\theta \phi$, we apply a first-order approximation by considering $I - \alpha \nabla^2_\theta \mathcal{L}(f_\theta) \approx I$, following the approach of [13]. Figure 1 displays a schematic diagram of our framework, outlining its two-stage process for adaptive bias discovery and mitigation.

Previous studies in bias mitigation have often employed intentionally biased models that were designed with a separate bias discovery stage. Consequently, these models primarily showcased dominant bias features. In contrast, in our method, we integrate the two stages into a unified learning framework reminiscent of the Model-Agnostic Meta-Learning (MAML) [13] algorithm.

At a high level, within ABD, the model parameters θ are guided to enhance robustness against detected biases, while the biased model's parameters ϕ are directed towards identifying undiscovered biases. These parameters ϕ are adapted to subsequent mini-batches based on the debiased parameters θ. We provide empirical evidence in our experiments that the intentionally biased model f_ϕ can discern various biased features throughout the learning iterations.

3.3 Theoretical Analysis

In this section, we delve deeper into the theoretical foundations of our proposed learning framework, comparing it with the traditional group DRO. The fundamental distinction between these two methods lies in the treatment of bias groups. While group DRO operates with predefined groups based on prior knowledge of biases, ABD adaptively discovers and updates a subset of biases at each iteration of the learning process. Our theoretical analysis focuses on the worst-case risk gap between group DRO and ABD, where ABD's bias discovery strategy introduces a dynamic component into the learning process. The group DRO objective is formulated as follows:

$$\min_\theta \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{\mathcal{P}}_g} \left[l(f_\theta; (x, y)) \right], \tag{7}$$

where the group set $\mathcal{G}$ is predefined with bias information. In contrast, ABD adaptively uncovers biases during, training and this adaptive nature can be viewed as a stochastic process, where the selection of groups occurs according to probabilistic mechanisms, rather than fixed, predefined groupings. Without making specific assumptions about the probabilities associated with the observation of bias features, we can formulate the objective of such an optimization process as:

$$\min_{\theta} \mathbb{E}_{\mathcal{G}' \sim \mathcal{B}} \left[\max_{g \in \mathcal{G}'} \mathbb{E}_{(x,y) \sim \hat{\mathcal{P}}_g} \left[l(f_\theta; (x, y)) \right] \right], \tag{8}$$

where $\mathcal{B}$ denotes a random variable representing a stochastic selection of groups from the set $\mathcal{G}$, and $\mathcal{G}'$ indicates the realization of $\mathcal{B}$. The expected difference between the optimization problems (7) and (8) arises from the worst-case risk as:

$$\begin{aligned} Gap(\theta) = \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{\mathcal{P}}_g} \left[l(f_\theta; (x, y)) \right] \\ - \mathbb{E}_{\mathcal{G}' \sim \mathcal{B}} \left[\max_{g \in \mathcal{G}'} \mathbb{E}_{(x,y) \sim \hat{\mathcal{P}}_g} \left[l(f_\theta; (x, y)) \right] \right]. \end{aligned} \tag{9}$$

The following theorem provides a bound on the worst-case gap between the two methods.

Theorem 1 (Bound on Worst-Case Risk Gap).
Let $\mathcal{L}_g(f_\theta) = \mathbb{E}_{(x,y) \sim \hat{\mathcal{P}}_g} \left[l(f_\theta; (x, y)) \right]$ *denote the expected loss for group* g*, and define*

$$\Delta(\theta) = \max_{g \in \mathcal{G}} \mathcal{L}_g(f_\theta) - \min_{g \in \mathcal{G}} \mathcal{L}_g(f_\theta)$$

as the maximum discrepancy in expected loss across the groups. Then, for any θ*, the gap given by Equation (9) is bounded by:*

$$Gap(\theta) \leq \frac{2^{m-1} - 1}{2^m - 1} \Delta(\theta),$$

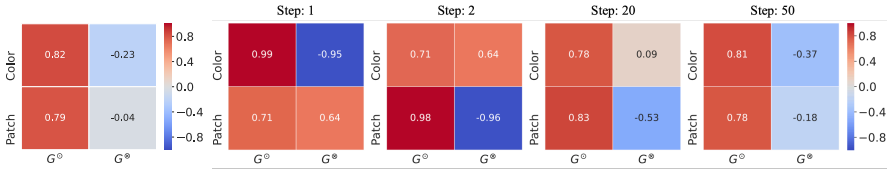
where m *denotes the cardinality of* $\mathcal{G}$*.*

Proof. See Supplementary Material.

