Discover and Mitigate Multiple Biased Subgroups in Image Classifiers

*Zeliang Zhang *Mingqian Feng †Zhiheng Li Chenliang Xu
University of Rochester
{zeliang.zhang,mingqian.feng,zhiheng.li,chenliang.xu}@rochester.edu

Abstract

Machine learning models can perform well on in-distribution data but often fail on biased subgroups that are underrepresented in the training data, hindering the robustness of models for reliable applications. Such subgroups are typically unknown due to the absence of subgroup labels. Discovering biased subgroups is the key to understanding models' failure modes and further improving models' robustness. Most previous works of subgroup discovery make an implicit assumption that models only underperform on a single biased subgroup, which does not hold on in-the-wild data where multiple biased subgroups exist.

In this work, we propose Decomposition, Interpretation, and Mitigation (DIM), a novel method to address a more challenging but also more practical problem of discovering multiple biased subgroups in image classifiers. Our approach decomposes the image features into multiple components that represent multiple subgroups. This decomposition is achieved via a bilinear dimension reduction method, Partial Least Square (PLS), guided by useful supervision from the image classifier. We further interpret the semantic meaning of each subgroup component by generating natural language descriptions using vision-language foundation models. Finally, DIM mitigates multiple biased subgroups simultaneously via two strategies, including the data- and model-centric strategies. Extensive experiments on CIFAR-100 and Breeds datasets demonstrate the effectiveness of DIM in discovering and mitigating multiple biased subgroups. Furthermore, DIM uncovers the failure modes of the classifier on Hard ImageNet, showcasing its broader applicability to understanding model bias in image classifiers. The code is available at https://github.com/ZhangAIPI/DIM.

1. Introduction

Machine learning models can achieve overall good performance on in-distribution data [10, 13]. However, they often underperform on certain biased subgroups that are underrepresented in the training data, undermining model robustness against group distributional shifts [39]. For instance, ResNet [9] trained on ImageNet [5] fails to recognize the "balance beam" when kids are not present, where the image classifier is biased towards the subgroup of "balance beam" when kids are present (*cf*. Fig. 5). Therefore, identifying and mitigating subgroup biases in image classifiers is crucial to improve models' reliability and robustness [16, 19, 28, 48, 49].

However, previous research on model bias has many limitations. First, most existing methods [3, 15, 50] for bias discovery require structured attributes, demanding expensive human labor to collect and label data. The impracticality of annotating every possible attribute also leads to potential biases remaining hidden [23]. Second, few works involve mining the unknown multi-bias. While Eyuboglu *et al.* [8] approach this problem by employing a Gaussian Mixed Model [2] to detect biases through clustering, their method relies on proxies of clusters to explain unknown biases, which can result in interpretation distortion (see the Domino part of Fig. 3 in Sec. 5.2). To address this issue,

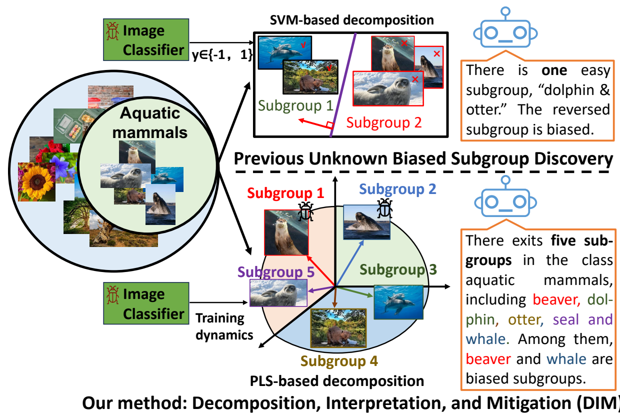

Figure 1. While the previous method [12] exploits the SVM to detect the single biased subgroup using the classification correctness on samples, we propose to integrate the training dynamics of biased image classifiers as the supervision into PLS decomposition to discover multiple unknown subgroups. This allows us to further subtly interpret discovered subgroups and precisely mitigate biases.

*Equal contribution.
†Project lead. Work done before Zhiheng Li joining Amazon.

recent work by Jain *et al.* [12] proposes directly distilling failure modes as firsthand directions in latent feature space and using cross-modal representation [51] for bias interpretation. However, this method only detects a single shortcut and falls short in multi-shortcut discovery. Compared with single bias, the discovery of multi-bias is more practical and challenging [21, 46]. The lack of efficient bias discovery techniques impedes the application of bias mitigation methods, which typically require the knowledge of bias attributes [39, 44]. Thus, efforts on multiple unknown biases are essential and challenging to enable practical and scalable solutions to trustworthy AI.

