Let Samples Speak: Mitigating Spurious Correlation by Exploiting the Clusterness of Samples

Weiwei Li[1] Junzhuo Liu[1] Yuanyuan Ren[2] Yuchen Zheng[2] Yahao Liu[1] Wen Li[1*]

[1]University of Electronic Science and Technology of China
[2]Shihezi University
{davelee.uestc,junzhuo.cs,lyhaolive,liwenbnu}@gmail.com, {yuanyuanren043,ouczyc}@outlook.com

Abstract

*Deep learning models are known to often learn features that spuriously correlate with the class label during training but are irrelevant to the prediction task. Existing methods typically address this issue by annotating potential spurious attributes, or filtering spurious features based on some empirical assumptions (*e.g.*, simplicity of bias). However, these methods may yield unsatisfactory performance due to the intricate and elusive nature of spurious correlations in real-world data. In this paper, we propose a data-oriented approach[1] to mitigate the spurious correlation in deep learning models. We observe that samples that are influenced by spurious features tend to exhibit a dispersed distribution in the learned feature space. This allows us to* identify *the presence of spurious features. Subsequently, we obtain a bias-invariant representation by* neutralizing *the spurious features based on a simple grouping strategy. Then, we learn a feature transformation to* eliminate *the spurious features by aligning with this bias-invariant representation. Finally, we* update *the classifier by incorporating the learned feature transformation and obtain an unbiased model. By integrating the aforementioned identifying, neutralizing, eliminating and updating procedures, we build an effective pipeline for mitigating spurious correlation. Experiments on image and NLP debiasing benchmarks show an improvement in worst group accuracy of more than 20% compared to standard empirical risk minimization (ERM).*

1. Introduction

Recent studies reveal that deep neural networks (DNNs) learn unintended decision rules from spurious correlations [15], also known as model bias. For example, researchers [7, 16, 36, 39, 52] report that models infer disease using cues of medical devices rather than symptoms. The sampling bias is blamed for introducing spurious correlations between attributes and class labels in data [2, 6, 47], as illustrated in Fig. 1. However, such unexpected biases are often masked by satisfying performance on i.i.d. test data [11, 44], and hard to annotate in advance, making methods [25, 40, 48] requiring bias attributes or labels less practical.

One possible way of addressing unknown model bias is to eliminate bias-relating features (known as spurious fea-

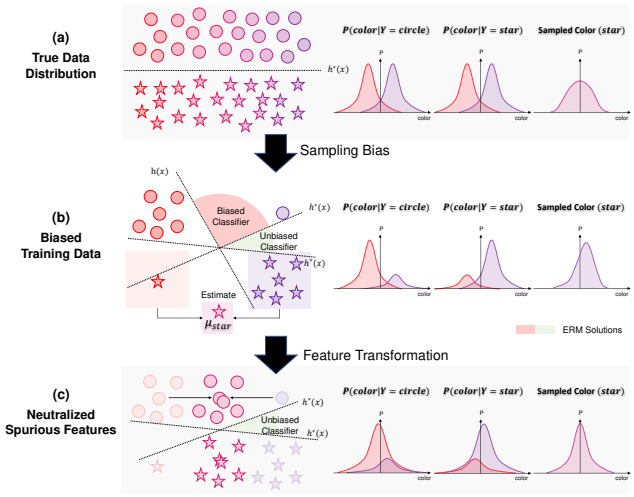


Figure 1. 1) Ideally, bias attributes (*e.g.*, color) should be evenly distributed and non-predictive of the class; 2) Sampling bias can introduce unintended patterns, like most circles being red and most stars being purple, causing some features to mistakenly correlate with class labels. Since ERM training minimizes the mean loss, an ERM-trained model is highly likely to fit these spurious correlations due to their large population in the data; 3) Intuitively, a transformation producing invariant representation for different values of bias attributes reduces the possible of learning bias.

*The corresponding author

[1]Codes and checkpoints are available at https://github.com/davelee-uestc/nsf_debiasing.

tures) to obtain a bias-invariant representation so that the classifier does not rely on them. The key challenge of such methods is to make proper assumptions of possible spurious features. Some methods [28, 46] use assumptions from the empirical manifestation of biases such as assuming simpler features, like color, to be spurious. Although seemingly straightforward, this assumption is task-specific, and has a clear limitation: they fail if the task changes, for example, changing from classifying numbers to colors in the Colored-MNIST dataset [3].

Instead, we estimate the bias-invariant representation using the fact that samples affected by spurious features tend to exhibit a dispersed distribution in the feature space. Specifically, we propose a task-independent assumption of spurious features, termed the strong spurious assumption, to estimate the bias-invariant representation without bias attributes or labels. This assumption is supported by recent findings: while DNNs produce both core and spurious features [19], DNNs favor strong features (high availability), regardless of whether they are core (predictive) or spurious (bias-related) [25, 41], meaning that the model becomes biased only when spurious features are stronger. This strong spurious assumption defines a characteristic in feature space for same class samples with different spurious feature values: the samples of majority values are closer to their centroid while the minority deviates. As illustrated in Fig. 1 (b), when the bias is most circles are red while most stars are purple, stronger spurious feature (color) makes purple circles closer to star class centroid rather than circle class centroid while the red circles don't, making samples of different spurious attribute separable. This characteristic enables us to (1) use the presence of a minority sample (purple circles) as the indicator of spurious features; (2) distinguish the majority (red circles) and minority groups (purple circles) and estimate a bias-invariant representation using the groups found. We provide theoretical proof using a widely accepted debiasing data model [41] for binary classification.

