SLIM: Spuriousness Mitigation with Minimal Human Annotations

Xiwei Xuan[1], Ziquan Deng[1], Hsuan-Tien Lin[2], and Kwan-Liu Ma[1]

[1] University of California, Davis
[2] National Taiwan University
{xwxuan, ziqdeng, klma}@ucdavis.edu, htlin@csie.ntu.edu.tw

Abstract. Recent studies highlight that deep learning models often learn spurious features mistakenly linked to labels, compromising their reliability in real-world scenarios where such correlations do not hold. Despite the increasing research effort, existing solutions often face two main challenges: they either demand substantial annotations of spurious attributes, or they yield less competitive outcomes with expensive training when additional annotations are absent. In this paper, we introduce *SLIM*, a cost-effective and performance-targeted approach to reducing spurious correlations in deep learning. Our method leverages a human-in-the-loop protocol featuring a novel attention labeling mechanism with a constructed attention representation space. *SLIM* significantly reduces the need for exhaustive additional labeling, requiring human input for fewer than 3% of instances. By prioritizing data quality over complicated training strategies, *SLIM* curates a smaller yet more feature-balanced data subset, fostering the development of spuriousness-robust models. Experimental validations across key benchmarks demonstrate that *SLIM* competes with or exceeds the performance of leading methods while significantly reducing costs. The *SLIM* framework thus presents a promising path for developing reliable models more efficiently. Our code is available in `https://github.com/xiweix/SLIM.git/`.

Keywords: Model spuriousness · Human-in-the-loop · Data distribution

1 Introduction

Spurious correlations, where models mistakenly rely on irrelevant features to make decisions, pose a significant challenge in machine learning. Consequently, certain groups may experience deceptively inflated performance metrics, while accuracy reduces for others. This prevalence of "right answers for wrong reasons" fosters a misleading impression of model proficiency and constrains its utility across diverse contexts. Meanwhile, rectifying the issue of spuriousness is a challenging task, particularly when seeking to enhance the performance of underperforming groups. A notable example is the ISIC skin cancer dataset [5], where a typical spurious correlation occurs when color patches frequently co-occur with benign-labeled data. As shown in Fig. 1, a biased model may focus
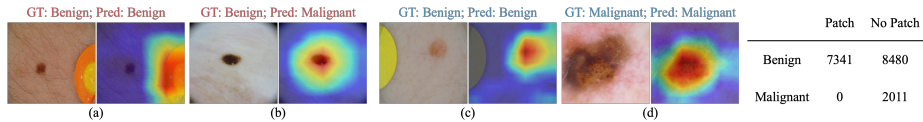
	Patch	No Patch
Benign	7341	8480
Malignant	0	2011

Fig. 1: (Left) GradCAM visualizations highlighting model attention on ISIC. (a, b) showcase a model biased towards patches, leading to a correct prediction with wrong reasons in (a) but an incorrect prediction despite correct focus in (b). (c, d) depict a spuriousness-robust model consistently focusing on core features. (Right) The table details the training data distribution by class (benign or malignant) and the presence of color patches, illustrating the imbalance and potential for spurious correlations.

on patches rather than skin lesions to identify melanoma, and fail when such patches are missing, while a desirable case would be the model captures correct skin regions regardless of spurious features' existence. Enhancing model robustness to address such dangerous misdiagnoses is crucial, driving a surge in research towards more effective solutions.

A typical approach in this context is to seek for optimal data distribution. GDRO [27] proves the efficacy of group-balanced data in enhancing model robustness. However, along with subsequent studies [14, 22], such an approach requires spuriousness labels at dataset scale, consuming a considerable cost. Our investigations reveal notable inconsistencies in human-annotated spuriousness labels, raising concerns about their relia-

Table 1: Annotation forms and quantities required by different methods on Waterbirds, highlighting the reduced workload of *SLIM*. (*Att.=Attention, T.=Training set, V.=Validation set*)

Method	Anno. Form	Anno. Amount
RRR [26]	Map	100%× T.
GradIA [9]	Att. Yes-No, Map	100%× T.
Energy Loss [24]	Bounding box	100%× T.
CRAYON [15]	Att. Yes-No	100%× T.
GDRO [27]	Spuriousness label	100%× T.
DFR [14]	Spuriousness label	100%× V.
SLIM (ours)	Att. Yes-No	2.5%×T. or 3%×V.

bility (refer to Table 6). There are other approaches with more direct human involvement, such as providing yes-no feedback on attention correctness [15, 36]; or outlining core features with maps [9, 26] or bounding boxes [24]. Despite different formats of queried information, they require extensive effort across the entire dataset, as summarized in Table 1. Considering the empirical applicability, our work aims to streamline this process by reducing human annotation workload and minimizing the ambiguity of collected data. Methods requiring spuriousness annotation tend to excel in performance at higher costs, while minimizing such expenses typically results in performance trade-offs [1, 6]. Facing this challenge, techniques such as JTT [17] and CNC [41] design training or data-augmenting phases to reduce spuriousness without additional annotations. However, the extra computational cost raised by multiple training iterations is undesired. We believe, as spuriousness is indeed caused by data biases, improving data quality is more essential than designing complex learning strategies.

Therefore, the remaining research gap calls for a data quality-oriented approach that balances performance and cost. This is where our framework, *SLIM*, steps in. *SLIM* offers a novel approach to mitigating spurious correlations by

intelligently leveraging a minimal set of human-annotated data and curating a feature-balanced data set to enhance model robustness without any significant computational overhead incurred by many existing methods. As shown in Fig. 2, inferred from a reference model, *SLIM* first constructs an attention representation space that reflects locally consistent data features and model attention, enabling efficient attention label estimation. — Representative instances can be sampled from this space for querying attention correctness labels and such labels can be expanded to neighbor instances. Subsequently, utilizing a visual explanation method that highlights model attention, like GradCAM [29], *SLIM* disentangles core and environment features within the latent space. Here, "core features" denote features critical for class determination, whereas "environment features" represent the surrounding contextual information. We agree with GDRO that balanced data distribution is crucial, but instead of using human-annotated labels to approximate spuriousness distribution, we leverage latent space features directly disentangled from model attention to balance data distribution, which is more aligned with model understanding and involving less human assumptions. Through this, *SLIM* assembles data groups with diverse feature combinations. Focused on data quality, a consistency-aware sampling mechanism is employed to curate a subset with optimal feature distributions — ensuring balanced core features across various environmental contexts. Finally, we conclude our framework with a straightforward training procedure.