In this paper, we are motivated by the insight that a single unknown bias can be identified in the latent space [12], and then move forward to addressing a more challenging problem, *multiple unknown biased subgroups*. Although models are designed to grasp predefined classes (labels) during training, they also inadvertently learn subgroups distinguished by unconscious attributes. The distribution difference between subgroups and the presence of multi-bias result in image classifiers exhibiting erratic performance across subgroups. As shown in Fig. 1, the class of "aquatic mammals" in the CIFAR-100 dataset [17] can be further divided into 5 subgroups, "beaver," "dolphin," "otter," "seal," and "whale." While a ResNet-18 [9] has an overall accuracy of 54.6% on this "aquatic mammals" class, it performs poorly on "beaver" (34%) and "whale" (37%). This disparity manifests the existence of subgroup bias and its impairment of the robustness of the model. Limited to simply using sample correctness, Jain *et al.* [12] can discover only two broad subgroups: those with superior and inferior performance. Specifically, it discerns the positive group of two well-performance subgroups, "dolphin" and "otter," through a unified direction in latent space. However, the opposite direction, supposed to embody biased subgroups (*i.e.*, "beaver" and "whale"), fails to maintain representativeness and interpretability (see the Jain *et al.* [12] part of Fig. 3 in Sec. 5.2).

Addressing the problem of *multiple unknown biased subgroups* presents three challenges. The first difficulty lies in the lack of explicit supervisory signals from the model to guide the discovery process. The second hurdle is interpreting identified directions and pinpointing biased ones among them. Third, beyond the identification and interpretation, there is a consequential line of inquiry into how these discerned subgroups can be harnessed for bias mitigation.

In response to these challenges, we propose an innovative *latent space-based multiple unknown biased subgroup discovery* method, named **DIM** (**D**ecomposition, **I**nterpretation, and **M**itigation). Initially, we integrate the training dynamics of the biased model into the partial least squares (PLS) [1] to supervise the decomposition of image features in latent space. Those features are decomposed into different subgroup directions, each aligning with different subgroups that the image classifier learned. Subsequently, these directions are utilized to generate pseudo-subgroup labels for the original dataset to distinguish biased subgroups that exhibit lower accuracy. Finally, upon identifying multiple subgroups, including biased ones, DIM employs cross-modal embeddings in the latent space to interpret these subgroups and annotate the data with subgroup information to mitigate biases.

Our work presents three key contributions as follows,

1. We formulate the problem of discovering multiple unknown biased subgroups and subsequent bias mitigation.
2. We propose **DIM** (**D**ecomposition, **I**nterpretation, and **M**itigation), a novel framework used to discover, understand, and mitigate multiple unknown biased subgroups learned by image classifies.
3. We conduct experiments on three datasets: CIFAR-100, Breeds, and Hard ImageNet. We verify DIM's ability to detect biased subgroups on CIFAR-100 and Breeds, where classes and ground-truth subgroups, including biased ones, are given. For Hard ImageNet, we apply DIM to discover biased subgroups implicitly learned by the image classifier, thereby illustrating its failure modes.

2. Related Work

Bias Identification Many works have been proposed to identify and explain the bias of deep learning models. Eyuboglu *et al.* [8] leverage cross-modal embeddings and a novel error-aware model to discover underperforming slices of samples. Singla and Feizi [42] use the activation maps for neural features to highlight spurious or core visual features and introduce an ImageNet-based dataset, Salient ImageNet, which contains masks of core visual and spurious features. Zhu *et al.* [52] propose a training-free framework, GSCLIP, to explain the dataset-level distribution shifts. Li and Xu [23] and Lang *et al.* [18] use a generative model to discover and interpret unknown biases. Jain *et al.* [12] harness the linear classifier to identify models' failure mode and uses CLIP [37] for the automatic caption to explain the failure mode. Previous methods usually fall on the single-shortcut problems. However, real-world scenes usually involve multiple biases, including multiple subgroup biases, posing challenges to existing methods. Our work proposes a novel approach to identify and explain multiple unknown subgroup biases, which is scalable to large real-world datasets.

Bias Mitigation There are many approaches proposed to mitigate the bias, such as the re-sampling and weighting strategy [20, 36], distributional robust optimization [43, 45], invariant risk minimization [29], and adversarial debiasing [24, 47]. Some work also exploits the identified bias to mitigate the bias of models, such as EIIL [4] and LfF [31]. However, these methods require the bias labels in the training dataset, which are usually unknown in practice, leading to poor scalability. To address this issue, some work has proposed research on bias mitigation without access to bias annotation. Nam *et al.* [32] propose to train a debiased model on samples, which is against the prejudice of the well-trained bias model. Pezeshki *et al.* [35] propose a regularization term to decouple failure learning dynamics. Li *et al.* [22] propose the debiasing alternate networks to discover unknown biases and unlearn the multiple identified biases for the classifier. Park *et al.* [34] propose the debiased contrastive weight pruning to investigate unbiased networks. In our work, we employ identified bias subgroup information to implement data-centric and model-centric strategies for bias mitigation, improving the model's robustness.

