# ShortcutProbe: Probing Prediction Shortcuts for Learning Robust Models

**Guangtao Zheng** , **Wenqian Ye** and **Aidong Zhang**

University of Virginia

{gz5hp, pvc7hs, aidong}@virginia.edu

## Abstract

Deep learning models often achieve high performance by inadvertently learning spurious correlations between targets and non-essential features. For example, an image classifier may identify an object via its background that spuriously correlates with it. This prediction behavior, known as spurious bias, severely degrades model performance on data that lacks the learned spurious correlations. Existing methods on spurious bias mitigation typically require a variety of data groups with spurious correlation annotations called group labels. However, group labels require costly human annotations and often fail to capture subtle spurious biases such as relying on specific pixels for predictions. In this paper, we propose a novel post hoc spurious bias mitigation framework without requiring group labels. Our framework, termed ShortcutProbe, identifies prediction shortcuts that reflect potential non-robustness in predictions in a given model's latent space. The model is then retrained to be invariant to the identified prediction shortcuts for improved robustness. We theoretically analyze the effectiveness of the framework and empirically demonstrate that it is an efficient and practical tool for improving a model's robustness to spurious bias on diverse datasets.

## 1 Introduction

Deep learning models have shown remarkable performance across domains, but this success is often achieved by exploiting spurious correlations [Sagawa *et al.*, 2020; Liu *et al.*, 2021; Yang *et al.*, 2023; LaBonte *et al.*, 2024; Ye *et al.*, 2024; Zheng *et al.*, 2024a] between spurious attributes or shortcut features [Geirhos *et al.*, 2020] and targets. For example, models have been found to use correlations between textures and image classes [Geirhos *et al.*, 2019] for object recognition instead of focusing on defining features of objects. This issue becomes even more problematic in high-stakes domains like healthcare. For instance, models predicting pneumonia were shown to rely on correlations between metal tokens in chest X-ray scans from different hospitals and the disease's detection outcomes [Zech *et al.*, 2018], rather than the pathological features of pneumonia itself. The tendency of using spurious correlations is referred to as *spurious bias*. Models with spurious bias often fail to generalize on data groups lacking the learned spurious correlations, leading to significant performance degradation and non-robust behaviors across different data groups. This robustness issue can have severe social consequences, especially in critical applications.

Mitigating spurious bias is crucial for robust generalization across data groups with varying spurious correlations. Existing methods on mitigating spurious bias [Sagawa *et al.*, 2020; Kirichenko *et al.*, 2023] rely on group labels. Group labels represent spurious correlations with class labels and spurious attributes. For example, (`waterbirds, water`) [Sagawa *et al.*, 2020] represents a spurious correlation between waterbirds and water backgrounds in the images of waterbirds with water backgrounds. Using group labels specifies explicitly the spurious correlations that a model should avoid. However, obtaining group labels requires expert knowledge and labor-intensive annotation efforts. Moreover, group labels fail to capture subtle spurious biases, such as using certain pixels in images for predictions.

In this paper, we propose a post hoc approach that can automatically detect and mitigate potential spurious biases in a model rather than relying on group labels. Our key innovation is reframing the task of identifying and mitigating spurious biases as detecting and leveraging *prediction shortcuts* in the model's latent space. Prediction shortcuts are latent features derived from input embeddings and predominantly contribute to producing the same prediction outcome across different classes. In essence, prediction shortcuts represent non-defining features of certain classes that the model heavily uses for predictions. By operating in the model's latent space, our approach leverages the expressiveness of latent embeddings, enabling direct identification of spurious biases across diverse input formats without requiring group labels.

We present our post hoc approach as a novel framework called *ShortcutProbe*, which first identifies prediction shortcuts in a given model and leverages them to guide model retraining for spurious bias mitigation. ShortcutProbe utilizes a probe set without group labels, typically containing a diverse mix of features, to uncover potential prediction shortcuts. These shortcuts are identified as latent features extracted from sample embeddings belonging to different classes but producing the same prediction outcome. By optimizing these

features to maximize the model's confidence in their corresponding predictions, ShortcutProbe effectively encodes spurious attributes in non-generalizable prediction shortcuts that the model overly relies on for predictions.

With the identified prediction shortcuts, ShortcutProbe mitigates spurious biases by retraining the model to be invariant to these shortcuts, as they are irrelevant to true prediction targets. This invariance is achieved by applying regularization during retraining, which ensures that the identified prediction shortcuts no longer contribute to the model's predictions of the true targets.

We theoretically demonstrate that when the spurious attributes in the training data are new to the model as reflected by the high prediction loss, the tendency of using spurious attributes for predictions is high after training on the data. Our method aims to revert the process of learning spurious attributes by retraining the model so that the learned spurious attributes induce high prediction losses, effectively unlearning the spurious attributes and reducing the influence of spurious correlations.

In summary, our contributions are as follows:

- We introduce ShortcutProbe, a novel post hoc framework for mitigating spurious bias without requiring group labels. ShortcutProbe identifies prediction shortcuts and leverages them as a form of regularization for training robust models.

- We provide a theoretical analysis revealing that our spurious bias mitigation approach effectively unlearns spurious attributes.

- Through extensive experiments, we show that our method successfully trains models robust to spurious biases without prior knowledge about these biases.

## 2 Related Works

**Shortcuts and spurious bias.** Shortcuts are decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions [Geirhos *et al.*, 2020], such as data with distributional shifts. Spurious correlations describe superficial associations between spurious (non-essential) features and targets in data, and they can be used by machine learning models as shortcuts [Geirhos *et al.*, 2019]. This shortcut-learning phenomenon in machine learning models results in spurious bias, the tendency to use spurious correlations in data for predictions. Shortcut features are spurious attributes that are used by models in predictions. Accurately identifying shortcut features involves domain knowledge and model interpretation techniques. In our work, we circumvent this by identifying prediction shortcuts that represent *potential* shortcut features in a model's latent space, making our method a general and useful method for spurious bias detection and mitigation.