The experimental results demonstrate *SLIM*'s advanced capability in spuriousness mitigation by performing comparably to, or exceeding state-of-the-art (SOTA) methods. Notably, *SLIM* accomplishes this with (1) a significantly reduced requirement for human annotations: 0.12%-2.5% of the training set size and 0.35%-4% of the validation set; and (2) a lightweight training for a robust model using our constructed data with 5%-30% of the original set. Our code will be published after the review process is complete. Our contributions are:

- **Data-quality-oriented spuriousness mitigation:** *SLIM* introduces an interactive, simple, and effective data construction pipeline to create high-quality data for spuriousness mitigation.
- **Cost-efficient framework:** In contrast to related works with leading performance, *SLIM* demonstrates a significant reduction in both annotation requirements and computational cost.
- **Competitive or superior model performance:** Without incurring excessive costs, *SLIM* achieves comparable or superior performance to existing methods, confirming its effectiveness in mitigating spurious correlations.

2 Related Work

Spuriousness mitigation with spuriousness labels. When a comprehensive set of spuriousness labels is accessible during training, methods like class balancing [11] and importance weighting [32] can be employed for mitigating spurious correlations, which utilizes data groups defined by class and spuriousness labels. Following the same schema, GDRO [27] refines the training approach

by focusing on the worst-performing groups as determined by ERM loss. Aiming to reduce the dependency on comprehensive spuriousness labeling, several methods [1, 6, 17, 21, 30, 32, 41] have emerged to approximate group information using a reference model. While these approaches are technically sophisticated, they often fail to achieve the same level of performance as approaches informed by spuriousness labels. To balance the tradeoff between performance and cost, a further research direction reduces the requirement for extensive labeling by leveraging spuriousness labels in smaller validation datasets [14, 22]. For example, SSA [22] adopts semi-supervised learning to predict group labels in training data according to information from validation, then applies GDRO to reduce spurious correlations. However, even with this reduced scale of validation sets, the labor-intensive nature of spuriousness labeling remains a barrier to scalability. Moreover, groups in these methods are formed based on human-defined criteria, risking a misalignment with the actual categories a model might discern. This drawback can hinder the true representation of spurious correlations and may lead to suboptimal mitigation outcomes.

Spuriousness mitigation with human-in-the-loop. Integrating human insights has proven beneficial in mitigating model spuriousness, as evidenced by various human-involved strategies. Considering the valuable human effort, [28] provides an efficient solution by querying data samples for issue mitigation with domain experts in the loop, which overlooks potential issues in the broader data scope. Another option for avoiding redundant human annotation is to query global spurious concepts [31, 38] from knowledgeable experts, which can be ambiguous and fail to cover all possibilities. More recent approaches [10, 24, 26] raise concerns about the inconsistency of spuriousness labels, introducing alternatives such as human-generated attention maps or bounding boxes outlining core features. These human-generated references help align the models' attention with the actual features relevant to decision-making. However, producing such annotations is challenging, and these techniques are resource-intensive as they necessitate annotations across the entire dataset. In response, more recent research research [2, 40] simplifies the annotation process to binary feedback on model's attention correctness. While less demanding, its requirement to annotate all instances still poses challenges in real practices.

3   Problem Formulation

We introduce the formal definition of the spuriousness mitigation problem [13]. Assume training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with input features $\boldsymbol{x}_i \in \mathbb{R}^d$ and labels $y_i \in [\mathcal{L}]$, where $\mathcal{L} = \{L_1, \ldots, L_l\}$ are the classes in $\mathcal{D}$. Machine learning models for classification are trained to minimize the empirical risk:

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} \ \mathbb{E}_{(\boldsymbol{x_i}, y_i) \in \mathcal{D}} [l(f(\boldsymbol{\theta}, \boldsymbol{x}_i), y_i)], \tag{1}$$

where $\boldsymbol{\theta}$ represent model parameters, $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$ and $l(*)$ are the model outputs and the loss function, respectively.

Groups. In spuriousness mitigation, we assume there exists a core feature set $\mathcal{C} = \{c_1, \ldots, c_l\}$ and a spurious feature set $\mathcal{S} = \{s_1, \ldots, s_k\}$ in $\mathcal{D}$. Each class L_i is uniquely identified by a core feature c_i, while spurious features s_j can be found across different classes. We consider an instance in the dataset as having both a core feature c_i and a spurious feature s_j, and each instance belongs to a unique group g_{c_i, s_j}. Core features are reliable class indicators in both training and testing sets, while spurious features with misleading co-occurrence with class in training may miss such correlations in the testing phase. If the model mistakenly links spurious features with classes in training, it may fail in testing when such spurious features are absent. This can lead to lower accuracy rates, especially for less common group variations that do not exhibit the spurious feature the model has learned to rely on.

Worst-group accuracy. The vital evaluation of spuriousness mitigation focuses on the model's performance in the subgroup where it is least effective – the worst-group accuracy. It is calculated by:

$$Acc_w = \min_{g \in \mathcal{G}}(\frac{1}{|g|} \sum_{(x_i, y_i) \in g} \mathbf{1}[f(x_i) = y_i]), \tag{2}$$

where $\mathcal{G} = \{\cup_{c_i, s_j} g_{c_i, s_j}\}$ is the union of subgroups defined by core feature c_i and spurious feature s_j, and $f(x_i)$ is the label predicted by the model. This metric is crucial because it directly assesses how effectively a model can generalize across different subgroups, especially those that might be mostly underrepresented because of spurious correlations. It ensures that improvements in model performance are not just reflective of gains on the majority or less challenging subsets of the data.