The bound on the worst-case risk gap, provided by Theorem 1, sheds light on the inherent differences between deterministic grouping in group DRO and our adaptive bias discovery. Furthermore, this theorem suggests that if ABD successfully identifies biases, it can achieve performance comparable to group DRO, which relies on prior bias knowledge. This underscores the intuition that the algorithm's effectiveness depends on its ability to discover biases. In line with this intuition, our empirical experiments, presented in the next section, demonstrate that ABD can dynamically identify biases.

Table 1: Test accuracy (%) of different algorithms on the Colored MNIST task over five trials (mean ± standard deviation). The value δ_{gap} indicates the generalization gap between i.i.d and OoD test environments.

Algorithm	Bias: Color			Bias: Color & Patch		
	Test (i.i.d) $p_e = 0.1$	Test (OoD) $p_e = 0.9$	δ_{gap}	Test (i.i.d) $p_e = 0.1$	Test (OoD) $p_e = 0.9$	δ_{gap}
ERM	88.6 ± 0.3	16.4 ± 0.8	-72.2	93.7 ± 0.3	14.0 ± 0.5	-79.7
IRM	71.4 ± 0.9	66.9 ± 2.5	-4.5	93.5 ± 0.2	13.4 ± 0.3	-80.1
Group DRO	89.2 ± 0.9	13.6 ± 3.8	-75.6	92.3 ± 0.3	14.1 ± 0.8	-78.2
PI	70.3 ± 0.3	70.2 ± 0.9	**-0.1**	85.4 ± 0.9	15.3 ± 2.7	-70.1
ABD (Ours)	70.5 ± 1.1	**70.7 ± 1.4**	**0.2**	68.3 ± 2.3	**62.3 ± 3.3**	**-6.0**
Optimal	75.0	75.0	0.0	75.0	75.0	0.0



(a) PI **(b)** ABD (Ours)

Fig. 2: Visualization of the Pearson correlation coefficients between the labels and bias features in each created group on the Colored MNIST task. ABD adaptively identifies diverse biases through the learning process while PI is only sensitive to a specific bias feature, *Color*.

4 Experimental Results

Our experiments aim to answer the following questions:

- Does ABD successfully address multiple biases through various learning dynamics, compared to conventional approaches?
- How effective is ABD in mitigating performance degradation on real-world tasks, compared to prior state-of-the-art baselines?

To answer those questions, we experiment with our framework on various tasks including synthetic tasks, Colored MNIST, and real-world applications such as MetaShift [22], Camelyon17-wilds [3,21], FMoW-wilds [10,21], CivilComments-wilds [9, 21], and MultiNLI [35] tasks.

For controlled experiments with synthetic datasets, we compare ABD with methods that utilize biased models trained independently of the main model, such as Predict then Interpolate (PI) [4], and Group DRO, which optimizes the worst-case group over a predefined group set. For evaluations on real-world applications, we compare our framework with diverse algorithms, including CORAL [33]

which aligns feature distributions across domains, Invariant Risk Minimization (IRM) [1] aimed at learning invariant features, and recent approaches like Just Train Twice (JTT) [23] which also utilizes fully-trained biased models. We also compare ABD with recent approaches like Common Gradient Descent (CGD) [27], and LISA [36], the latter being a state-of-the-art method for the WILDS dataset [21].

4.1 Synthetic Task with Multiple Biases

In this section, we demonstrate the capability of our framework in identifying multiple biases using a synthetic task based on the Colored MNIST dataset [1]. We introduce further complexity by adding an additional bias feature to the Colored MNIST setting. Specifically, we inject small patches of noise at the corners of the images, ensuring that both the colors and locations of these patches exhibit strong correlations with the labels. In the setting of Colored MNIST, we consider two training environments with noise probabilities for bias features, $p_e = 0.1$ and $p_e = 0.2$, which correspond to 90% and 80% of bias features matching the labels, respectively. Additionally, label noise is added to the shape attribute with probability $p_c = 0.25$, resulting in a reduced correlation with the target label; only 75% of the shape features match the target.