**Spurious bias mitigation with group labels.** Group labels in training data indicate the presence of spurious correlations across different subsets of the data. Exploiting group labels during training helps mitigate the reliance on the specified spurious correlations. Existing approaches for spurious

bias mitigation utilize group labels to balance data distributions during training [Cui *et al.*, 2019; He and Garcia, 2009; Byrd and Lipton, 2019], to formulate a distributionally robust optimization objective [Sagawa *et al.*, 2019], or to progressively expand group-balanced training data [Deng *et al.*, 2023]. Although these methods achieve remarkable success in spurious bias mitigation, the reliance on training group labels becomes a barrier in practice, as obtaining such labels often requires domain knowledge and labor-intensive annotation efforts. Our method does not require group labels during training; instead, it automatically detects prediction shortcuts that encode potential spurious biases the model has developed in training.

**Reducing reliance on group labels.** To alleviate the dependency on training group labels, existing approaches propose inferring training group labels through various means, such as identifying misclassified samples [Liu *et al.*, 2021], clustering hidden representations [Zhang *et al.*, 2022], employing invariant learning techniques [Creager *et al.*, 2021], or training group label estimators using a few group-labeled samples [Nam *et al.*, 2022]. Nevertheless, validation group labels are still needed to specify which biases to address. Recent works [Zheng *et al.*, 2024b; Zheng *et al.*, 2024c] use vision-language models to infer group labels. Last-layer retraining methods [Kirichenko *et al.*, 2023; LaBonte *et al.*, 2024] leverage group-balanced validation data to fine-tune the last layer of a model. While our approach also involves retraining the last layer, it does not require group labels. Instead, we introduce a novel strategy that utilizes identified prediction shortcuts during retraining, leading to improved model robustness.

## 3 Preliminary

A *spurious correlation* is the correlation between a *spurious attribute* present in the training samples and a prediction target. For example, the class `waterbird` and the attribute `water background` might form a spurious correlation in the images of `waterbird`, where some of them have water backgrounds, e.g., pond or river, and some do not. Spurious attributes are not truly predictive of the targets. A group label, e.g., (`waterbird`, `water background`), consists of a prediction target and a spurious attribute.

Consider a model $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ with parameter $\theta$ trained on a training dataset $\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$ with $N$ sample-label pairs, where $x_i \in \mathcal{X}$ denotes a sample in the input set $\mathcal{X}$, $y_i \in \mathcal{Y}$ denotes a label in the label set $\mathcal{Y}$, and $|\cdot|$ denotes the size of a set. The model $f_\theta = e_{\theta_1} \circ h_{\theta_2}$ can be considered as a feature extractor $e_{\theta_1} : \mathcal{X} \to \mathbb{R}^D$ followed by a classifier $h_{\theta_2} : \mathbb{R}^D \to \mathbb{R}^{|\mathcal{Y}|}$, where $\theta = \theta_1 \cup \theta_2$, $h_{\theta_2}$ is the last layer of the model, and $D$ denotes the number of dimensions.

Due to the existence of spurious attributes in $\mathcal{D}_{\text{tr}}$, the model can exploit them for predictions, such as recognizing waterbirds by detecting the existence of water backgrounds [Sagawa *et al.*, 2019]. This presents a challenge: It is hard to determine whether a high-performing model is truly robust or simply "right for the wrong reasons", i.e., relying on spurious attributes. Although models with spurious biases typically exhibit degraded performance when the learned spurious attributes are absent from input data, e.g., a waterbird on
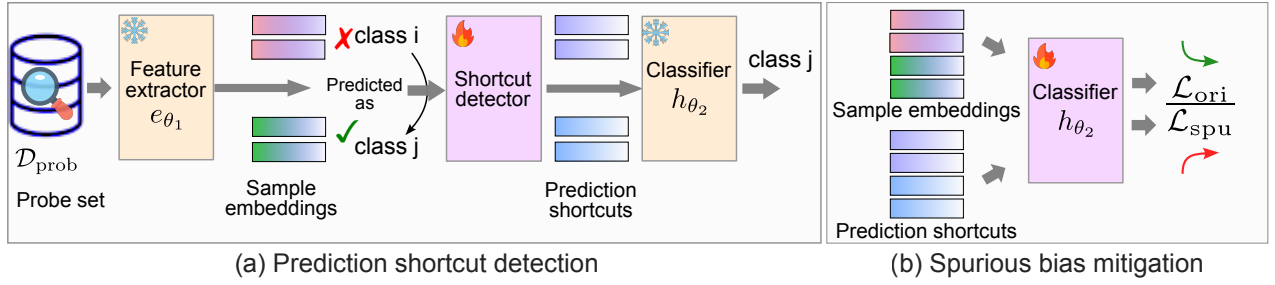
Figure 1: Illustration of ShortcutProbe. (a) The framework uses a set of probe data $\mathcal{D}_{\text{prob}}$ to identify prediction shortcuts by learning a shortcut detector to extract similar features from samples of different classes $i$ and $j$ that are all predicted as the same class $j$. Feature extractor $e_{\theta_1}$ and classifier $h_{\theta_2}$ are frozen during this stage. (b) ShortcutProbe then retrains the classifier with the probe data (the loss of the probe data $\mathcal{L}_{\text{ori}}$) while using the identified prediction shortcuts as regularization (the loss of the prediction shortcuts $\mathcal{L}_{\text{spu}}$).

a land background, it remains challenging to identify the specific spurious attributes encoded by the model without group labels, which hinders the development of effective spurious bias mitigation strategies.