4   Method

In this section, we detail the method behind *SLIM*. In alignment with related works requiring no spuriousness label, our approach also follows a two-stage procedure: inferring groups with an ERM-trained reference model f_r, and then leveraging the inferred groups to train a robust model. As an approach oriented by data quality, *SLIM* framework comprises a three-phase data construction pipeline, as illustrated in Fig. 2. Firstly, we build an attention representation space to group instances with both similar data features and model attention, preparing for the data sampling and label propagation. Following this, we sample typical instances for human to evaluate model's attention correctness with yes-no feedback, then estimate neighboring instances' attention labels accordingly. In the last phase, we filter out data with wrong attention and leverage attention matrix to disentangle core and environment features. With constructed feature sets, we apply consistency-weighted data sampling to curate a feature-balanced subset, with a desired core-environment balanced distribution. Finally, *SLIM* framework is concluded by a training process with the curated subset, resulting in a model exhibiting robustness against spurious correlations.
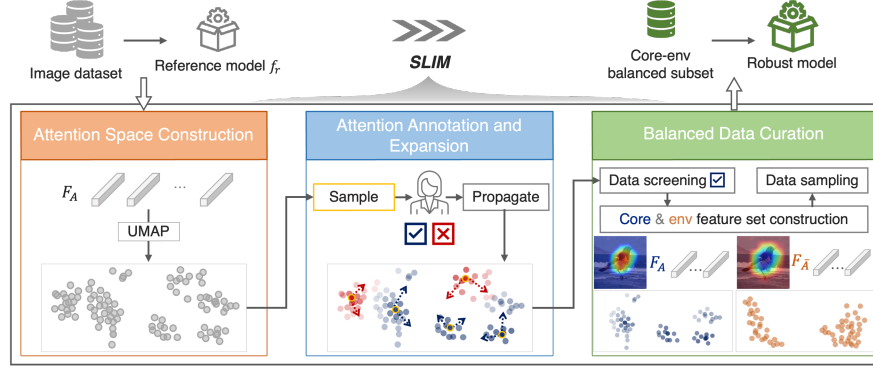
Fig. 2: Overview of *SLIM* framework, with a data construction pipeline consisting of three phases: (1) **Attention Space Construction**, creating a space with data features and model attention aligned locally; (2) **Attention Annotation and Expansion**, where instances sampled from the attention space are labeled by human for attention correctness, and labels are propagated to neighboring instances; and (3) **Balanced Data Curation**, which filters out instances with incorrect attention, and utilize attention-weighted (F_A) and inverse-attention-weighted ($F_{\bar{A}}$) feature vectors to create core and environment feature sets, forming subgroups to assemble a feature-balanced subset for training a spuriousness-robust model.

4.1   Attention Space Construction

Considering the benefits of being aware of spuriousness information and drawbacks of the annotation cost, in our framework, we deploy a binary labeling scheme to simplify descriptive labeling tasks into yes-no feedback on model's attention correctness. Moreover, we aim to further facilitate this process by avoiding the necessity of exhaustive labeling. Our empirical observations indicate that a model would have consistent attention mechanisms on specific data subgroups. If such subgroups can be found, it would largely aid the labeling workload by enabling label estimations among neighbors.

Driven by this, we aim to find a representation space where instances with similar attention outcomes (as indicated by highlighted image features) are closely grouped. However, our analysis reveals that conventional feature representation spaces formed directly from feature vectors do not fulfill this requirement – such spaces only maintain local consistency in data features but fail to capture the nuances of model attention (refer to Sec. 5.3).

Thus, we have developed a process utilizing model attention in the latent space to support the establishment of such space, as illustrated in Fig. 3. For each image, we obtain its feature vector F and model attribution A in the latent space at first. Note that A represents the model's attention focus and can be linearly interpolated to visualize an attention heatmap. We then use A as a weight matrix for F, producing an attention-weighted feature vector F_A. Finally, with the dimensionality reduction method UMAP [20], we project vectors F_A into a 2D space, namely the attention space, which ensures local consistency encompasses both data features and the model's attention mechanisms.
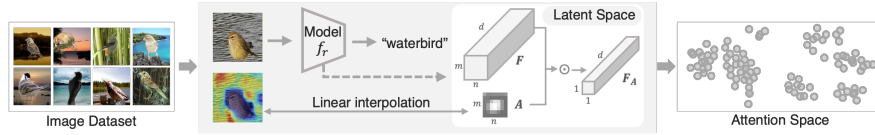
Fig. 3: Construction of the attention space. In the latent space, feature vector F and model's attribution vector A are extracted, representing input features and model's attention, representatively. By weighting F with A, the attention-weighted feature vector F_A emphasizes features of model's top interest. All F_A are then projected to form the attention space featuring locally consistent features and model attention.

Note that for a fair comparison, in alignment with baseline methods, experiments in this paper involve the CNN model (ResNet) for classification and GradCAM for obtaining model attention. But our framework is fully applicable to (1) other vision model architectures such as vision transformers (ViT) [8] or Swin transformers [18], and (2) other heatmap-based model visualization techniques, such as RISE [23] or ScoreCAM [33].

4.2 Attention Annotation and Expansion

Given that our attention space ensures locally consistent model attention, it allows for annotating a selectively sampled subset of instances and then estimating neighboring labels accordingly, thereby enabling efficient transfer of human expertise. The criteria for selecting these instances are twofold: (1) *diversity*, to encompass a broad covering of various data features and model attention patterns; (2) *typicality*, ensuring they are representative of their neighboring instances. To achieve this, we utilize the k-means clustering algorithm to partition our constructed attention space into n clusters (*diversity*). We then select n instances that are closest to each cluster's center (*typicality*). The choice of n is aimed at optimizing the balance between minimizing intra-cluster variance in data features and model attention patterns, and keeping the annotation budget manageable. Our empirical analysis, detailed in Sec. 5.3, reveals that the intra-cluster consistencies increase with n, but this trend levels off upon reaching a specific value of n–the elbow point. Thus, selecting n at this elbow point optimizes consistency without unnecessarily increasing the annotation workload.

We gather binary yes-no feedback from human experts on the correctness of the model's attention for these n instances, feeding this feedback along with our attention space embedding into the label spreading algorithm. This algorithm leverages data similarity and density to extend attention labels to previously unlabeled data [43]. Ultimately, this process yields a probability score for correct attention for each instance across the dataset.

4.3 Balanced Data Curation

Following Sec. 4.2 that provides attention accuracy scores, we proceed to construct our data subset to support model's unbiased learning, as described below:

Data screening. Inspired by works [3,7,12], our theoretical analysis further supports the premise that data receiving correct attention from a reference model are more likely to enhance the learning of core features in a robust model within a more balanced data distribution, as substantiated in Appx. A.2. Furthermore, our study indicates that even a reference model with inherent biases has the potential to accurately identify core features for a data subset. Such correctly captured features typically exhibit diversity, ensuring good coverage across the spectrum of the core feature set. This understanding informs our approach to curating a data subset that: (1) selects instances based on accurate model attention, and (2) ensures a balanced representation of core features across various environment features. Thus we first apply data screening to filter out instances with wrong attention according to attention accuracy scores.