Building on this foundation, we introduce Neutralizing Spurious Features (NSF), a debiasing method that does not require prior knowledge of bias attributes. NSF consists of four key steps: *(1) Identifying Bias Presence*: Minority samples that deviate from the class centroid are identified, as such deviations indicate the presence of spurious features. *(2) Neutralizing Spurious Feature for Bias-Invariant Features*: Use identified groups to estimate a bias-invariant representation for each class. *(3) Eliminating Spurious Feature*: Learn a common transformation across all classes that aligns all training samples within a class to the estimated bias-invariant features. This transformation eliminates spurious features while preserving core features. *(4) Updating Classifier*: Finetune the classifier on these bias-invariant features, forcing reliance on core features alone.

To validate the effectiveness of the proposed method, we conduct experiments on multiple popular benchmarks. Experiments across four image and text tasks with known spurious correlations and one medical dataset show an average improvement of 20% in Worst Group Accuracy (WGA) compared to Standard training via ERM, achieving state-of-the-art with a very fast speed (within a few minutes). We performed ablation studies and qualitative analysis to validate the key components and intuition.

The contributions of this paper are as follows.

- We leverage the separable of samples affected by spurious features in feature space, enabling an estimation method for bias-invariant representation. We provide theoretical proof of its correctness.
- We introduce a novel, non-intrusive debiasing framework with four steps—bias identification, neutralization, spurious feature elimination, and classifier update—enabling robust learning on core features without bias labels.
- Extensive experiments across multiple challenging benchmarks and comprehensive ablation studies validate our approach, demonstrating its state-of-the-art performance, effectiveness of individual components.

2. Related Works

Removing Known Spurious Correlations Spurious correlations are common in real-world datasets [15] due to natural relationships [15] or selection bias [47]. A common assumption is that bias attributes are known and labeled. This ideal assumption is the basis of the group-robustness methods [40] and retraining-based methods [8, 12, 25], which are the upper bound of debiasing methods, despite the impractical of labeling bias. A more reasonable assumption is that only the bias attributes are known, and we do not have access to the bias label. The task-specified biases mostly explored are that low-level features [5, 9, 10, 14, 18, 23, 24, 33, 34] such as the color and texture, and background shortcut [28, 42], these biases are common in image classification. Additionally, some work [1] uses zero-shot insights from language models to obtain priors of harmful biases. Leveraging these priors of specific biases, methods such as data augmentation, can be designed and applied to address specific biases.

Removing Unknown Spurious Correlations Biases vary across tasks and manifest as unexpected patterns, making unknown biases a more common problem. As we know nothing about the bias, an assumption of bias is required. One representative assumption is the simplicity bias [45, 46], which refers to features that are easy-to-learn or learned at the early training phase are more likely to be a bias. It is also common to assume that an ERM-trained model is biased and leverages the learned bias for debiasing. Methods such as Lff [34] and Echoes [21] learn a model that differs from the biased model by minimizing the mutual alignment of the two. It is also natural to extend this idea

to learning a set of diverse hypotheses, as seen in [27, 45]. Furthermore, some works [9, 29] use bias-adversarial augmentation for model debiasing. Digging bias-conflicting samples is another solution for addressing unknown biases, considering that the spurious correlations could be effectively eliminated if the effect of the minority groups, known as the bias-conflicting samples, were amplified. A straightforward idea is reweighting those bias-conflicting samples [35, 44], or reforming the representation space using those samples [20, 51, 53]. However, those methods usually rely on an empirical set of bias-conflicting samples such as false positives, and require to retrain the model. The limitations of these methods arise because these assumptions use indirect manifestations of the biases and do not capture the intrinsic nature of the biases, thus being costly in training and yielding unsatisfactory performance. In contrast, the proposed method directly eliminates the spurious features by adding a single linear transformation after the frozen encoder, leveraging a derived conclusion from the bias-fitting mechanism.

3. Problem Statement

We start by defining our problem setting. Formally, let $\mathcal{D} = \{(\vec{x}_i, y_i, g_i)\}_{i=1}^n$ denote the dataset, where $\vec{x}_i$ is the input feature, y_i is the corresponding label, and $g_i = (a_i, y_i) \in G$ is the corresponding group defined by the label y and a spurious attribute $a \in A$ that spuriously correlates with the label (i.e., $G = A \times Y$).

3.1. Masked Poor Generalization on Minorities

The ERM is a common practice of training DNNs, which aims to find the hypothesis $h^* \in \mathcal{H}$ that minimizes the empirical risk under the loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$:

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(\vec{x}_i), y_i), \tag{1}$$

where $\mathcal{H}$ is the hypothesis class, $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces respectively. The ERM selects a model from a hypothesis class that minimizes average loss over training data, and results in majority groups having a greater impact, so the ERM makes a good average performance, but underperforms for minority groups [48].

Learning Goal: Optimizing Worst Group Accuracy Our goal is to learn a model $h : \mathcal{X} \to \mathcal{Y}$ that maximizes the least accurately predicted subgroup, also known as Worst Group Accuracy (WGA). This goal ensures that the performance of minority groups is taken into account. The WGA is defined as:

$$\text{Acc}_{wg}(h) = \min_{j \in \{1,2,\ldots,k\}} \frac{1}{|\mathcal{D}^j|} \sum_{(\vec{x}_i, y_i) \in \mathcal{D}^j} \mathbb{1}[h(\vec{x}_i) = y_i], \tag{2}$$

where $\mathcal{D}^j$ denote j-th group, $\mathbb{1}[\cdot]$ is the indicator function, h is the classifier.