## 4 Methodology

### 4.1 Method Overview

We introduce *ShortcutProbe*, a post hoc framework that automatically detects and mitigates prediction shortcuts without requiring group labels to specify which spurious biases to address. The framework comprises two key steps: (1) **Prediction shortcut detection**, where a probe set is used to identify prediction shortcuts, and (2) **Spurious bias mitigation**, where the identified shortcuts are used in retraining the model to mitigate spurious biases in the model.

We provide an overview of our framework in Fig. 1. The process begins with a set of probe data $\mathcal{D}_{\text{prob}}$, which typically contains samples with various spurious attributes that reflect a model's non-robustness to spurious correlations. These samples are mapped to the latent embedding space of the model $f_\theta$ through its feature extractor $e_{\theta_1}$, allowing us to model *any* prediction shortcuts the model might use for predictions. This strategy bypasses the need to explicitly define spurious correlations through group labels, a task that is especially challenging for subtle features, such as specific pixels in images. More concretely, as shown in Fig. 1(a), we first train a shortcut detector that extracts potential prediction shortcuts from samples of different classes but having the same prediction. The identified prediction shortcuts encode spurious attributes shared across classes and capture the model's non-robustness across different data groups. Next, in Fig. 1(b), the model is retrained with $\mathcal{D}_{\text{prob}}$ to mitigate spurious biases by unlearning the identified prediction shortcuts.

### 4.2 Prediction Shortcut Detection

Given a model $f_\theta$ and a probe set $\mathcal{D}_{\text{prob}}$, we aim to detect prediction shortcuts by learning a shortcut detector $g_\psi : \mathbb{R}^D \to \mathbb{R}^D$ to identify prediction shortcuts from input embeddings. Intuitively, prediction shortcuts can be identified from samples of different classes but having the same prediction outcome, an indication that similar features exist in these samples but are irrelevant to classes. In other words, prediction shortcuts can be shared among samples from different

classes, necessitating a shared representational space to encode diverse prediction shortcuts. We formalize this intuition in the following definition.

**Definition 1** (Prediction shortcuts). *Given input sample-label pairs $(x, y)$ and $(x', y')$, where $y \neq y'$, a trained model $f_\theta = e_{\theta_1} \circ h_{\theta_2}$, sample embeddings $\mathbf{v} = e_{\theta_1}(x)$ and $\mathbf{v}' = e_{\theta_1}(x')$ for $x$ and $x'$, respectively, a vector space $\mathcal{V} \subset \mathbb{R}^D$ spanned by $K$ base column vectors in $\mathbf{A} \in \mathbb{R}^{D \times K}$, prediction shortcuts $\mathbf{s}_x \in \mathcal{V}$ and $\mathbf{s}_{x'} \in \mathcal{V}$ for the two samples satisfy the following conditions:*

- $Pred\big(h_{\theta_2}(e_{\theta_1}(x)) = y,\ Pred\big(h_{\theta_2}(\mathbf{s}_x)\big) = y,\ and$

- $Pred\big(h_{\theta_2}(e_{\theta_1}(x'))\big) = y,\ Pred\big(h_{\theta_2}(\mathbf{s}_{x'})\big) = y,$

*where $Pred(f_\theta(x)) = \arg\max_{\mathcal{Y}} f_\theta(x)$, $\mathbf{s}_x = \mathbf{P_A v}$, $\mathbf{s}_{x'} = \mathbf{P_A v}'$, and*

$$\mathbf{P_A} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \tag{1}$$

*is the projection matrix such that the prediction shortcut $\mathbf{s}_x$ is the best estimate of $\mathbf{v}$ in the vector space $\mathcal{V}$ in the sense that the distance $\|\mathbf{s}_x - \mathbf{v}\|_2^2$ is minimized.*

In Definition 1, we define a prediction shortcut as a projection of a sample embedding. It exists in the vector space $\mathcal{V}$ shared by samples of different classes with $K$ degrees of freedom. Here, $K$ governs the complexity of the vector space representing prediction shortcuts. A smaller $K$ results in a less expressive vector space that may fail to adequately capture prediction shortcuts. Conversely, a larger $K$ provides greater flexibility in representing prediction shortcuts but may encode irrelevant information. The optimal value of $K$ depends on the complexity of the probe data; values that are too small or too large can impede learning and lead to suboptimal performance. We treat $K$ as a tunable hyperparameter.

By representing prediction shortcuts as vectors, we can in theory capture any spurious bias, even the intricate ones. For instance, a vector $\mathbf{s}$ might correspond to features of water backgrounds that are used to predict waterbirds in any image with water backgrounds, revealing a spurious bias in the model. Alternatively, $\mathbf{s}$ could represent a specific feature corresponding to certain pixels in input images, capturing the prediction shortcut based on low-level pixel values—an aspect that is challenging to articulate through group labels.

**Learning the shortcut detector.** Based on the definition of the prediction shortcut, we design the shortcut detector $g_\psi$ as a function that implements the projection operation defined in Eq. (1) with the learnable parameter $\psi = \mathbf{A} \in \mathbb{R}^{D \times K}$, that is for a sample embedding $\mathbf{v} \in \mathbb{R}^D$, $g_\psi(\mathbf{v}) = \mathbf{P_A} \mathbf{v}$. Learning $g_\psi$ essentially learns a shared vector space spanned by $\mathbf{A}$ that could cover prediction shortcuts in samples in the probe set.

To effectively learn $g_\psi$, for each class $y$, we first collect samples from the probe set $\mathcal{D}_{\text{prob}}$ to formulate two non-overlapping sets $\mathcal{D}^y_{\text{cor}}$ and $\mathcal{D}^y_{\text{pre}}$, i.e.,

$$\mathcal{D}^y_{\text{cor}} = \{(x,y) | (x,y) \in \mathcal{D}_{\text{prob}}, \text{Pred}(f_\theta(x)) = y\}, \quad (2)$$

and

$$\mathcal{D}^y_{\text{pre}} = \{(x',y') | (x',y') \in \mathcal{D}_{\text{prob}}, \text{Pred}(f_\theta(x')) = y \neq y'\}, \quad (3)$$

where $\mathcal{D}^y_{\text{cor}}$ and $\mathcal{D}^y_{\text{pre}}$ contain samples that are correctly and incorrectly predicted as $y$. The two sets together demonstrate a possibility that certain features shared across classes are incorrectly associated with prediction targets, allowing us to extract these features as potential prediction shortcuts.