Disentangling core and environment features. After screening, the retained data subset mainly comprises instances with correct attention. This is reflected by attention masks that highlight core features, meaning that the corresponding attribution vector A in the latent space also highlights core features of feature vector F. Moreover, existing studies have found that a model's feature vectors are representative of most input features, regardless of whether they are used for decision-making [14, 37]. Therefore, utilizing A as the weight matrix, we disentangle core and environment features in the latent space — weighting F by attention masks A and inverse attention masks $\bar{A}$. Subsequently, k-means clustering is applied to partition all attention-weighted feature vectors F_A, and inverse-attention-weighted feature vectors $F_{\bar{A}}$, respectively, facilitating the formation of representative subgroups for the core feature set ($\hat{\mathcal{C}} = \hat{c}_1, \ldots, \hat{c}_N$) and the environment feature set ($\mathcal{E} = e_1, \ldots, e_M$). The number of clusters, determined through the elbow method, culminates in $M \times N$ subgroups $g_{\hat{c}_i, e_j}$, each characterized by a distinct combination of core and environment features $(\hat{c}_i, e_j)$.

Data sampling. Given the unequal distribution of core and environment features, the internal consistency across each constructed subgroup $g_{\hat{c}_i, e_j}$ varies. Our sampling strategy aims to mitigate over-representation of particular features to prevent overfitting, while ensuring broad coverage of diverse core and environment features. We assess cluster consistency as a metric for determining the sample size from each cluster. For a cluster $\hat{c}_i$ in the core feature space, denote the distance between an instance x and the cluster center as $d_{\hat{c}_i}(x)$. The cluster consistency is defined as the reciprocal of the average distance per cluster:

$$\rho_{\hat{c}_i} = \frac{1}{\frac{1}{|\hat{c}_i|} \sum_{x \in \hat{c}_i} d_{\hat{c}_i}(x)}. \tag{3}$$

A lower $\rho_{\hat{c}_i}$ signals less consistent features, necessitating a higher sample weight for adequately representing them. A cluster's sampling weight is defined as $w_{\hat{c}_i} = \frac{1}{\rho_{\hat{c}_i}}$ and the number of samples to be drawn from a core cluster $\hat{c}_i$ is:

$$n_{\hat{c}_i} = N \cdot \frac{1/\rho_{\hat{c}_i}}{\sum_{j=1}^{|\hat{c}|} 1/\rho_{\hat{c}_j}}, \tag{4}$$

where N is the sampling budget. Moving to the environment feature space, the $n_{\hat{c}_i}$ instances from a core feature cluster $\hat{c}_i$ may distribute across multiple environment clusters $\{e_1, \ldots, e_M\}$ and formulate subgroups $\{g_{\hat{c}_i, e_1}, \ldots, g_{\hat{c}_i, e_M}\}$. A similar methodology is employed for a certain group $g_{\hat{c}_i, e_i}$. The average distance for $g_{\hat{c}_i, e_i}$ in the environment representation space is:

$$\rho_{\hat{c}_i, e_i} = \frac{1}{\frac{1}{|\hat{c}_i, e_i|} \sum_{x \in (\hat{c}_i, e_i)} d_{e_i}(x)}. \tag{5}$$

Similarly, we have the sampling weight $w_{\hat{c}_i, e_i} = \frac{1}{\rho_{\hat{c}_i, e_i}}$ and the number of samples to be drawn from a certain group $g_{\hat{c}_i, e_i}$ is finally defined by:

$$n_{\hat{c}_i, e_i} = n_{\hat{c}_i} \cdot \frac{w_{\hat{c}_i, e_i}}{\sum_{j=1}^{|\mathcal{E}(\hat{c}_i)|} w_{\hat{c}_j, e_i}} = N \cdot \frac{1/\rho_{\hat{c}_i}}{\sum_{j=1}^{|\hat{\mathcal{C}}|} 1/\rho_{\hat{c}_j}} \cdot \frac{1/\rho_{\hat{c}_i, e_i}}{\sum_{j=1}^{|\mathcal{E}(\hat{c}_i)|} 1/\rho_{\hat{c}_i, e_j}}, \tag{6}$$

where $|\hat{\mathcal{C}}|$ denotes the number of core clusters and $|\mathcal{E}(\hat{c}_i)|$ denotes the number of subgroups formulated by instances from cluster $\hat{c}_i$. To summary, our sampling strategy takes feature consistencies and sampling budget into consideration, and ensures fair representation of different core-environment feature combinations across the feature spaces.

4.4 Training

In our pursuit to mitigate spurious correlations, we emphasize the pivotal role of data quality by ensuring a balanced distribution of core features across various environmental contexts within our dataset. Given this focus, we employ a basic training approach for our model to learn from this well-prepared dataset.

We apply our data curation pipeline separately to the training ($SLIM_{Tr}$) and validation ($SLIM_{Val}$) sets. For the data constructed from the training set, we adopt the conventional ERM. For the dataset derived from the validation set, we employ the training technique presented by DFR [14]. It involves retraining only the final linear layer of a classification model that was initially trained using ERM. Note that our approach diverges from DFR practices by utilizing our curated dataset instead of a metagroup-balanced dataset.

5 Experiments

5.1 Experimental Setup

Datasets. Our study explores a range of datasets with inherent spurious correlations to challenge image classification models, including: (1) **Waterbirds** [27]. Contrasting landbirds and waterbirds against congruent/incongruent backgrounds, this dataset tests models' ability to focus on relevant features for bird type classification. (2) **CelebA** [19]. A comprehensive celebrity image dataset where the focus is on distinguishing hair color (blond vs. non-blond) to examine

biases in associating hair color with gender. (3) **ISIC** [5]. Aiming to differenti-
ate benign from malignant melanoma in dermoscopic images, we target colorful
patches as spurious features, aligning with previous studies [25]. (4) **NICO**. De-
rived from NICO++ [42], this dataset features various object categories in shifted
contexts to probe spurious correlations. (5) **ImageNet-9** [35]. This dataset is
crafted from ImageNet to test models' robustness against background variations.
Detailed information is available in Appx. A.4.

Baselines. We benchmark *SLIM* against a spectrum of SOTA spuriousness mit-
igation techniques, categorized as follows: (1) Techniques not reliant on spurious-
ness labels: EIIL [6], PGI [1], GEORGE [30], LfF [21], CIM [32], JTT [17], and
CNC [41]; (2) Techniques requiring additional information such as predefined
spuriousness concepts: DISC [34]; (3) Methods utilizing spuriousness labels from
validation sets: SSA [22] and DFR [14]; (4) Approaches utilizing spuriousness
labels from training sets: GDRO [27] and PDE [7].

Evaluation. Aligning with baselines, our experiments involve CNN models with
ResNet50 architecture and GradCAM for visualizing model attention. For the
capability of performance enhancement, we evaluate the worst-group and aver-
age accuracy. Additionally, we consider the necessity for extra labels and the
additional (add.) training cost to evaluate computational efficiency. Note that
the add. training cost is aligned with [13, 39], measuring add. data amounts re-
quired for second-stage training, where 1 means the amount of data equals the
training set size. Furthermore, we examine the attention precision to evaluate
whether model attention is corrected. All quantification experiments are based
on five independent runs with different random seeds.