3.2. The Challenge of Unknown Biases

Considering the ERM objective, poor generalization on minority groups indicates that the model learns patterns aligned with the majority group but not consistently true within each class. These spurious correlations—valid for the majority but not the minority—lead models to make inaccurate predictions for minority groups [15].

If the groups within the dataset are known, we can address the issue of poor performance on minority groups by using group reweighting or resampling such as [25, 40]. These techniques ensure that the model pays equal attention to all groups, thereby improving performance on the least accurately predicted subgroup. However, in many real-world scenarios, the groups are not explicitly labeled, and how the groups formed is unknown, making it hard to improve the performance of the model on minorities.

In summary:

- ERM minimizes average loss over the dataset but often overfits to spurious features prevalent in majority group data, masking poor generalization on minority groups.
- Models fitting spurious correlations cause incorrect predictions for minority groups. However, in many real-world scenarios, the groups are not explicitly labeled, and how the groups formed is unknown, making it hard to apply group reweighting or resampling techniques.

4. Proposed Method

4.1. Overview

To address the model bias, especially considering the bias attributes are hard to foresee before the model is trained, we propose NSF, a novel debiasing method that does not require prior knowledge of bias attributes. We develop a method to estimate mean values of the true data distribution without accessing the bias labels. Thus, NSF can eliminate the spurious features by transforming features to align training samples with estimated unbiased mean values of the true data distribution, and debias the classifier through fine-tuning, as in Fig. 2.

4.2. The Biased-Sampled Data Model

To address these spurious correlations, it is essential to understand their origins. Here, we use the widely used [8, 12, 41] assumption of sampling bias. It refers to certain members of a population being systematically more likely to be selected than others, leading to non-representative samples [2]. Sampling bias can cause labels to mistakenly correlate with a specific attribute because the samples are not representative of the entire population, as in Fig. 1. We adopt a data generation process from [41] to model the joint

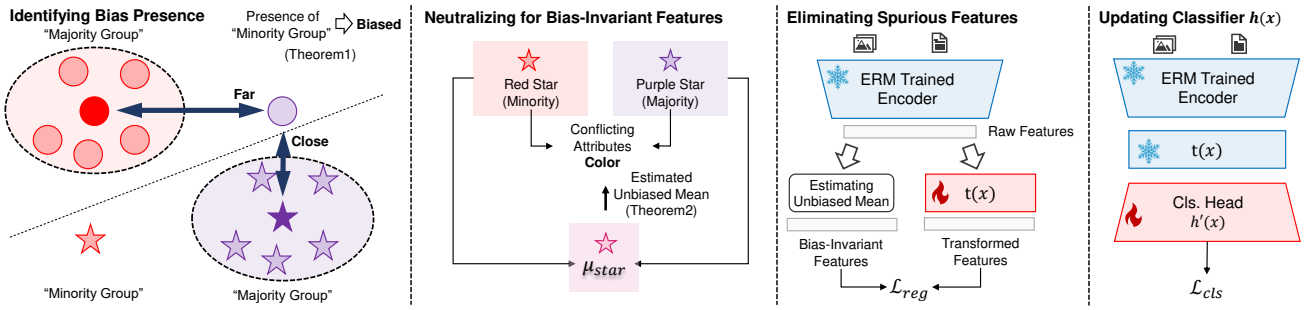


Figure 2. NSF leverages a task-independent strong spurious assumption, enabling us to (1) use the presence of minority sample (purple circles) as the indicator of spurious features; (2) distinguish majority (purple stars) and minority groups (red stars) and estimate a bias-invariant representation using the groups found. NSF mitigates biases by (3) first eliminating the bias attributes by transforming features to align training samples with estimated unbiased mean values of the true data distribution, then (4) debiasing classifiers through fine-tuning.

data distribution $(X_\rho, Y_\rho, A_\rho) \sim p_\rho$ under spurious correlation. The label $y \in Y_\rho$ follows the Uniform distribution over $\{1, -1\}$, the data point $\vec{x} = [Ba, y, \vec{\delta}] \in X_\rho$ and the spurious feature $a \in A_\rho$ are generated as follow:

$$a \sim \begin{cases} P(a = k|y = k) = \rho \\ P(a = -k|y = k) = 1 - \rho \end{cases}, \vec{\delta} \sim \mathcal{N}(\vec{0}, \vec{I}^{D-2}),$$

where $\mathcal{N}$ is the normal distribution, D is the dimension of $\vec{x}$, $\rho \in (0.5, 1)$, and $B \geq 1$ is scalar constants. And The conditional expectation of $\vec{x}$ is

$$\mathbb{E}(\vec{x}|y = k) = [(2\rho - 1)Bk, k, \vec{0}]. \tag{3}$$

This data generation process models the characteristics of the bias-sampled data, if $\rho = 0.5$, then it means that the sampling is fair, a has no correlation with the label, or else the sampling procedure is biased toward a specific value. This creates a majority group with a high concentration of a particular value, while minority groups, known as bias-conflicting samples, form a small population [47]. In this data model, the input $\vec{x}$ consists of core feature y, spurious feature a and noise $\vec{\delta}$:

- **Spurious Feature** Ba : The spurious feature a correlate with the label y with a probability of ρ. The scalar constant B controls the impact of the bias attribute a. The bias-sampled data has a biased conditional expectation $(2\rho - 1)Bk$ for the spurious feature, and it is zero for the true data distribution ($\rho = 0.5$).
- **Core Feature** y : The core features of samples only relate to their class labels, so the label y is used as the core feature since the informative is equivalent. Its conditional expectation is independent of ρ and B.
- **Noise** $\vec{\delta}$: Other features that not correlate with the labels. Its conditional expectation is independent of ρ and B.