Next, we propose the following objective to identify prediction shortcuts:

$$\mathcal{L}_{\text{det}} = \mathop{\mathbb{E}}_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{(x,y) \in \mathcal{D}^y_{\text{cor}} \cup \mathcal{D}^y_{\text{pre}}} \ell\big(h_{\theta_2}(g_\psi(\mathbf{v})), y\big), \quad (4)$$

where $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ is the loss function, $\psi = \mathbf{A}$, and $\mathbf{v} = e_{\theta_1}(x)$. To ensure that prediction shortcuts are relevant to the given samples, we regularize the semantic similarity between $g_\psi(\mathbf{v})$ and $\mathbf{v}$ as follows,

$$\mathcal{L}_{\text{reg}} = \mathop{\mathbb{E}}_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{(x,y) \in \mathcal{D}^y_{\text{cor}} \cup \mathcal{D}^y_{\text{pre}}} \|g_\psi(\mathbf{v}) - \mathbf{v}\|_2^2. \quad (5)$$

The overall learning objective for $\psi$ is

$$\psi^* = \arg\min_\psi \mathcal{L}_{\text{det}} + \eta \mathcal{L}_{\text{reg}}, \quad (6)$$

where $\eta > 0$ represents the regularization strength. The objective in Eq. (6) aims to encode the properties of prediction shortcuts in Definition 1 into the shortcut detector $g_\psi$ while maintaining the relevance of the prediction shortcuts to input samples. Training $g_\psi$ is lightweight as there are only $DK$ learnable parameters. With $g_\psi$, we can identify multiple prediction shortcuts from samples in $\mathcal{D}_{\text{prob}}$.

## 4.3 Spurious Bias Mitigation

Mitigating spurious biases in a model requires that the spurious attributes captured during training are no longer predictive of the targets. Although identifying spurious attributes can be challenging, the shortcut detector introduced in the previous section identifies prediction shortcuts as potential spurious attributes utilized by the model, providing valuable guidance for addressing spurious biases.

In the following, we formulate an optimization objective that incorporates the above constraint to learn a robust model using the probe set $\mathcal{D}_{\text{prob}}$. First, a general requirement is that the trained model should produce correct and consistent predictions on training samples. To this end, for each class $y$, we sample from $\mathcal{D}^y_{\text{cor}}$ and $\mathcal{D}^y_{\text{mis}}$, where $\mathcal{D}^y_{\text{cor}}$ is the set of correctly

predicted sample-label pairs defined in Eq. (2), and $\mathcal{D}^y_{\text{mis}}$ is defined as follows,

$$\mathcal{D}^y_{\text{mis}} = \{(x',y) | (x',y) \in \mathcal{D}_{\text{prob}}, \text{Pred}(f_\theta(x')) \neq y\}, \quad (7)$$

representing misclassified sample-label pairs from the class $y$. Then, the training objective $\mathcal{L}_{\text{ori}}$ is as follows,

$$\mathcal{L}_{\text{ori}}(\mathcal{D}_{\text{prob}}; \theta) = \mathop{\mathbb{E}}_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{(x,y) \in \mathcal{D}^y_{\text{cor}} \cup \mathcal{D}^y_{\text{mis}}} \ell(f_\theta(x), y), \quad (8)$$

which mitigates potential spurious biases by ensuring consistent predictions in $\mathcal{D}^y_{\text{mis}}$ and $\mathcal{D}^y_{\text{cor}}$.

Moreover, to ensure that the training targets at mitigating the spurious biases in the model, we further formulate a regularization term using the prediction shortcuts identified by our shortcut detector to guide the training process. Specifically, as the identified prediction shortcuts are not predictive of the targets, we aim to maximize the loss on the prediction shortcuts defined as follows,

$$\mathcal{L}_{\text{spu}}(\mathcal{D}_{\text{prob}}; \theta) = \mathop{\mathbb{E}}_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{(x,y) \in \mathcal{D}^y_{\text{cor}} \cup \mathcal{D}^y_{\text{mis}}} \ell(h_{\theta_2}(g_\psi(\mathbf{v})), y), \quad (9)$$

where $\mathbf{v} = e_{\theta_1}(x)$.

We combine the two loss terms above to formulate the **overall training objective** as follows,

$$\theta_2^* = \arg\min_{\theta_2} \lambda \mathcal{L}_{\text{ori}} / \mathcal{L}_{\text{spu}}, \quad (10)$$

where $\lambda > 0$ is the regularization strength. Here, we retrain only the final classification layer of the model while keeping the feature extractor frozen. This approach significantly reduces computational complexity and allows us to reuse the previously learned sample embeddings. Details of the training algorithm are provided in Appendix.

## 4.4 Choice of the Probe Set

The probe set plays a crucial role in both detecting prediction shortcuts and mitigating spurious biases. To achieve the full potential of ShortcutProbe, we use a held-out dataset—unseen by the model—to construct the probe set $\mathcal{D}_{\text{prob}}$. This choice is essential because the model may have memorized the training samples, making it difficult to identify prediction shortcuts based on discrepancies in predictions for samples of the same class. Moreover, we select samples with high predication confidence (measured by output entropy) from the held-out dataset to construct $\mathcal{D}_{\text{prob}}$ so that prediction shortcuts can be easily detected from these samples. Details of constructing $\mathcal{D}_{\text{prob}}$ and results on different choices of $\mathcal{D}_{\text{prob}}$ are provided in Section 5.