More experimental results with other architecture (ViT) and datasets (MetaShift
[16] and FMoW [4]) are provided in Appx. A.8.

Table 2: Worst-group and average accuracy (%) comparison on binary classification
tasks, highlighting **best** and <u>second-best</u> performances. ♦ and ◊ indicate methods
requiring spuriousness concepts and attention labels, respectively.

Method	Spuriousness label	Add. training cost	Waterbirds		CelebA		ISIC	
			Worst	Avg	Worst	Avg	Worst	Avg
ERM	✗	0×	$62.6_{\pm0.3}$	$97.3_{\pm1.0}$	$47.7_{\pm2.1}$	$94.9_{\pm0.3}$	$56.7_{\pm2.2}$	$79.0_{\pm1.3}$
EIIL	✗	0×	$83.5_{\pm2.8}$	$94.2_{\pm1.3}$	$81.7_{\pm0.8}$	$85.7_{\pm0.1}$	$72.3_{\pm3.9}$	$86.4_{\pm1.7}$
PGI	✗	0×	$79.5_{\pm1.9}$	$95.5_{\pm0.8}$	$85.3_{\pm0.3}$	$87.3_{\pm0.1}$	$67.4_{\pm1.9}$	$85.7_{\pm1.1}$
GEORGE	✗	1×	$76.2_{\pm2.0}$	$95.7_{\pm0.5}$	$54.9_{\pm1.9}$	$94.6_{\pm0.2}$	$58.8_{\pm2.1}$	$83.2_{\pm1.4}$
LfF	✗	1×	$77.3_{\pm2.3}$	$91.4_{\pm1.7}$	$76.8_{\pm1.2}$	$84.5_{\pm0.7}$	$72.1_{\pm2.1}$	$85.4_{\pm0.5}$
CIM	✗	1×	$77.8_{\pm1.5}$	$94.7_{\pm1.1}$	$83.8_{\pm1.0}$	$90.5_{\pm0.6}$	$71.7_{\pm1.6}$	$85.9_{\pm1.1}$
JTT	✗	5×-6×	$83.1_{\pm3.5}$	$90.6_{\pm0.3}$	$81.5_{\pm1.7}$	$88.1_{\pm0.3}$	$28.1_{\pm4.3}$	$86.2_{\pm0.3}$
CNC	✗	2×-12×	$88.5_{\pm0.3}$	$90.9_{\pm0.1}$	$88.8_{\pm0.9}$	$89.9_{\pm0.5}$	$68.2_{\pm0.2}$	$84.6_{\pm0.2}$
DISC	✗ ♦	1×	$88.7_{\pm0.7}$	$93.8_{\pm0.7}$	$85.7_{\pm1.1}$	$89.9_{\pm0.7}$	$73.4_{\pm0.4}$	$84.7_{\pm0.4}$
$SLIM_{Tr}$	✗ ◊	0.1×-0.2×	$89.1_{\pm0.6}$	$91.7_{\pm0.5}$	<u>$89.4_{\pm0.3}$</u>	$90.1_{\pm0.2}$	**$79.1_{\pm0.5}$**	$88.5_{\pm0.4}$
$SLIM_{Val}$	✗ ◊	0.05×-0.3×	**$91.0_{\pm0.4}$**	$93.8_{\pm0.4}$	<u>$89.4_{\pm0.5}$</u>	$90.7_{\pm0.2}$	<u>$74.0_{\pm0.2}$</u>	$85.4_{\pm0.2}$
SSA	validation	1.5×-5×	$89.0_{\pm0.6}$	$92.2_{\pm0.9}$	**$89.8_{\pm1.3}$**	$92.8_{\pm0.1}$	$61.7_{\pm0.7}$	$77.7_{\pm1.0}$
DFR	validation	0.05×-0.3×	$89.6_{\pm0.7}$	$92.1_{\pm0.2}$	$87.3_{\pm1.0}$	$90.2_{\pm0.6}$	$71.9_{\pm0.7}$	$84.3_{\pm0.3}$
GDRO	training	0×	<u>$89.9_{\pm0.6}$</u>	$92.6_{\pm0.1}$	$88.9_{\pm1.3}$	$93.9_{\pm0.1}$	$66.8_{\pm3.5}$	$80.5_{\pm1.9}$
PDE	training	0×	$85.9_{\pm2.7}$	$90.5_{\pm1.8}$	$85.0_{\pm4.0}$	$92.3_{\pm0.8}$	$66.4_{\pm2.3}$	$82.2_{\pm1.3}$

Table 3: Worst-group and average accuracy (%) comparison on multi-classification tasks, highlighting **best** and <u>second-best</u> results. $\Diamond$ indicates methods requiring attention labels. - represents inapplicable methods because such methods require spuriousness labels while the dataset does not provide them.

Method	Spuriousness label	Add. training cost	NICO		ImageNet9	
			Worst	Avg	Worst	Avg
ERM	✗	0×	$11.9_{\pm 0.2}$	$70.9_{\pm 1.2}$	$55.3_{\pm 2.0}$	$75.6_{\pm 1.1}$
CNC	✗	5×-12×	$72.6_{\pm 1.7}$	$89.5_{\pm 0.6}$	$58.9_{\pm 0.6}$	$74.3_{\pm 0.2}$
$SLIM_{Tr}$	✗ $\Diamond$	0.1×-0.2×	$\mathbf{78.0_{\pm 0.8}}$	$\mathbf{92.1_{\pm 0.1}}$	$\mathbf{68.1_{\pm 0.3}}$	$\mathbf{74.5_{\pm 0.2}}$
$SLIM_{Val}$	✗ $\Diamond$	0.1×-0.2×	$73.8_{\pm 0.4}$	$90.7_{\pm 0.2}$	$64.6_{\pm 0.4}$	$74.2_{\pm 0.3}$
DFR	validation	0.4×	$70.9_{\pm 2.1}$	$85.8_{\pm 0.6}$	-	-
GDRO	training	0×	$73.3_{\pm 1.0}$	$90.3_{\pm 0.5}$	-	-

Table 4: *SLIM*'s primary cost, including the data amount required for attention annotation *(Att.)*, and the constructed data size for model training *(Training)*.