So we can find that the model fits a data distribution that deviates from the true data distribution due to the sampling bias, more specifically, in the spurious feature.

4.3. Confirming Bias Presence

In this section, we demonstrate that if the spurious features are sufficiently strong, the presence of bias can be confirmed using the relative distance to the class centroid. We start with the definition of C_k^ρ and the relative distance $d(\vec{x}_i)$. The conditional mean $C_k^\rho = \mathbb{E}[X_\rho|Y = k]$, also known as the centroid, can be estimated as

$$C_k^\rho = \frac{1}{\sum_{i=1}^N \mathbb{1}[\vec{y}_i = k]} \sum_{i=1}^N \mathbb{1}[\vec{y}_i = k] * \vec{x}_i, \tag{4}$$

where $\vec{x}_i$ is the feature. $\forall(\vec{x}_i, y_i) \in p_\rho$, the relative distance between $\vec{x}_i$ and its corresponding centroid $C_{y_i,\rho}$, compared to the nearest centroid of another class is given by $d(\vec{x}_i, \rho) = (\vec{x}_i - C_{y_i}^\rho)^2 - (\vec{x}_i - C_{q_i}^\rho)^2$, where $\vec{q}_i = \arg\min_{u \neq y_i}\{(\vec{x}_i - C_u^\rho)^2\}$.

Based on the data model, we found that if some samples significantly deviate from the mean values, indicating the existence of spurious features, these samples are in the minority group. This assumption is natural for the bias-sampled data, taking the example of classifying circles v.s. stars, the color is a strong spurious feature. However, this spurious correlation isn't valid for all samples due to exceptions like purple circles, noticing the bias is that most circles are red. Using this conclusion, we can separate the circles of different spurious features (red and purple) in the feature space. Based on this insight, we propose a method for separating the minority group from the majority group by the relative distance $d(\vec{x}_i, \rho)$ between data points and sample means, as

Theorem 1 *If* $1 - (2\rho - 1)^2 B^4 < 0$, *then*

$$\forall \vec{y}_i = \vec{y}_j, d(\vec{x}_i, \rho) \times d(\vec{x}_j, \rho) < 0 \iff a_i \neq a_j$$

For detailed proof, please refer to the Appendix. Theorem 1 implies that we can confirm the existence of spurious feature $(a_i \neq a_j)$ by checking if any samples are deviating

from their conditioned sample mean (pair of instance i, j satisfying $d(\vec{x}_i, \rho) \times d(\vec{x}_j, \rho) < 0$).

4.4. Estimating Bias-Invariant Representation

To mitigate the impact of spurious features, a feasible solution is to neutralize them to obtain a bias-invariant feature, which requires estimating the unbiased mean of spurious features depending on the inaccessible bias label a. However, Theorem 1 implies that, if the spurious feature a is strong enough, we can separate the minority group from the majority group using the sign of relative distance $d(\vec{x}_i)$, without knowing the exact value of a. This enables us to estimate the unbiased conditional mean $C_k = C_k^{0.5}$ of $\vec{x}$ in the true data distribution. We start by identifying the majority and minority groups in each class.

Identifying Majority and Minority Groups We make a soft-assignment $q_i \in Q$ for each point as $q_i = \arg\min_u ||\vec{x}_i - C_u^\rho||_2$. And it has $y_i = q_i \Rightarrow d(\vec{x}_i, \rho) < 0$, and $y_i \neq q_i \Rightarrow d(\vec{x}_i, \rho) > 0$. Then, we can split class X_k into $U_k = \{\vec{x}_i \mid q_i \neq y_i, (\vec{x}_i, y_i) \in p_\rho\}$ and $V_k = \{\vec{x}_i \mid q_i = y_i, (\vec{x}_i, y_i) \in p_\rho\}$ by Q. Since not all class k satisfies that $|U_k| > 0$ and $|V_k| > 0$ (when the spurious features not strong enough), which indicates that we cannot estimate C_k for those classes. We exclude those classes with a mask $o \in O$ as $o_i = (|U_{y_i}| > 0) \wedge (|V_{y_i}| > 0)$. **Estimating Bias-Invariant Representation** Using the groups formed by the sign of relative distance $d(\vec{x}_i, \rho)$, we can estimate the value of C_k as follows.

Theorem 2 *If* $1 - (2\rho - 1)^2 B^4 < 0$, *then*

$$C_k = \mathbb{E}\left(\frac{1}{2|U_k|}\sum_i^{|U_k|} \vec{u}_i + \frac{1}{2|V_k|}\sum_j^{|V_k|} \vec{v}_j\right)$$

where $U_k = \{\vec{x} \mid (\vec{x}, y) \in p_\rho, y = k, d(\vec{x}, \rho) > 0\}$, $V_k = \{\vec{x} \mid (\vec{x}, y) \in p_\rho, y = k\} \setminus U_k$, $\vec{u}_i \in U_k$, $\vec{v}_i \in V_k$.

For detailed proof, please refer to the Appendix.

4.5. Eliminating Unknown Spurious Features

Using the estimated unbiased mean $C_y = u(a, y, \vec{\delta})$ of the true data distribution, spurious features can be eliminated, as in Fig. 3, by learning a channel-wise transformation $t(\vec{x}) = \vec{w}(\vec{x} - \vec{b}) + \vec{b}$ where $\vec{w} \in R^{1\times D}$ and $\vec{b} \in R^{1\times D}$ to make all data points close to their corresponding conditioned mean value.