## 4.5 Theoretical Analysis

We theoretically demonstrate that minimizing the proposed objective $\mathcal{L}_{\text{ori}} / \mathcal{L}_{\text{reg}}$ effectively unlearns the spurious correlations between spurious attributes and their associated targets captured in the model. Without loss of generality, we analyze this in the context of a general linear regression setting. Consider an input sample $x \in \mathcal{X}$, a prediction target $y \in \mathcal{Y}$, and a spurious-only sample $\tilde{x}$ that lacks any defining features in $x$ related to $y$. Let $x_1, \ldots, x_N$ denote $N$ training samples, $\varphi : \mathcal{X} \to \mathbb{R}^D$ be a generic feature map, and

$J_{\mathbf{w}}(x) = \varphi(x)^T \mathbf{w}$ represent a generalized linear regression model with parameters $\mathbf{w} \in \mathbb{R}^D$.

We denote the correlation between the model output for a spurious-only sample $\tilde{x}$ and a prediction target $y$ as $\rho(J_{\mathbf{w}}(\tilde{x}), y)$. The following lemma [Bombari and Mondelli, 2024] gives an upper bound on the correlation.

**Lemma 1.** *The correlation between the model output for the spurious-only sample $\tilde{x}$ and the prediction target $y$ is upper bounded as follows:*

$$\rho(J_{\mathbf{w}}(\tilde{x}), y) \leq \gamma_\varphi \sigma_{\mathcal{Y}} \sqrt{\mathcal{R}_{\mathcal{X}}}, \tag{11}$$

*where $\mathcal{R}_{\mathcal{X}}$ is the generalization error, $\sigma_{\mathcal{Y}}$ is the standard deviation of prediction targets, and*

$$\gamma_\varphi = \mathbb{E}_{\tilde{x},x}\left[\frac{\varphi(\tilde{x})^T \mathbf{O}\varphi(x)}{\|\mathbf{O}\varphi(x)\|_2^2}\right], \tag{12}$$

*where $\mathbf{O} = \mathbf{I} - \mathbf{V}^T(\mathbf{V}\mathbf{V}^T)^{-1}\mathbf{V}$ is the orthogonal projection matrix, and $\mathbf{V} = [\varphi(x_1), \ldots, \varphi(x_N)]^T \in \mathbb{R}^{N \times D}$ is the feature matrix.*

Given that $\mathcal{R}_{\mathcal{X}}$ and $\sigma_{\mathcal{Y}}$ are independent of $\tilde{x}$, the spurious-only sample $\tilde{x}$ affects the correlation upper bound via the feature alignment term $\gamma_\varphi$ between the spurious attribute of $\tilde{x}$ and the original feature of $x$. We further interpret this term in the following proposition.

**Proposition 1.** *The feature alignment term $\gamma_\varphi$ is the ratio between the expected prediction error on the spurious sample $\tilde{x}$ and the expected prediction error on the original sample $x$:*

$$\gamma_\varphi = \mathbb{E}_{\tilde{x},x}\left[\frac{\varphi(\tilde{x})^T \mathbf{O}\varphi(x)}{\|\mathbf{O}\varphi(x)\|_2^2}\right] = \frac{\mathbb{E}_{\tilde{x}}[\|\mathbf{O}\varphi(\tilde{x})\|_2]}{\mathbb{E}_x[\|\mathbf{O}\varphi(x)\|_2]}. \tag{13}$$

The term $\|\mathbf{O}\varphi(x)\|_2$ in Eq. (13) denotes the error term for the sample $x$, while $\|\mathbf{O}\varphi(\tilde{x})\|_2$ denotes the error term for the spurious sample $\tilde{x}$. Since $x$ and $\tilde{x}$ are independent, the feature alignment term $\gamma_\varphi$ is the ratio between the loss on spurious-only samples and the loss on the original samples. We provide a detailed proof in Appendix.

The prediction shortcuts obtained by our shortcut detector approximate the spurious-only features. Thus, the loss $\mathcal{L}_{\text{spu}}$ approximates the nominator of the regularization term in Eq. (13). Moreover, $\mathcal{L}_{\text{ori}}$ approximates the denominator in Eq. (13). Note that Lemma 1 gives the upper bound of the spurious correlation *before* learning it. The objective in Eq. (4) trains the shortcut detector to learn the correlation by minimizing $\gamma_\phi$. In the mitigation step, the objective in Eq. (10) unlearns the correlation by maximizing $\gamma_\phi$.

## 5 Experiments

### 5.1 Datasets

**Image datasets.** **Waterbirds** [Sagawa *et al.*, 2019] contains waterbird and landbird classes selected from the CUB dataset [Welinder *et al.*, 2010]. The two bird classes are mixed with water and land backgrounds from the Places dataset [Zhou *et al.*, 2017]. **CelebA** [Liu *et al.*, 2015] is a large-scale image dataset of celebrity faces. Images showing two hair colors, non-blond and blond, are spuriously correlated with gender. **CheXpert** [Irvin *et al.*, 2019] is a chest X-ray dataset containing six spurious attributes from the combination of race (White, Black, Other) and gender (Male, Female). Two diagnose results, i.e., "No Finding" (positive) and "Finding" (negative) are the labels. **ImageNet-9** [Ilyas *et al.*, 2019] is a subset of ImageNet [Deng *et al.*, 2009] and contains nine super-classes. It is known to have correlations between object classes and image textures. We prepared the training and validation data as in [Kim *et al.*, 2022] and [Bahng *et al.*, 2020]. **ImageNet-A** [Hendrycks *et al.*, 2021] is a dataset of real-world images, adversarially curated to test the limits of classifiers such as ResNet-50. While these images are from standard ImageNet classes [Deng *et al.*, 2009], they are often misclassified in multiple models. We used this dataset to test the robustness of a classifier after training it on ImageNet-9. **NICO** [He *et al.*, 2021] is designed for out-of-distribution image classification, simulating real-world scenarios where testing distributions differ from training ones. It labels images with both main concepts (e.g., cat) and contexts (e.g., at home). We used the Animal super-class in NICO and followed the setting in [Bai *et al.*, 2021; Tiwari and Shenoy, 2023a] for data preparation.