Method	Waterbird		CelebA		ISIC		NICO		ImageNet 9	
	Att.	Training	Att.	Training	Att.	Training	Att.	Training	Att.	Training
$SLIM_{Tr}$	120 (2.5%)	0.1×	200 (0.12%)	0.1×	120 (0.7%)	0.2×	40 (1.8%)	0.2×	180 (0.4%)	0.1×
$SLIM_{Val}$	40 (3%)	0.1×	70 (0.35%)	0.05×	40 (2%)	0.3×	20 (2.7%)	0.2×	50 (1%)	0.1×

5.2   Spuriousness Mitigation Performance

In this section, we present the performance evaluation by comparing different methods across multiple binary and multiclass classification datasets mentioned in Sec. 5.1. Key findings are demonstrated below.

Superior worst-group performance. The metric of worst-group accuracy (Acc_w) is pivotal for spuriousness mitigation methods (refer to Sec. 3), and the results in Tables 2 and 3 showcase *SLIM*'s superior capability in this regard. Across various classification tasks, *SLIM* consistently outperforms competing methods in achieving the highest Acc_w while maintaining high average accuracy. Specifically, in binary classification scenarios (Table 2), *SLIM* leads in performance for Waterbirds and ISIC and secures a close second in CelebA. In multi-classification tasks (Table 3), $SLIM_{Tr}$ and $SLIM_{Val}$ rank as the top performers across the board, underscoring the effectiveness of our approach in spuriousness mitigation.

Cost-efficient spuriousness mitigation. Methods in Tables 2 and 3 are divided into four groups from top to bottom, as per the categorization established in Sec. 5.1 according to the computational cost. The *"Add. training cost"* columns denote the additional training cost for mitigating spuriousness apart from the initial training, such as additional learning phases (CNC, SSA) or data augmentation (JTT). Results in the table substantiate that methods with higher costs (higher *Add. training costs* or more spuriousness labels) tend to outperform others, aligning with prior research findings. Nevertheless, among the well-performed methods, it is worth mentioning that *SLIM* is more cost-efficient: $SLIM_{Tr}$ used 0.1-0.2 proportion of training sets and $SLIM_{Val}$ used 0.05-0.3 proportion of validation sets for additional training, which is the same as DFR and much smaller than others. Moreover, *SLIM* requires labels for attention correctness, which is more cost-efficient to obtain than spuriousness labels (refer to Sec. 5.3 for more details). Table 4 presents *SLIM*'s primary cost, including the training data amount for mitigating spuriousness, which is the detailed version of *"Add. train-*

Table 5: Attention accuracy evaluation. (Left) GradCAM visualization of model attention. (Right) Worst group and average attention accuracy of selected methods.

Waterbirds	AIoU	
	Worst	Avg
ERM	$.25_{\pm.05}$	$.47_{\pm.03}$
CNC	$.48_{\pm.05}$	$.49_{\pm.04}$
GDRO	$.49_{\pm.03}$	$.52_{\pm.03}$
PDE	$.50_{\pm.02}$	$.52_{\pm.03}$
$SLIM_{Tr}$	$.53_{\pm.02}$	$.61_{\pm.02}$

ImageNet 9	AIoU	
	Worst	Avg
ERM	$.33_{\pm.07}$	$.55_{\pm.03}$
CNC	$.44_{\pm.03}$	$.61_{\pm.03}$
$SLIM_{Tr}$	$.58_{\pm.04}$	$.68_{\pm.03}$

ing cost" in Tables 2 and 3, and the data amount requiring attention labels *(Att.)*. Notably, our approach only requires no more than 3% of the corresponding data split (train or validation) for attention-label annotation, further proving the cost-efficiency of *SLIM*.

Enhanced attention accuracy. Beyond classification accuracy, we further evaluate the performance of spuriousness mitigation methods via attention accuracy. Specifically, for models resulting from each method, we randomly sample data instances to visualize model attention, and measure attention accuracy scores in both worst-group and dataset levels, consistent with our evaluation of classification accuracy. Align with the related work [40], we employ the Adjusted Intersection-over-Union (AIoU) score as our attention accuracy measurement, which is an adjusted IoU score with a threshold, aiming at measuring the overlap between model's highest attributed region and the ground-truth bounding box. Details related to AIoU are given in Appx. A.7. Table 5 presents the evaluation results. The left GradCAM examples showcase *SLIM*'s capability of correcting model's wrong attention, and the right table compares *SLIM* with SOTA methods, where we can find our method secures the best performance in terms of both worst-group and average attention accuracy.

5.3   Attention Labeling Performance

To compare the processes of annotating spuriousness versus attention correctness, we employed crowdsourcing tasks using the Waterbirds and NICO datasets, as both of which come with ground-truth spuriousness labels. For this study, we selected a random set of 120 images from each dataset and established two separate tasks: one for spuriousness labeling and one for attention correctness labeling. For each task, we recruited 60 workers independently from the Prolific platform to prevent learning biases from cross-task participation. We meticulously recorded the time taken to label each image and reviewed all submissions to exclude any outliers with unusually long or short annotation times. Ensuring ethical research standards, our study avoided collecting personally identifiable information and excluded any potentially offensive content. Detailed instructions provided to the crowdsourcing participants can be found in Appx. A.6.

Evaluation Consistency. Table 6 presents metrics for this study. While the annotation time for Waterbirds remains similar across tasks, attention correctness for NICO is annotated more quickly than the other. This difference may stem

Table 6: Comparison of crowd-sourced labeling tasks: spuriousness vs. attention correctness.

Waterbirds	Avg. Time	Avg. Consis.
Spuriousness	$4.9_{\pm2.3}$ sec	$81.9_{\pm14.8}\%$
Att.	$4.5_{\pm1.4}$ sec	$94.6_{\pm10.5}\%$
NICO	Avg. Time	Avg. Consis.
Spuriousness	$7.9_{\pm3.8}$ sec	$70.1_{\pm19.1}\%$
Att.	$5.3_{\pm1.6}$ sec	$95.1_{\pm6.1}\%$
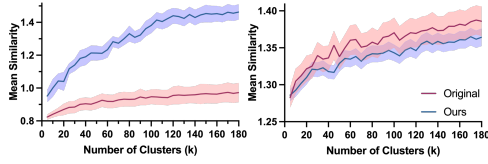
Fig. 4: Comparison of neighbor consistency over original feature space and our attention space on Waterbirds, w.r.t. model attention (left) and data features (right).

from NICO's broader array of spuriousness choices versus the binary attention annotations. To measure the annotation consistency—whether a single receives divergent labels—we use the percentage of the most frequent label as a consensus measure. A higher percentage suggests greater consistency. Our findings indicate that attention correctness annotations exhibit a notably higher consistency, suggesting that it is less ambiguous to determine correct attention than to identify spurious features.