We here provide a simple verification that the corresponding channels of core features are kept unchanged, while those of spurious features are eliminated after transformation. Using Eq. (3) (with $\rho = 0.5$ in the true data distribution), we have $C_y = u(a, y, \vec{\delta}) = [0, y, \vec{0}], x = [Ba, y, \delta]$, and we learn a $t(\vec{x}) = \vec{w}(\vec{x} - \vec{b}) + \vec{b} = [w_1(Ba - b_1), w_2(y - b_2), \vec{w}_3(\vec{\delta} - \vec{b}_3)]$ to make $t(x) = C_y$, then

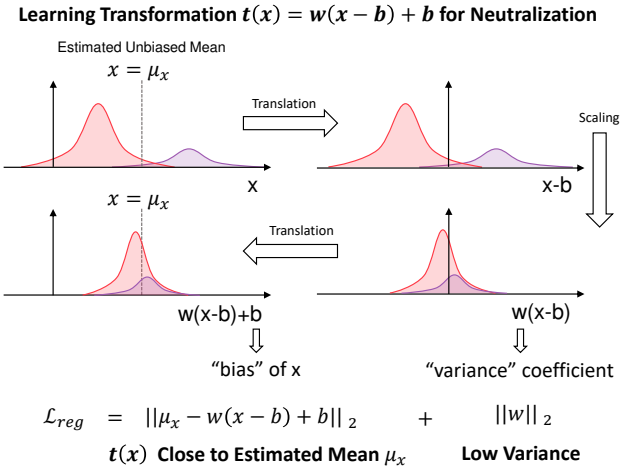


Figure 3. A linear transformation t(x) is learned to eliminate the spurious features by shifting them to their unbiased mean and reducing variance so that the correlation between the spurious feature and label is removed.

- Spurious: $w_1(Ba - b_1) = 0 \xleftarrow{optimal} w_1 = 0, b_1 = 0$
- Core: $w_2(y - b_2) = y \xleftarrow{optimal} w_2 = 1, b_2 = 0$
- Noise: $\vec{w}_3(\vec{\delta} - \vec{b}_3) = \vec{0} \xleftarrow{optimal} w_3 = 0, b_3 = 0$

So optimal $t(x) = [0, y, \vec{0}]$, making core features kept unchanged, and spurious features and noise are eliminated after the transformation.

Learning the Feature Transformation t(x) The optimization objective is to minimize the Euclidean distance between the unbiased mean and the transformed data points

$$\mathcal{L}_{reg} = \lambda||\vec{w}||_2 + \frac{1}{N}\sum_i^N o_i||t(\mathbf{sg}[\vec{x}_i]) - C_{y_i}||_2, \quad (5)$$

where $\mathbf{sg}[\cdot]$ is the stop-gradient operation. This objective ensures that the core feature remains unchanged, while the impact of spurious features and noise is reduced.

4.6. Debiasing the Classifier

Even with transformed features, a classifier can still make false predictions due to high coefficients on spurious features. Here we train a new classifier on the balanced-sampled data aligning with the true data distribution.

Minority Sampling For better debiasing, the training data should be sampled fairly from the majority and minority groups, which are found in Section 4.3. Formally, let $\mathcal{D} = \{(\vec{x}_i, y_i)\}_{i=1}^N$ be the entire dataset, and $\mathcal{M}_1, \mathcal{M}_2 \subset \mathcal{D}$.

$$\mathcal{M}_1 = \{(\vec{x}_{i_1}, y_{i_1}) \mid d(\vec{x}_i, \rho) > 0 \wedge d(t(\vec{x}_i), 0.5) < 0\} \quad (6)$$

$$\mathcal{M}_2 = \{(\vec{x}_{i_1}, y_{i_1}) \mid d(\vec{x}_i, \rho) < 0\} \quad (7)$$

The sampling process can be defined as $\mathcal{S}_{M_i} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_B, y_B) \mid (\vec{x}_k, y_k) \in \mathcal{M}_i, 1 \leq$

	Labels		Waterbirds		CelebA		MultiNLI		CivilComments		Mean
	Tr.	Val	I.I.D.	WGA	I.I.D.	WGA	I.I.D.	WGA	I.I.D.	WGA	WGA
ERM	✗	✗	97.30	72.60	95.60	47.20	82.09	68.11	92.34	57.06	61.24
JTT[30]	✗	✗	93.30	86.70	88.00	81.10	78.60	72.60	91.10	69.30	77.43
MT[4]	✗	✗	93.00	86.40	91.30	78.00	——— Not Applicable ———				
CNC[53]	✗	✗	90.90	88.50	89.90	**88.80**	-	-	81.70	68.90	-
AFR[38]	✗	✗	94.20	90.40	91.30	82.00	81.40	**73.40**	89.80	68.70	78.63
Ours	✗	✗	95.65± 0.0011	**91.12**± 0.0063	88.70± 0.0036	84.27± 0.0047	80.43± 0.0003	73.12± 0.0008	87.19± 0.0010	**79.51**± 0.0022	**82.01**
GDRO[40]	✓	✓	93.50	91.40	92.90	88.90	81.40	77.70	88.90	69.90	81.98
DFR[25]	✗	✓	94.20	92.90	91.30	88.30	82.10	74.70	87.20	70.10	81.50
uLA[48]	✗	✓	91.50	86.10	93.90	86.50	——— Not Applicable ———				

Table 1. Results of mean (i.i.d) and worst-group accuracy (WGA) and standard deviation over 10 random seeds on four image and text debiasing benchmarks. The proposed method, which does not require bias labels from the training (Tr.) and validation (Val) set, achieves superior or comparable WGA compared to methods like GroupDRO and DFR which require bias labels.