**Text datasets.** **MultiNLI** [Williams *et al.*, 2017] is a text classification dataset with 3 classes: neutral, contradiction, and entailment, representing the natural language inference relationship between a premise and a hypothesis. The spurious feature is the presence of negation, which is highly correlated with the contradiction label. **CivilComments** [Borkan *et al.*, 2019] is a binary text classification dataset aimed at predicting whether a comment contains toxic language. Spurious features involve references to eight demographic identities: male, female, LGBTQ, Christian, Muslim, other religions, Black, and White.

### 5.2 Experimental Setup

**Constructing the probe set.** From the chosen data source, such as the training or validation set, we sorted the samples within each class by their prediction losses and divided them into two equal halves: a high-loss set and a low-loss set. This approach approximates the incorrectly and correctly predicted samples, respectively, while ensuring that the incorrectly predicted set is non-empty, even when all samples are correctly classified. Within each set, we then selected the top $r\%$ of samples with the highest prediction confidence (i.e., those with the lowest output entropy).

**Training details.** We first trained a base model initialized with pretrained weights using empirical risk minimization (ERM) on the training dataset. Then, we retrained the model on half of the validation set using various bias mitigation methods. For our method, we first constructed the probe set using the same half of the validation set and used the probe set for shortcut detection and mitigation. The remaining half of the validation set was used for model selection and hyperparameter tuning. For experiments on the Waterbirds, CelebA, and CheXpert datasets, we used ResNet-50 as the backbone network, and we used ResNet-18 on the ImageNet-9/A and NICO datasets to ensure a fair comparison with baseline methods. For text datasets, we used a pretrained BERT model [Kenton and Toutanova, 2019]. We trained models us-

| Method | Waterbirds | | | CelebA | | | CheXpert | | |
|---|---|---|---|---|---|---|---|---|---|
| | WGA ($\uparrow$) | Average ($\uparrow$) | Gap ($\downarrow$) | WGA ($\uparrow$) | Average ($\uparrow$) | Gap ($\downarrow$) | WGA ($\uparrow$) | Average ($\uparrow$) | Gap ($\downarrow$) |
| ERM [Vapnik, 1999] | $80.3_{\pm3.1}$ | $93.3_{\pm0.4}$ | 13.0 | $45.6_{\pm2.9}$ | $95.2_{\pm0.2}$ | 49.6 | $22.0_{\pm1.6}$ | $90.8_{\pm0.1}$ | 68.8 |
| JTT [Liu *et al.*, 2021] | $86.7_{\pm1.0}$ | $93.3_{\pm0.2}$ | 6.6 | $40.6_{\pm1.2}$ | $88.6_{\pm0.2}$ | 48.0 | $60.4_{\pm4.9}$ | $75.2_{\pm0.8}$ | 14.8 |
| DFR [Kirichenko *et al.*, 2023] | $90.3_{\pm2.1}$ | $95.0_{\pm1.3}$ | 4.7 | $72.2_{\pm2.0}$ | $92.9_{\pm0.1}$ | 20.7 | $72.7_{\pm1.5}$ | $78.7_{\pm0.4}$ | 6.0 |
| AFR [Qiu *et al.*, 2023] | $88.7_{\pm4.2}$ | $95.0_{\pm1.0}$ | 6.3 | $77.8_{\pm1.5}$ | $91.0_{\pm0.4}$ | 13.2 | $72.4_{\pm2.0}$ | $76.8_{\pm1.1}$ | 4.4 |
| **ShortcutProbe (Ours)** | $\mathbf{90.8_{\pm0.6}}$ | $\mathbf{95.0_{\pm0.3}}$ | **4.2** | $\mathbf{83.4_{\pm0.9}}$ | $91.4_{\pm0.1}$ | **8.0** | $\mathbf{75.0_{\pm0.7}}$ | $\mathbf{79.0_{\pm0.2}}$ | **4.0** |

Table 1: Comparison of worst-group accuracy (WGA) and average accuracy (%) with baseline methods on the Waterbirds, CelebA, and CheXpert datasets. The best results are highlighted in **boldface**. All bias mitigation methods use the same half of the validation set.

| Method | MultiNLI | | | CivilComments | | |
|---|---|---|---|---|---|---|
| | WGA ($\uparrow$) | Average ($\uparrow$) | Gap ($\downarrow$) | WGA ($\uparrow$) | Average ($\uparrow$) | Gap ($\downarrow$) |
| ERM [Vapnik, 1999] | $67.0_{\pm0.4}$ | $82.2_{\pm0.2}$ | 15.2 | $58.5_{\pm1.3}$ | $92.2_{\pm0.1}$ | 33.7 |
| JTT [Liu *et al.*, 2021] | $71.6_{\pm0.8}$ | $80.7_{\pm0.4}$ | 9.1 | $68.3_{\pm0.9}$ | $\mathbf{89.0_{\pm0.3}}$ | 20.7 |
| DFR [Kirichenko *et al.*, 2023] | $72.6_{\pm1.7}$ | $81.8_{\pm0.4}$ | 9.2 | $76.6_{\pm0.8}$ | $85.8_{\pm0.5}$ | 9.2 |
| AFR [Qiu *et al.*, 2023] | $66.6_{\pm0.3}$ | $82.2_{\pm0.2}$ | 15.6 | $74.6_{\pm5.1}$ | $84.7_{\pm2.5}$ | 10.1 |
| **ShortcutProbe (Ours)** | $\mathbf{74.3_{\pm0.7}}$ | $\mathbf{82.6_{\pm0.3}}$ | **8.3** | $\mathbf{79.9_{\pm0.6}}$ | $88.5_{\pm0.2}$ | **8.6** |

Table 2: Comparison of worst-group accuracy (WGA) and average accuracy (%) with baseline methods on the MultiNLI and Civilcomments datasets. Best results are highlighted in **boldface**. All bias mitigation methods use the same half of the validation set.

ing each method three times with different random seeds and reported the average results as well as their standard deviations. Detailed training settings are provided in Appendix[1].