Attention Consistency. *SLIM* requires spreading labels in a space where attention patterns are consistently similar in nearby areas. To evaluate its feasibility, we compared how similar neighbors are in the original feature space versus the attention space created by *SLIM*. We used k-means clustering with varying numbers of clusters to measure the intra-cluster similarity for both the model's attention matrices and the feature vectors. A higher similarity score means the space has more consistent patterns, which is essential for the accurate spreading of labels. As Fig. 4 shows, our attention space significantly improves consistency in the model's attention, while also maintaining a good level of data feature consistency. Increasing the number of clusters generally increases consistency, but there is a point where adding more clusters does not lead to further gains, as shown by the curve flattening at k=120 in the left plot.

In summary, we have demonstrated that the annotation module of *SLIM* is less time-consuming and yields more consistent labeling feedback. Additionally, the attention space engineered by *SLIM* establishes a solid foundation for propagating attention labels, considering the consistent model attention observed among neighboring instances.

5.4   Ablation Study

We investigate two critical factors of *SLIM*, which are detailed below. Other factors such as the dimension of the attention space are discussed in Appx. A.8.
Influence of data amount for attention annotations. We change the data amount for attention annotations to investigate their effect on performance, keeping other variables constant. Table 7 illustrates that both the worst-group and average accuracy tend to rise with more annotations, while such incremental benefit stabilizes after reaching a certain point, helping us identify an optimal number of annotations that balance performance with annotation cost.
Influence of training set size. This size refers to the size of the data subset constructed in *SLIM*'s final phase for training a robust model. By only varying

Table 7: Performance variation with different attention annotation amounts in *SLIM*. Marks in table: **best** and <u>second best</u> performance; the amounts used in Table 2.

Method	Waterbirds			CelebA			ISIC		
	Amount	Worst	Avg	Amount	Worst	Avg	Amount	Worst	Avg
$SLIM_{Tr}$	60	$78.2_{\pm0.5}$	$86.0_{\pm0.3}$	100	$86.4_{\pm0.2}$	$90.2_{\pm0.1}$	60	$75.3_{\pm0.2}$	$86.6_{\pm0.9}$
	90	$85.9_{\pm0.9}$	$89.3_{\pm0.3}$	150	$87.8_{\pm0.4}$	$91.1_{\pm0.7}$	90	$74.4_{\pm0.4}$	$87.4_{\pm0.7}$
	120	$89.1_{\pm0.6}$	$91.7_{\pm0.5}$	200	$89.4_{\pm0.4}$	$90.1_{\pm0.2}$	120	$79.1_{\pm0.5}$	$88.5_{\pm0.4}$
	150	$\mathbf{89.4_{\pm0.6}}$	$91.7_{\pm0.2}$	250	$\underline{89.4_{\pm0.2}}$	$90.9_{\pm0.1}$	150	$\mathbf{79.9_{\pm0.8}}$	$88.8_{\pm0.7}$
	180	$\underline{89.3_{\pm0.3}}$	$91.0_{\pm1.5}$	300	$\mathbf{89.5_{\pm0.1}}$	$91.2_{\pm0.2}$	180	$\underline{79.8_{\pm0.3}}$	$88.6_{\pm0.3}$
$SLIM_{Val}$	10	$78.9_{\pm8.7}$	$86.5_{\pm5.1}$	40	$84.7_{\pm0.4}$	$87.4_{\pm0.3}$	10	$70.5_{\pm0.5}$	$83.3_{\pm1.2}$
	25	$89.2_{\pm0.6}$	$93.4_{\pm1.0}$	55	$87.6_{\pm0.3}$	$90.4_{\pm0.6}$	25	$72.7_{\pm0.1}$	$85.2_{\pm0.8}$
	40	$91.0_{\pm0.4}$	$93.8_{\pm0.4}$	70	$89.4_{\pm0.4}$	$90.7_{\pm0.2}$	40	$74.0_{\pm0.2}$	$85.4_{\pm0.2}$
	55	$\mathbf{91.5_{\pm0.3}}$	$93.8_{\pm0.1}$	85	$\mathbf{89.7_{\pm0.1}}$	$90.5_{\pm0.4}$	55	$\underline{74.0_{\pm0.2}}$	$85.9_{\pm0.5}$
	70	$90.7_{\pm0.3}$	$94.0_{\pm0.3}$	100	$\underline{89.7_{\pm0.2}}$	$90.5_{\pm0.2}$	70	$\mathbf{74.1_{\pm0.2}}$	$85.8_{\pm0.2}$

Table 8: Performance variation with different training sizes. $SLIM_{Tr}$ and $SLIM_{Val}$ are trained on different proportions (Prop) of training sets and validation sets, respectively.

Method	Waterbirds			CelebA			ISIC		
	Prop%	Worst	Avg	Prop%	Worst	Avg	Prop%	Worst	Avg
$SLIM_{Tr}$	5	$82.4_{\pm1.0}$	$83.6_{\pm1.3}$	5	$88.9_{\pm0.1}$	$90.0_{\pm0.1}$	15	$77.6_{\pm2.1}$	$87.1_{\pm1.2}$
	10	$\mathbf{89.1_{\pm0.6}}$	$91.7_{\pm0.5}$	10	$\mathbf{89.4_{\pm0.3}}$	$90.1_{\pm0.2}$	20	$\mathbf{79.1_{\pm0.5}}$	$88.5_{\pm0.4}$
	12	$82.3_{\pm1.3}$	$87.8_{\pm0.3}$	15	$89.4_{\pm0.4}$	$91.0_{\pm0.4}$	25	$77.4_{\pm1.5}$	$88.2_{\pm1.2}$
	15	$83.5_{\pm2.7}$	$89.6_{\pm0.5}$	20	$88.9_{\pm0.1}$	$90.8_{\pm0.3}$	30	$78.3_{\pm1.5}$	$88.3_{\pm0.9}$
$SLIM_{Val}$	5	$84.2_{\pm3.6}$	$89.2_{\pm2.6}$	3	$85.7_{\pm1.3}$	$88.5_{\pm1.1}$	25	$68.2_{\pm2.1}$	$81.9_{\pm0.8}$
	10	$\mathbf{91.0_{\pm0.4}}$	$93.8_{\pm0.4}$	5	$\mathbf{89.4_{\pm0.4}}$	$90.7_{\pm0.2}$	30	$\mathbf{74.0_{\pm0.2}}$	$85.4_{\pm0.2}$
	12	$87.6_{\pm0.6}$	$93.9_{\pm1.3}$	10	$88.3_{\pm1.4}$	$90.3_{\pm0.5}$	35	$72.5_{\pm0.5}$	$85.3_{\pm0.1}$
	15	$86.8_{\pm2.0}$	$94.6_{\pm0.5}$	15	$87.2_{\pm0.4}$	$90.3_{\pm0.4}$	40	$70.9_{\pm1.2}$	$85.4_{\pm3.5}$

this parameter (Table 8), we find that the optimal size is inherently linked to the size of our constructed subgroups, $|g_{\hat{c}_i, e_i}|$. An optimal size is reached when data from various groups are sufficiently sampled without redundant oversampling. These findings guide the hyperparameter selection in this phase, ensuring an effective cost allocation.