	Bias Label		Chexpert		
	Train	Val	I.I.D	WGA	Gain
ERM	✗	✗	89.78	31.21	-
JTT[30]	✗	✗	75.20	60.40	+29.19
Ours	✗	✗	**81.94**	**70.21**	**+40.00**
GDRO[40]	✓	✓	78.90	74.50	+43.29

Table 2. Results on Chexpert Dataset.

$k \leq B\}$, where $\mathcal{S}_{M_1}$ and $\mathcal{S}_{M_2}$ are the sampled sets from $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The new classifier h' is trained using Cross-Entropy (CE) loss, as follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(h'(\mathbf{sg}[t(\vec{x}_i)]), y_i) + \mathcal{L}_{CE}(h'(\mathbf{sg}[t(\vec{x}'_i)]), y'_i), \tag{8}$$

where $(\vec{x}_i, y_i)$ and $(\vec{x}'_i, y'_i)$ are the i-th example in the $\mathcal{S}_{M_1}$ and $\mathcal{S}_{M_2}$, and $\mathbf{sg}[\cdot]$ is the stop-gradient operation. We also conduct ablation of sampling in Sec. 5.4.

5. Experiments

5.1. Experiment Settings

In this section, we provide a detailed description of the experimental setup.

Datasets & Metrics To validate the proposed methods, we conduct experiments on both image and text tasks using five common benchmark datasets: The Waterbirds [40] dataset comprises 4,795 images, with class labels of landbirds and waterbirds, and a spurious attribute of background. CelebA [31] contains over 200,000 images, with hair color labeled and gender as a spurious attribute. MultiNLI [49] includes over 400,000 sentence pairs categorized into entailment, contradiction, or neutral, with the presence of negation words as a spurious attribute. CivilComments-WILDS [26] also has over 400,000 text samples, labeled toxic or non-toxic, with spurious attributes related to demographic identities, including male, female, LGBTQ, Christian, Muslim, other religions, Black, and White. Finally, the CheXpert dataset includes over 200,000 chest radiographs labeled as ill or non-ill, without a specified spurious attribute. We report the WGA and mean accuracy on the test set using checkpoint from the last epoch.

Network Architecture We use ResNet-50 [17] for image classification tasks, and BERT-base-uncased [13] for text tasks as in most previous works.

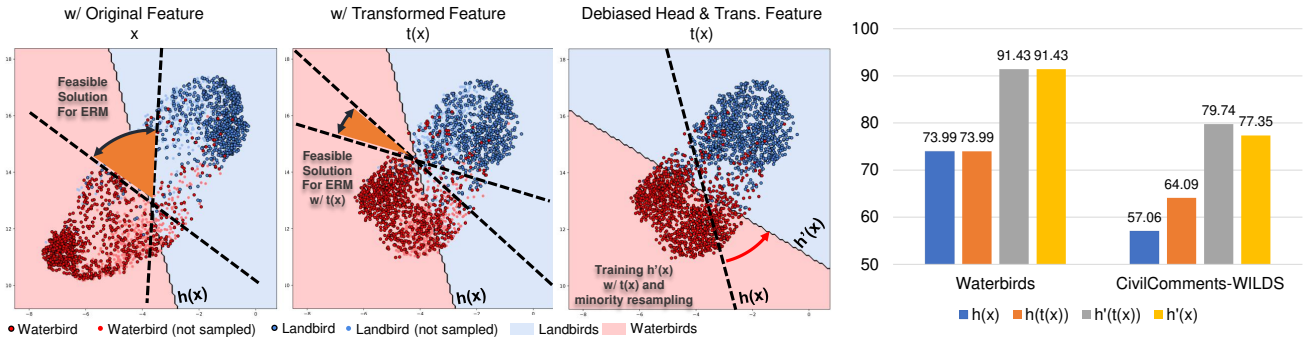
Implementation Details All experiments were conducted on Nvidia GeForce 3090 GPU with 24GB VRAM using PyTorch [37]. We first train an ERM model as in the DFR [25] study. For the CheXpert dataset, we follow the preprocessing and groups in the SubPopBench [50] for a fair comparison. Then a transformation is learned using the pre-extracted embeddings from the ERM-trained model and data in the validation set. The AdamW optimizer is used with a learning rate of 1e-3 and zero weight decay.

5.2. Comparison with the state-of-the-art Methods

We compare the NSF against several state-of-the-art methods, including JTT [30], MaskTune (MT) [4], CNC [53], AFR [38], GroupDRO (GDRO) [40], DFR [25] and uLA [48], and report mean (i.i.d) and worst-group accuracy (WGA) and standard deviation (std) over 10 random seeds on four image and text debiasing benchmarks. As in Tab. 1, the proposed method outperforms competing methods with an average of 82.01% and low std.

5.3. Mitigating Bias in Medical Domain

We conduct experiments on the Chexpert [22] dataset to validate the proposed method in the field of medical imaging, since unknown biases are fatal for automatic diagnostic, as in Tab. 2. The proposed method shows a significant improvement compared to baseline methods.



(a) Visualization of Features and Decision Boundary of Classifier **(b) Combination of Classifiers & Features**

Figure 4. **(a)** Features and the decision boundary using models trained on the waterbirds dataset, using group-balanced sampling for better visualization. 1) Original features $\vec{x}$ allows more biased solutions for the ERM training; 2) Elimination of spurious features using $t(\vec{x})$ leaves a smaller room for biased solutions; 3) Finetuning $h'(\vec{x})$ using $t(\vec{x})$ results in a unbiased classifier. **(b)** Combinations of classifiers of ERM $h(\vec{x})$ and debiased $h'(\vec{x})$, and the features of raw $\vec{x}$ and transformed $t(\vec{x})$. The debiased classifier $h'(\vec{x})$ performs well using the original features $\vec{x}$ indicating $h'(\vec{x})$ relies on core rather than spurious features.