**Metrics.** Without group labels, we used the worst-class accuracy [Yang *et al.*, 2023] for model selection, which is defined as the worst per-class accuracy on an evaluation set. For performance evaluation on the Waterbirds, CelebA, CheXpert, MultiNLI, and Civilcomments datasets, we adopted the widely accepted metric, *worst-group accuracy*, which is the lowest accuracy across multiple groups of the test set with each group containing a certain spurious correlation. We also calculated the *accuracy gap* defined as the standard average accuracy minus the worst-group accuracy to measure the degree of a classifier's spurious biases. A high worst-group accuracy with a low accuracy gap indicates that the classifier is robust to spurious biases and can fairly predict samples from different groups. We adopted *average accuracy* for the evaluations on the NICO, ImageNet-9, and ImageNet-A datasets as these datasets are specifically constructed to evaluate the robustness against distributional shifts.

### 5.3 Analysis of Probe Set

Our method relies on a probe set for detecting prediction shortcuts and mitigating spurious bias. A good probe set can be used to effectively reveal and mitigate spurious biases in a model, such as those curated with group labels [Sagawa *et al.*, 2019]. Here, we show that our method, ShortcutProbe, performs effectively without group labels when a probe set is carefully selected from *readily available sources*, such as the training data and held-out validation data.

To demonstrate, we constructed a probe set as described in Section 5.2, using the training set or half of the validation set (with the other half reserved for model selection) as the data

[1]Code is available at https://github.com/gtzheng/ShortcutProbe.

source. For each data source, we varied the proportion $r\%$ from 20% to 100%. This adjustment created different probe sets, ranging from those containing only samples with highly confident predictions (small $r$) to those including all samples from the selected data source ($r = 100$).

Fig. 2(a) shows the performance of ShortcutProbe measured by worst-group accuracy (WGA) under different probe sets, while Fig. 2(b) presents the sizes of these probe sets. Compared to ERM models, we observe that using the training data for retraining results in minimal improvement on the Waterbirds dataset. This is because most of the training data can be correctly predicted by the model, resulting in a probe set that is not informative for learning prediction shortcuts.

In contrast, by leveraging a relatively small amount of the held-out data compared to the size of the training data, our method demonstrates a significant improvement in robustness to spurious biases. Additionally, our approach benefits most from samples with high prediction confidence, i.e., when $r$ is small. However, as shown in Fig. 2(a), setting $r$ too small results in an insufficient number of training samples, leading to suboptimal WGA performance. In the following, unless otherwise specified, we use half of the validation set to construct the probe set and treat $r$ as a tunable hyperparameter.

### 5.4 Main Results

We focus on a challenging and practical setting where group labels are unavailable in both training and validation data. This scenario requires detecting and mitigating spurious biases using only the available data and model in a standard ERM training setup. As baselines, we selected state-of-the-art last-layer retraining methods DFR [Kirichenko *et al.*, 2023] and AFR [Qiu *et al.*, 2023], as well as JTT [Liu *et al.*, 2021], which retrains the entire model. For DFR, which typically requires group labels, we used class labels instead. All baseline methods listed in Tables 1 and 2 utilized half of the
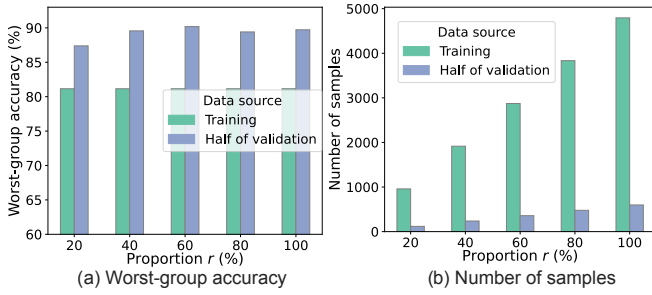
Figure 2: Analyses on different probe sets constructed from the Waterbirds dataset. (a) Worst-group accuracy comparison between models trained with training data and half of the validation data. (b) Numbers of samples in respective probe sets.

| Method | Accuracy ($\uparrow$) |
|---|---|
| ERM | 75.9 |
| REx [Krueger *et al.*, 2021] | 74.3 |
| Group DRO [Sagawa *et al.*, 2019] | 77.6 |
| JiGen [Carlucci *et al.*, 2019] | 85.0 |
| Mixup [Zhang *et al.*, 2017] | 80.3 |
| CNBB [He *et al.*, 2021] | 78.2 |
| DecAug [Bai *et al.*, 2021] | 85.2 |
| SIFER [Tiwari and Shenoy, 2023b] | 86.2 |
| **ShortcutProbe (Ours)** | **90.5**$_{\pm 0.6}$ |

Table 3: Comparison of average accuracy (%) on the NICO dataset.

validation data for training. Similarly, we used the same half of the validation data to construct a probe set, derived as a subset of this portion. The remaining half of the validation data was reserved for model selection.

As shown in Tables 1 and 2, our method achieves the highest WGA and the smallest accuracy gap between average accuracy and WGA, indicating its effectiveness in balancing performance across different data groups. Methods that achieve high average accuracy, such as DFR on the CelebA dataset, prioritize learning spurious features in the probe set. Although maintaining good predictivity on average, DFR still suffers from prediction shortcuts, as shown by its low WGA. Our method remains effective on larger backbone networks beyond ResNet-50, such as ResNet-152 and ViT (see Appendix). Unlike other baseline methods, our method uses only a portion of available data for training, highlighting the effectiveness of the probe set in detecting and mitigating spurious biases. Notably, when we applied the same probe set with baseline methods, this led to degraded performance in both WGA and accuracy gap, underscoring the unique advantages of our approach.