For the algorithm evaluation, we consider both independent and identically distributed (i.i.d) data and a distributionally shifted (or OoD) test environment with $p_e = 0.9$ where the correlation between color and label is reversed compared to the training environments. The model that exploits *Color* or *Patch* attributes for prediction can achieve high accuracy on the i.i.d test environment, but it will fail on the OoD test. The purpose of the Colored MNIST task is to classify digits solely based on the shape, without relying on other bias features such as colors and patches, in order to achieve robust generalization.

Results. Tab. 1 exhibits the evaluation results on the Colored MNIST task according to the types of synthetic bias, only with *Color* and with both *Color* and *Patch*, respectively. Since the experimental setting does not provide explicit bias information, group DRO cannot achieve robustness for distributional shifts, a limitation stemming from its reliance on specific bias information. In all cases, ABD achieves state-of-the-art OoD performance on Colored MNIST. For the case with multiple types of biases (*Color* & *Patch*), PI shows degraded performance on the OoD test environment not better than ERM, while ABD outperforms PI by 47 pp. This clearly demonstrates that ABD can more effectively de-bias multiple types of spurious correlations in the dataset.

Analysis. To verify the effectiveness of adaptive bias discovery for datasets with multiple underlying biases, we measure Pearson correlation coefficients between the label and bias features (*Color*, *Patch*) of groups $G^{\odot}$ and $G^{\otimes}$ and visualize them in Fig. 2. Compared to PI, which relies on prebuilt biased models, ABD discovers multiple biases based on its adaptive bias discovery, dynamically

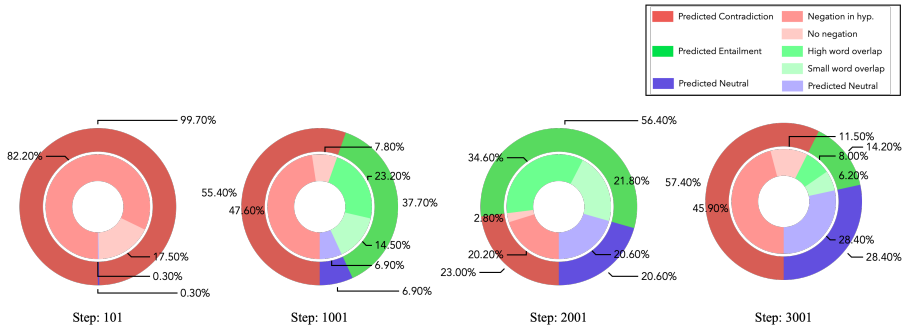


Fig. 3: Visualization of the composition of bias features for the misclassified group $G^{\otimes}$.

producing new data partitions for each training step, as shown in Fig. 2b. Specifically, PI mainly focuses on the *Color* instead of *Patch* as shown in Fig. 2a. ABD captures *the Color* bias during the first learning step and mitigates its effect. As a result, during the second learning step, ABD is able to discover and address *Patch* bias, resulting in the superior performance on multiple bias setting as shown in Tab. 1.

4.2 Real-World Tasks

In this section, we evaluate our framework on large-scale practical applications. Since the annotation of bias features is not always available for real-world datasets, we consider two cases for these experiments: datasets with bias annotations, allowing group DRO to leverage this information, and datasets without bias knowledge.

Dataset with Bias Information In this experimental setting, we compare ABD, which does not utilize bias knowledge, to group DRO that operates with bias-annotated groups. We utilize two datasets: CivilComments-wilds and MultiNLI. CivilComments-wilds [9, 21] serves as a benchmark for measuring the impact of biases in a binary text toxicity classification task. Specifically, biases relating to comments that mention particular demographic identities have been observed to spuriously associate toxicity with those demographics, leading to degraded model performance on other subpopulations. Thus, robustness against biases is measured by the worst-case test accuracy across 16 distinct demographic groups. MultiNLI [35] is a large-scale natural language dataset composed of sentence pairs (premise and hypothesis) and their corresponding textual entailment labels. The main goal of the task is to predict the relationship between a premise and a hypothesis as one of *entailment*, *contradiction*, or *neutral*. There are two types of well-known biases in the MultiNLI dataset: the first is that the presence of negation words is highly correlated with the label *contradiction* [15, 28],

Table 2: Comparison of worst-case test accuracy (%) for various algorithms on the CivilComments-wilds dataset. The *Group* column details the demographic information used by each algorithm for grouping.