6   Conclusion

In this paper, we introduce *SLIM*, a novel framework for spuriousness mitigating in deep learning. With minimal human annotation effort, *SLIM* provides a data construction pipeline to create a feature-balanced subset for training a robust model. Extensive evaluation demonstrates the superior cost-effectiveness of *SLIM*, which matches or outperforms state-of-the-art methods with far less costs. The success of *SLIM* underscores the potential of data quality and efficient human supervision in the development of robust models. We believe that *SLIM* will inspire further research in this direction, ultimately leading to more cost-effective machine learning systems for reliable AI.

References

1. Ahmed, F., Bengio, Y., Van Seijen, H., Courville, A.: Systematic generalisation with group invariant predictions. In: International Conference on Learning Representations (2020)
2. Bontempelli, A., Teso, S., Tentori, K., Giunchiglia, F., Passerini, A.: Concept-level debugging of part-prototype networks. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=oiwXWPDTyNk`
3. Chen, Z., Deng, Y., Wu, Y., Gu, Q., Li, Y.: Towards understanding the mixture-of-experts layer in deep learning. Advances in neural information processing systems **35**, 23049–23062 (2022)
4. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: CVPR (2018)
5. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172 (2018). `https://doi.org/10.1109/ISBI.2018.8363547`
6. Creager, E., Jacobsen, J.H., Zemel, R.: Environment inference for invariant learning. In: International Conference on Machine Learning. pp. 2189–2200. PMLR (2021)
7. Deng, Y., Yang, Y., Mirzasoleiman, B., Gu, Q.: Robust learning with progressive data expansion against spurious correlation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), `https://openreview.net/forum?id=9QEVJ9qm46`
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Gao, Y., Sun, T.S., Zhao, L., Hong, S.R.: Aligning eyes between humans and deep neural network through interactive attention alignment. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2) (nov 2022). `https://doi.org/10.1145/3555590`, `https://doi.org/10.1145/3555590`
10. Gao, Y., Sun, T.S., Zhao, L., Hong, S.R.: Aligning eyes between humans and deep neural network through interactive attention alignment. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–28 (2022)
11. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009)
12. Jelassi, S., Li, Y.: Towards understanding how momentum improves generalization in deep learning. In: International Conference on Machine Learning. pp. 9965–10040. PMLR (2022)
13. Joshi, S., Yang, Y., Xue, Y., Yang, W., Mirzasoleiman, B.: Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. arXiv preprint arXiv:2306.11957 (2023)
14. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. arXiv preprint arXiv:2204.02937 (2022)
15. Lee, S., Payani, A., Chau, D.H.P.: Towards mitigating spurious correlations in image classifiers with simple yes-no feedback. In: AI and HCI Workshop at the 40th International Conference on Machine Learning (2023)

16. Liang, W., Zou, J.: Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=MTex8qKavoS`

17. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: International Conference on Machine Learning. pp. 6781–6792. PMLR (2021)

18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)

20. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

21. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems **33**, 20673–20684 (2020)

22. Nam, J., Kim, J., Lee, J., Shin, J.: Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. arXiv preprint arXiv:2204.02070 (2022)

23. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)

24. Rao, S., Böhle, M., Parchami-Araghi, A., Schiele, B.: Using explanations to guide models. arXiv preprint arXiv:2303.11932 (2023)

25. Rieger, L., Singh, C., Murdoch, W.J., Yu, B.: Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)

26. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 2662–2670 (2017). `https://doi.org/10.24963/ijcai.2017/371`, `https://doi.org/10.24963/ijcai.2017/371`

27. Sagawa*, S., Koh*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020), `https://openreview.net/forum?id=ryxGuJrFvS`

28. Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K., Kersting, K.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence **2**(8), 476–486 (2020)

29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

30. Sohoni, N., Dunnmon, J., Angus, G., Gu, A., Ré, C.: No subclass left behind: Fine-grained robustness in coarse-grained classification problems. Advances in Neural Information Processing Systems **33**, 19339–19352 (2020)

31. Srivastava, M., Hashimoto, T., Liang, P.: Robustness to spurious correlations via human annotations. In: International Conference on Machine Learning. pp. 9109–9119. PMLR (2020)

32. Taghanaki, S.A., Choi, K., Khasahmadi, A.H., Goyal, A.: Robust representation learning via perceptual similarity metrics. In: International Conference on Machine Learning. pp. 10043–10053. PMLR (2021)

33. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 24–25 (2020)

34. Wu, S., Yuksekgonul, M., Zhang, L., Zou, J.: Discover and cure: Concept-aware mitigation of spurious correlation. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)

35. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. ArXiv preprint arXiv:2006.09994 (2020)

36. Xuan, X., Ono, J.P., Gou, L., Ma, K.L., Ren, L.: Attributionscanner: A visual analytics system for metadata-free data-slicing based model validation. arXiv preprint arXiv:2401.06462 (2024)

37. Xue, Y., Payani, A., Yang, Y., Mirzasoleiman, B.: Eliminating spurious correlations from pre-trained models via data mixing. arXiv preprint arXiv:2305.14521 (2023)

38. Yan, S., Yu, Z., Zhang, X., Mahapatra, D., Chandra, S.S., Janda, M., Soyer, P., Ge, Z.: Towards trustable skin cancer diagnosis via rewriting model's decision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11568–11577 (2023)

39. Yang, Y., Gan, E., Karolina Dziugaite, G., Mirzasoleiman, B.: Identifying spurious biases early in training through the lens of simplicity bias. In: Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 238, pp. 2953–2961. PMLR (02–04 May 2024)

40. Yang, Y., Nushi, B., Palangi, H., Mirzasoleiman, B.: Mitigating spurious correlations in multi-modal models during fine-tuning. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)

41. Zhang, M., Sohoni, N.S., Zhang, H.R., Finn, C., Ré, C.: Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. arXiv preprint arXiv:2203.01517 (2022)

42. Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., Cui, P.: Nico++: Towards better benchmarking for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16036–16047 (2023)

43. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in neural information processing systems **16** (2003)