We further tested the out-of-distribution generalization of our method on the NICO dataset. Images of each class in the test set are associated with an unseen context. Our method achieves the best classification accuracy without using group labels (Table 3), demonstrating its effectiveness in mitigating the reliance on contexts. We present additional results on the ImageNet-9 and ImageNet-A datasets in Appendix to demonstrate our method's effectiveness in combating distributional shifts and the capability of achieving good tradeoff between
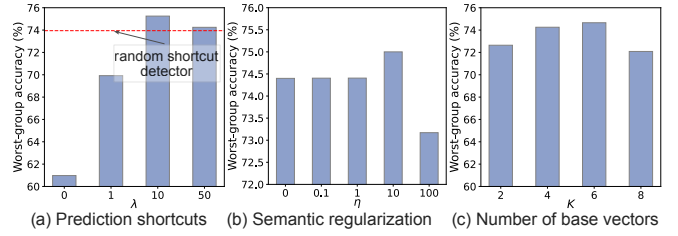


Figure 3: Analyses on how (a) prediction shortcuts as well as their regularization strength $\lambda$, (b) semantic regularization strength $\eta$, and (c) number of base vectors $K$ affect a model's robustness to spurious bias. We report the worst-group accuracy on the CheXpert dataset.

in-distribution and out-of-distribution performance.

## 5.5 Ablation Studies

**Prediction shortcuts.** We evaluated the effectiveness of prediction shortcuts in Fig. 3(a). We began by using a randomly initialized shortcut detector to optimize the objective in Eq.(10). Additionally, we tested the performance without the spurious bias regularization term $\mathcal{L}_{\text{spu}}$, represented as $\lambda = 0$ in Fig. 3(a), as well as for $\lambda$ values of 1, 10, and 50. Our results indicate that the model performs the worst when prediction shortcuts are not used as regularization. Interestingly, our method still demonstrates effectiveness even when the prediction shortcuts are random. As $\lambda$ increases, the regularization strength decreases. The model achieves optimal performance when an appropriately balanced $\lambda$ is selected.

**Semantic regularization.** We analyzed the impact of the semantic regularization strength $\eta$ defined in Eq. (5). Fig. 3(b) shows that incorporating this regularization with a moderate value of $\eta$ enhances the model's robustness.

**Number of base vectors.** The number of base vectors, $K$, determines the representational capacity of the shortcut detector. A value of $K$ that is too small will limit the detector's ability to identify spurious attributes effectively, while an excessively large $K$ may lead to overfitting on the probe data. Notably, as shown in Fig. 3(c), the optimal value of $K$ is 6, which coincides with the true number of spurious attributes in the CheXpert dataset.

## 6 Conclusion

In this work, we proposed a novel post hoc framework to mitigate spurious biases without requiring group labels. Our approach first learns a shortcut detector in the latent space of a given model via a diverse probe set. To mitigate spurious biases, we retrained the model to be invariant to detected prediction shortcuts using a novel regularized training objective. We theoretically demonstrated that this objective effectively unlearns the spurious correlations captured during training. Experiments confirmed that our method successfully mitigates spurious biases and enhances model robustness to distribution shifts. Future work may explore constructing a more diverse probe set to further enhance the detection and mitigation of spurious biases.

## Acknowledgments

## References

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

[Bai *et al.*, 2021] Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6705–6713, 2021.

[Bombari and Mondelli, 2024] Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and ntk features. In *International Conference on Machine Learning*, 2024.

[Borkan *et al.*, 2019] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[Byrd and Lipton, 2019] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

[Carlucci *et al.*, 2019] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[Creager *et al.*, 2021] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Deng *et al.*, 2023] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. *arXiv preprint arXiv:2306.04949*, 2023.

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

[Geirhos *et al.*, 2020] 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.

[He and Garcia, 2009] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[He *et al.*, 2021] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[Ilyas *et al.*, 2019] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, 32, 2019.

[Irvin *et al.*, 2019] 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*, volume 33, pages 590–597, 2019.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Kim *et al.*, 2022] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. *Advances in Neural Information Processing Systems*, 35:18403–18415, 2022.

[Kirichenko *et al.*, 2023] 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*, 2023.

[Krueger *et al.*, 2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[LaBonte *et al.*, 2024] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for

group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36, 2024.

[Liu *et al.*, 2015] 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.

[Liu *et al.*, 2021] 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.

[Nam *et al.*, 2022] 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.

[Qiu *et al.*, 2023] 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.

[Sagawa *et al.*, 2019] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

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

[Tiwari and Shenoy, 2023a] Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34330–34343. PMLR, 23–29 Jul 2023.

[Tiwari and Shenoy, 2023b] 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.

[Vapnik, 1999] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 1999.

[Welinder *et al.*, 2010] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[Williams *et al.*, 2017] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[Yang *et al.*, 2023] 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.

[Ye *et al.*, 2024] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

[Zech *et al.*, 2018] 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.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[Zhang *et al.*, 2022] 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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26484–26516. PMLR, 17–23 Jul 2022.

[Zheng *et al.*, 2024a] Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Benchmarking spurious bias in few-shot image classifiers. In *European Conference on Computer Vision*, pages 346–364. Springer, 2024.

[Zheng *et al.*, 2024b] Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Learning robust classifiers with self-guided spurious correlation mitigation. In *The 33rd International Joint Conference on Artificial Intelligence*, 2024.

[Zheng *et al.*, 2024c] Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Spuriousness-aware meta-learning for learning robust classifiers. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4524–4535, 2024.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.