Algorithm	**Worst-case accuracy**	**Group**
ERM	56.0	*None*
IRM	66.3	*(label × Black)*
Group DRO	69.1	*(label)*
Group DRO	70.0	*(label × Black)*
JTT	69.3	*None*
PI	61.1	*None*
ABD (Ours)	**71.1**	*None*

Table 3: Worst-group and its test accuracy (%) for each algorithm on MultiNLI dataset. The worst-group indicates worst-group information as {*Label*, *Negation*, *Overlap*}, for example, {*neu., neg., sml.*} means a group with label *neutral*, negation words in hypothesis, and small overlap between premise and hypothesis. Note that the result of Group DRO* is evaluated with groups hand-crafted using prior knowledge of biases.

Algorithm	**Worst-case acc.**	**Worst-group**
ERM	61.8	{*neu., neg., sml.*}
Group DRO	62.7	{*neu., neg., sml.*}
JTT	63.2	{*neu., neg., sml.*}
PI	61.5	{*ent., neg., sml.*}
ABD (Ours)	**67.1**	{*ent., neg., sml.*}
Group DRO*	67.5	{*neu., neg., sml.*}

and the second is significant word overlap between the premise and hypothesis, highly correlated with the label *entailment* [25].

Results. The experimental results for CivilComments-wilds and MultiNLI are presented in Tab. 2 and Tab. 3, respectively. As the results demonstrate in Tab. 2, ABD is able to outperform group DRO, even without access to any bias information. In the results shown in Table 3, group DRO exhibits almost no performance improvement compared to ERM when the prior knowledge of biases is not given. Additionally, PI fails to achieve robust generalization compared to group DRO with hand-crafted groups, as the dataset contains multiple intricate biases. In contrast, our ABD achieves the highest performance, yielding results close to those of group DRO using ground truth group annotations.

Table 4: OoD test accuracy (%) on the Camelyon17-wilds (average accuracy) and FMoW-wilds (worst-region accuracy).

Algorithm	Camelyon17-wilds	FMoW-wilds
ERM	70.3 ± 6.4	32.3 ± 1.3
IRM	59.5 ± 7.7	31.7 ± 1.2
Group DRO	68.4 ± 7.3	30.8 ± 0.8
CORAL	59.5 ± 7.7	32.8 ± 0.7
JTT	63.8 ± 1.4	33.4 ± 0.9
PI	71.7 ± 7.5	31.2 ± 0.3
CGD	69.4 ± 7.9	32.0 ± 2.3
LISA	77.1 ± 6.5	**35.5 ± 0.7**
ABD (Ours)	**81.1 ± 4.8**	34.1 ± 2.5

Analysis. We also analyze the prediction results of an intentionally biased model to demonstrate the performance of ABD in discovering multiple biases. Specifically, we analyze data samples in the group $G^{\otimes}$, which contains wrongly predicted instances, according to their predicted class and bias annotations. In Fig. 3, each outer circle of the chart represents predictions of the biased model f_ϕ, and each inner circle shows the proportion of samples containing bias features for the corresponding class. In the early stage of training, our method discovers the most prominent bias, *negation*. However, as the training progresses, the ratio of the negation bias decreases, and another bias, termed *large overlap* begins to occupy more portions in the group $G^{\otimes}$. This analysis supports our conclusion that ABD's adaptive bias discovery can progressively uncover multiple underlying biases in the dataset, starting with the most pronounced and then revealing subtler biases as training continues.

Dataset without Bias Information In scenarios where bias information is not annotated within the dataset, understanding and mitigating biases become more complex tasks. These cases are common in real-world applications, where recognizing and handling biases can be challenging. In this experimental setting, we evaluate the efficacy of ABD in managing datasets that lack explicit bias knowledge.

To illustrate the capability of ABD in such scenarios, we utilize two datasets: Camelyon17-wilds and FMoW-wilds. Camelyon17-wilds [3,21] and FMoW-wilds [10, 21] datasets serve as benchmarks for OoD generalization, with the main goal of classifying real-world image data.

Results. Tab. 4 presents the experimental results on the Camelyon17-wilds and FMoW-wilds datasets. In this setting, where bias knowledge is not available, group DRO exhibits degraded performance, even falling below that of ERM. Additionally, PI fails to achieve robust performance in these complex real-world

Table 5: Test accuracy (%) across diverse distributional shifts in the MetaShift benchmark, where distance indicates the distributional difference between training and test domains.