	Feature		Finetuning		
Method	$\vec{x}$	$t(\vec{x})$	ERM	MR	WGA
Baseline	✓		✓		72.60
ERM+$t(\vec{x})$		✓	✓		87.85
MR+$\vec{x}$	✓			✓	86.92
Ours		✓		✓	91.43

Table 3. Ablation of components using the waterbirds dataset. Finetuning the classifier using the transformed features $t(\vec{x})$ without resampling results in an increase of 15.25% in WGA, proving that eliminating spurious features reduces the possibility of model bias. The similar good performance using minority sampling illustrates the correctness of the majority and minority groups found, and bias-sampled data makes a biased model.

	Groups		h(t($\vec{x}$))	
Method	Random	U_k, V_k	i.i.d	WGA
Baseline			97.30	72.60
Random	✓		90.60	73.99
U_k, V_k		✓	95.65	91.12

Table 4. Ablation of the U_k and V_k used in learning $t(\vec{x})$ using the waterbirds dataset. Replacing each sample in the U_k and V_k with a random one (equal to the sample mean) causes a significant drop in mean accuracy of 6%, indicating that transforming to a biased mean corrupts the representation. Taking the example of dog vs. cat, it transforms a white cat to black (color of the majority group).

5.4. Ablation Study

We conduct experiments on two debiasing benchmarks, the Waterbirds (by default) and CivilComments-WILDS to validate each component in the proposed NSF, and the worst group accuracy is presented in Tab. 3 and Fig. 4(b). For a better understanding, we visualize the features and decision boundary in the waterbirds dataset using UMAP [32].

The ERM & Sampling Bias As shown in Fig. 4(a), we plot group-balance sampled data from pre-extracted embeddings in the waterbird dataset, and borders are added as the ratio they are sampled in the training set for better visualization. It can be seen that such a sampling bias in the training data results in a larger feasible solution space for the ERM training, and most of them leverage the spurious feature for separating those classes, leading to a model bias.

Ablation of Components As in Tab. 3, both proposed components improve WGA, and the combination of them is the best, demonstrating their effectiveness.

The Estimation of Bias-Invariant Feature Replacing examples in the U_k and V_k with randomly selected examples used in learning $t(\vec{x})$, as in Tab. 4 cause a drop of mean accuracy to 90.60%, indicating that transforming to a biased mean corrupts the representation, highlighting the importance of using unbiased mean values for neutralizing the spurious features.

Transformed Feature $t(\vec{x})$ As shown in Tab. 3, finetuning the classifier with transformed features $t(\vec{x})$ improves WGA by 15.25%, demonstrating that eliminating spurious features reduces model bias. Fig. 4(b) shows both $h'(\vec{x})$ and $h'(t(\vec{x}))$ perform well with the debiased classifier, with the transformed features achieving even better results, highlighting the effectiveness of eliminating spurious features.

Channels of Spurious Features Discard some channels of feature by the lowest of the coefficient w in the transformation $t(x)$ outperform random selection of same proportion. This validate low w highly correlate with spurious features so that they are eliminated, as in Fig. 5.

Debiased Classifier $h'(\vec{x})$ To validate if the classifier successfully removes the bias, we test combinations of classifiers of ERM $h(\vec{x})$ and debiased $h'(\vec{x})$, and the features

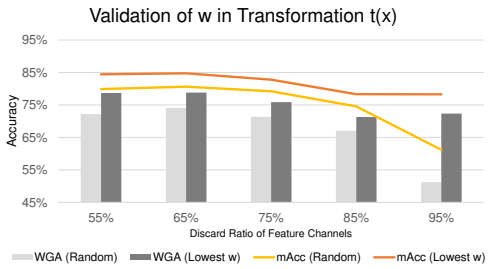


Figure 5. The WGA and mAcc of discarding channels by the lowest of the coefficient w in the transformation $t(x)$ are higher than choosing randomly, validating lower w highly correlate with spurious features so that they are eliminated, as in Fig. 3.

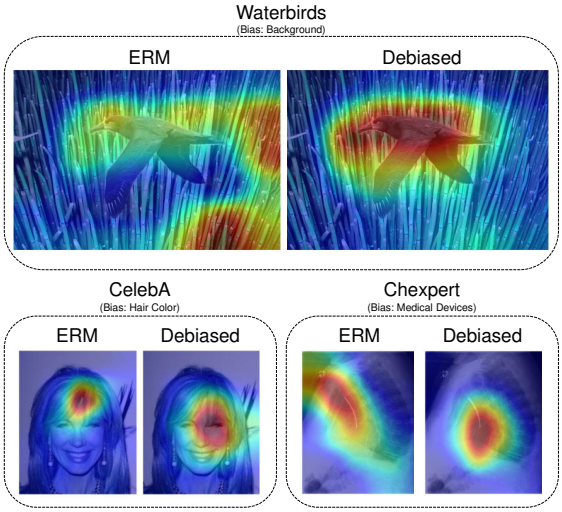


Figure 6. Biases found on the Waterbirds, CelebA, and CheXpert Datasets. The found biases are aligned with known biases (background, hair color, and medical devices).

of raw $\vec{x}$ and transformed $t(\vec{x})$, as in Fig. 4(b). The improvement of $h'(\vec{x})$ using the original features $\vec{x}$ indicates the successful removal of dependence on spurious features.