Distance	**0.44**	**0.71**	**1.12**	**1.43**
ERM	80.1	68.4	52.1	33.2
IRM	79.5	67.4	51.8	32.0
Group DRO	77.0	68.9	51.9	34.2
LISA	**81.3**	69.7	54.2	37.5
ABD (Ours)	80.4	**71.8**	**55.2**	**41.8**

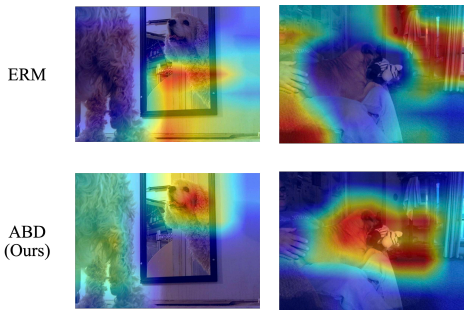


Fig. 4: GradCAM visualizations of trained model with ERM and ABD on the MetaShift test data for distributional distance is 1.43.

applications, as it cannot be assumed that only one superficial bias in the dataset. Compared to LISA, the state-of-the-art baseline for the WILDS benchmark [21], our ABD demonstrates better performance on Camelyon17-wilds and comparable performance on FMoW-wilds.

4.3 Robustness for Distributional Shifts

In this section, we evaluate the performance of ABD across various distributional shift scenarios using the MetaShift [22] dataset. MetaShift enables the assessment of performance degradation in dogs and cats classification tasks due to discrepancies between training and test distributions. Table 5 presents the performance of different learning frameworks under varying degrees of distributional distance. Our observations indicate that ABD significantly outperforms all other methods.

GradCAM Visualizations. We investigate GradCAM visualizations of the main classifier, f_θ, and the biased model, f_ϕ, trained with ABD, as shown in Figure 4

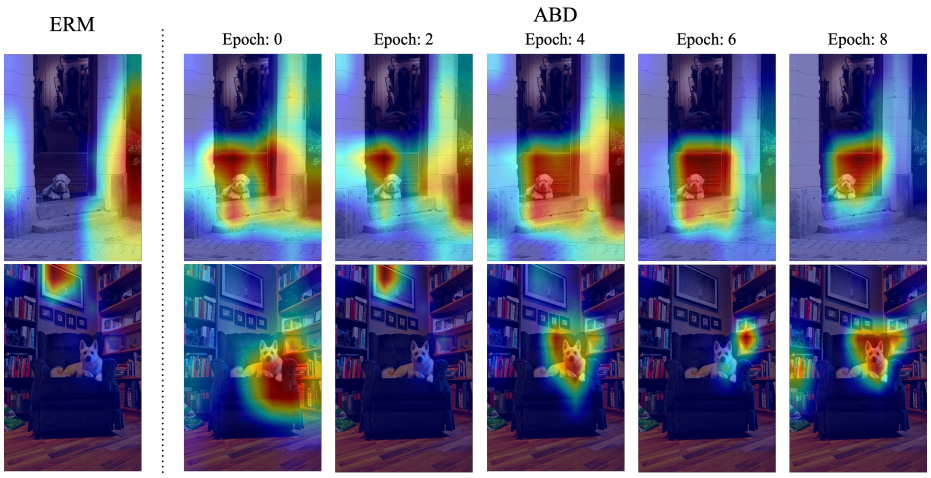


Fig. 5: GradCAM visualizations of ERM-trained model and biased models, f_ϕ, trained with ABD on the MetaShift test data for diverse learning steps.

and 5, respectively. As demonstrated in Figure 4, ERM tends to rely on background features for predicting dogs, leading to lower performance, whereas ABD bases its predictions more on the object itself, resulting in more accuracte classifications. Additionally, we observe that the biased model, trained with adaptive bias discovery strategy, focuses on diverse features according to the learning epochs, as shown in Figure 5.

5 Conclusion

We propose ABD, a novel end-to-end debiasing learning framework, designed to autonomously address biases in datasets. Our empirical results demonstrate that ABD can effectively identify and mitigate multiple unidentified biases during the training process, without requiring prior knowledge of these biases. Extensive evaluations on both synthetic and real-world experiments validate the efficacy and robustness of ABD across various scenarios. We believe that our work offers a promising new direction for future research on addressing complex bias features in datasets, potentially improving the fairness and generalization of deep neural networks in real-world applications.

Acknowledgements This research was supported by the Core Research Institute Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03043144) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2023-00241123).

References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019) 4, 8, 9
2. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017) 5
3. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging **38**(2), 550–560 (2018) 8, 12
4. Bao, Y., Chang, S., Barzilay, R.: Predict then interpolate: A simple algorithm to learn stable classifiers. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 640–650. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/bao21a.html 2, 3, 8
5. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 456–473 (2018) 1
6. Belinkov, Y., Poliak, A., Shieber, S.M., Van Durme, B., Rush, A.M.: Don't take the premise for granted: Mitigating artifacts in natural language inference. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 877–891 (2019) 3
7. Belinkov, Y., Poliak, A., Shieber, S.M., Van Durme, B., Rush, A.M.: On adversarial removal of hypothesis-only bias in natural language inference. In: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019). pp. 256–262 (2019) 3
8. Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Management Science **59**(2), 341–357 (2013) 2
9. Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L.: Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference. pp. 491–500 (2019) 8, 10
10. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180 (2018) 8, 12
11. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4069–4082 (2019) 3
12. Duchi, J.C., Glynn, P.W., Namkoong, H.: Statistics of robust optimization: A generalized empirical likelihood approach. Mathematics of Operations Research **46**(3), 946–969 (2021) 2
13. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. pp. 1126–1135. PMLR (2017) 6
14. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018) 1

15. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A.: Annotation artifacts in natural language inference data. In: 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 (2018) 10
16. de Haan, P., Jayaraman, D., Levine, S.: Causal confusion in imitation learning. In: Advances in Neural Information Processing Systems. pp. 11698–11709 (2019) 1
17. He, H., Zha, S., Wang, H.: Unlearn dataset bias in natural language inference by fitting the residual. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). pp. 132–142 (2019) 3
18. Hu, W., Niu, G., Sato, I., Sugiyama, M.: Does distributionally robust supervised learning give robust classifiers? In: International Conference on Machine Learning. pp. 2029–2037. PMLR (2018) 2, 3
19. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems. pp. 125–136 (2019) 1
20. Kim, N., Hwang, S., Ahn, S., Park, J., Kwak, S.: Learning debiased classifier with biased committee. Advances in Neural Information Processing Systems **35**, 18403–18415 (2022) 2, 3
21. Koh, P.W., Sagawa, S., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning. pp. 5637–5664. PMLR (2021) 1, 8, 9, 10, 12
22. Liang, W., Zou, J.: Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=MTex8qKavoS 8, 13
23. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: International Conference on Machine Learning. pp. 6781–6792. PMLR (2021) 3, 8
24. Mahabadi, R.K., Belinkov, Y., Henderson, J.: End-to-end bias mitigation by modelling biases in corpora. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8706–8716 (2020) 3
25. McCoy, T., Pavlick, E., Linzen, T.: Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3428–3448 (2019) 10
26. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems **33**, 20673–20684 (2020) 2, 3
27. Piratla, V., Netrapalli, P., Sarawagi, S.: Focus on the common good: Group distributional robustness follows. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=irARV_2VFs4 9
28. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B.: Hypothesis only baselines in natural language inference. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 180–191 (2018) 10
29. Sagawa*, S., Koh*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxGuJrFvS 2, 3, 5

30. Sanh, V., Wolf, T., Belinkov, Y., Rush, A.M.: Learning from others' mistakes: Avoiding dataset biases without modeling them. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=Hf3qXoiNkR 2, 3
31. Schuster, T., Shah, D., Yeo, Y.J.S., Ortiz, D.R.F., Santus, E., Barzilay, R.: Towards debiasing fact verification models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3419–3425 (2019) 3
32. Scimeca, L., Oh, S.J., Chun, S., Poli, M., Yun, S.: Which shortcut cues will DNNs choose? a study from the parameter-space perspective. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=qRDQi3ocgR3 2, 3
33. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016) 8
34. Utama, P.A., Moosavi, N.S., Gurevych, I.: Towards debiasing nlu models from unknown biases. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7597–7610 (2020) 2, 5
35. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: 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 (2018) 8, 10
36. Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., Finn, C.: Improving out-of-distribution robustness via selective augmentation. In: International Conference on Machine Learning. pp. 25407–25437. PMLR (2022) 9
37. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: HellaSwag: Can a machine really finish your sentence? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4791–4800. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1472, https://aclanthology.org/P19-1472 5