6. Discussion

The Biases Found on Image Datasets As the CAM [43] in Fig. 6, for Waterbirds, the ERM model focuses on the background, while the debiased model centers on the bird's body. In CelebA, the ERM model highlights the hair, whereas the debiased model focuses on the face. In CheXpert, the ERM model targets medical devices, while the debiased model concentrates on clinically relevant areas. These visualizations show that the found biases are aligned with the known bias (the background, the hair color, and medical devices) in those datasets and debiasing leads to models using more relevant patterns.

Results on Different Architecture ViT-s with ours achieves +5.56 in mAcc (90.78→96.34) and +21.96 in WGA (67.13→89.09) on Waterbirds compared to ERM.

COMP.	WB	CA	MN	CC	MEAN
$t(\vec{x})$	1s	2s	15s	2s	5s
$h'(\vec{x})$	1s	1s	1s	1s	1s

Table 5. Comparison of training time on different datasets.

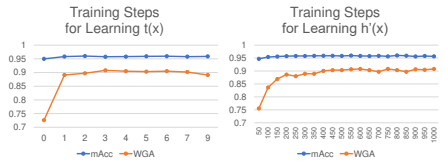


Figure 7. The impact of the training steps.

Swapping the Target y and Spurious Attribute a on CelebA results in +1.87 in WGA (90.56→92.43) and -2.79 in mAcc (98.60→95.81) compared to ERM.

Training Efficiency & Convergence As shown in Tab. 5, it takes only a few seconds for the proposed method to remove the bias, highlighting advantages in training time and cost. We abbreviate WB for Waterbirds, CA for CelebA, MN for MultiNLI, and CC for CivilComments-WILDS. Fig. 7 shows a clear trend of improvement in performance with increased training then begins to plateau, indicating a successful convergence.

Limitations The NSF follows findings in [19], which suggest models favor strong features even when less predictive. This assumption may limit the applicability of NSF under weaker biases, which we should address in future works.

7. Conclusion

This work introduces NSF, a novel method for mitigating unknown biases in DNNs. NSF effectively debias models by identifying and eliminating spurious features, while reducing their influence in the classifier, using insights from the model's bias-fitting mechanism. Extensive experiments across multiple benchmarks show that NSF achieves state-of-the-art results in both vision and text tasks, with minimal computational cost, demonstrating its efficiency. Due to its easy integration, non-intrusive nature, and high efficiency, NSF provides a new choice for addressing unknown biases in DNNs.

Acknowledgment: This work is supported by the National Natural Science Foundation of China (No.62176047), the Sichuan Science and Technology Program (No.2021YFS0374) and Sichuan Natural Science Foundation (No.2024NSFTD0041).

References

[1] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *International Conference on Learning Representations*, 2024. 2

[2] Jing An, Lexing Ying, and Yuhua Zhu. Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients. In *International Conference on Learning Representations*, 2021. 1, 3

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[4] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. *Advances in Neural Information Processing Systems*, 35:23284–23296, 2022. 6

[5] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 2

[6] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pages 456–473, 2018. 1

[7] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[8] Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[9] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13022–13032, 2022. 2, 3

[10] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, David Suter, and Alireza Bab-Hadiashar. An information-theoretic method to automatic shortcut avoidance and domain generalization for dense prediction tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[11] Sander De Coninck, Sam Leroux, and Pieter Simoens. Mitigating bias using model-agnostic data attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 235–243, 2024. 1

[12] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 6

[14] 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*, 2018. 2

[15] 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, 2020. 1, 2, 3

[16] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[18] Xilin He, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Feng Liu, and Linlin Shen. Shift from texture-bias to shape-bias: Edge deformation-based augmentation for robust object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1526–1535, 2023. 2

[19] Katherine Hermann, Hossein Mobahi, Thomas FEL, and Michael Curtis Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, 2024. 2, 8

[20] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34: 26449–26461, 2021. 3

[21] Rui Hu, Yahan Tu, and Jitao Sang. Echoes: Unsupervised debiasing via pseudo-bias labeling in an echo chamber. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1616–1624, 2023. 2

[22] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597, 2019. 6

[23] 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*, pages 14992–15001, 2021. 2

[24] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020. 2

[25] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 6

[26] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 6

[27] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *International Conference on Learning Representations*, 2023. 3

[28] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023. 2

[29] 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*, pages 3832–3841, 2023. 3

[30] 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 *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 6

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 6

[32] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7

[33] Shlok Kumar Mishra, Anshul Shah, Ankan Bansal, Janit Anjaria, Jonghyun Choi, Abhinav Shrivastava, Abhishek Sharma, and David Jacobs. Learning visual representations for transfer learning by suppressing texture. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 300. BMVA Press, 2022. 2

[34] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 2

[35] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022. 3

[36] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020. 1

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[38] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pages 28448–28467. PMLR, 2023. 6

[39] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR, 2020. 1

[40] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 1, 2, 3, 6

[41] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 2, 3

[42] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. 2

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[44] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16742–16751, 2022. 1, 3

[45] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772, 2022. 2, 3

[46] Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *International Conference on Machine Learning*, pages 34330–34343. PMLR, 2023. 2

[47] A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011. 1, 2, 4

[48] Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. Group robust classification without any group information. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 6

[49] Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122, 2018. 6

[50] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, pages 39584–39622. PMLR, 2023. 6

[51] 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*, pages 4794–4804, 2023. 3

[52] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. 1

[53] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, pages 26484–26516. PMLR, 2022. 3, 6