3. Problem Formulation

Consider an image classifier trained on the data $\boldsymbol{x} \in \mathcal{X}$ with annotated ground-truth label $y \in \mathcal{Y}$. In this context, assume there are L annotated classes in the dataset, making $\mathcal{Y} = \{1, 2, ..., L\}$. We hypothesize each class contains G subgroups[1] that share certain characteristics or features, a total of $L \times G$ subgroups. For each input $\boldsymbol{x}$, we denote its subgroup membership as $g \in \{1, 2, ..., G\}$. The input data, labels, and subgroups are drawn from a joint distribution P.

An ideal classifier without subgroup bias should maintain consistent performance across all G subgroups. Conversely, a biased classifier is characterized by its inferior performance on specific subgroups. The primary objective is twofold: first, to discover a total of $L \times G$ subgroups, and second, to identify the biased k subgroups (out of G) in each class, which exhibit lower classification accuracy than the median. The insight that mitigating one bias in an image classifier with multiple biases can inadvertently amplify others [21] underscores the importance of addressing multiple unknown biases. Hence, we particularly focus on multiple unknown biased subgroups ($k \geq 2$), which implies $G \geq 4$.

Following the identification, the subsequent tasks involve interpreting and mitigating these biases. This progression is key to not only understanding but also enhancing the robustness of image classifiers against subgroup biases.

4. Method

In the quest to address multiple unknown biased subgroups, we propose a novel method, **DIM** (**D**ecomposition, **I**nterpretation, and **M**itigation), as shown in Fig. 2.

4.1. Decomposition

Discovering Multiple Subgroups. The previous study [25] has identified that *it is theoretically impossible to derive invariant features from the heterogeneous data without environment information* [27, 38]. This implies that without any additional information, unsupervised biased subgroup discovery is also impossible. Motivated by this, we integrate the model supervision into the partial least square (PLS) [1] method to decompose the image features into multiple subgroup directions in the latent space. Unlike the principal component analysis (PCA) [14], which is unsupervised decomposition, PLS can be used to derive the components within the inputs mostly aligned with the supervision.

Concretely, there are three steps to discovering multiple unknown subgroups in a specific class (fixing $y = l$). Initially, we use the CLIP image encoder, a function $f_{\text{image}}: \mathcal{X} \to \mathbb{R}^d$, where d denotes the latent feature dimension, to encode images in this class, which can be considered as a variable that follows specific empirical class distribution $\hat{P}_l$. The images are encoded into input embeddings $\widehat{\boldsymbol{x}} := f_{\text{image}}(\boldsymbol{x})$ in the latent space. Next, for each image $\boldsymbol{x}$, we collect some information provided by the studied model, denoted as $\boldsymbol{z}_{\boldsymbol{x}} \in \mathbb{R}^M$, where M is the number of used information. The information, including the loss, correctness, logit, *etc.*, serves as supervision guiding the decomposition. Subsequently, we apply the PLS method to model the decomposition of $\widehat{\boldsymbol{x}}$ supervised by $\boldsymbol{z}_{\boldsymbol{x}}$. The core idea is to search for subgroup directions in the latent space that maximize the correlation with principal components in the supervision. In discovering the first subgroup, $i = 1$, setting $\widehat{\boldsymbol{x}}_1 := \widehat{\boldsymbol{x}}$ and $\boldsymbol{z}_{\boldsymbol{x},1} := \boldsymbol{z}_{\boldsymbol{x}}$, it can be formulated as the following optimization target,

$$\max_{\boldsymbol{w}_i,\, \boldsymbol{h}_i} \mathbb{E}_{\hat{P}_l}(u_i v_i) = \max_{\boldsymbol{w}_i,\, \boldsymbol{h}_i} \mathbb{E}_{\hat{P}_l}\left((\boldsymbol{w}_i^T \widehat{\boldsymbol{x}}_i)(\boldsymbol{h}_i^T \boldsymbol{z}_{\boldsymbol{x},i})\right), \tag{1}$$

where $\boldsymbol{w}_i \in \mathbb{R}^d$, $\boldsymbol{h}_i \in \mathbb{R}^M$ are the normalized directions and $u_i := \boldsymbol{w}_i^T \widehat{\boldsymbol{x}}_i$, $v_i := \boldsymbol{h}_i^T \boldsymbol{z}_{\boldsymbol{x},i}$ represent the latent scores on the directions. For each image, a high latent score implies a high similarity to the discovered subgroup. Upon the optimization, $\widehat{\boldsymbol{x}}_i$ and $\boldsymbol{z}_{\boldsymbol{x},i}$ are updated by subtracting the discovered information. It can be achieved by regressing $\widehat{\boldsymbol{x}}_i$ on u_i,and regressing $\boldsymbol{z}_{\boldsymbol{x},i}$ on v_i,

$$\widehat{\boldsymbol{x}}_i = u_i \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i,\ \boldsymbol{z}_{\boldsymbol{x},i} = v_i \boldsymbol{\beta}_i + \boldsymbol{\eta}_i, \tag{2}$$

where $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$ are regression coefficients and $\boldsymbol{\epsilon}_i$, $\boldsymbol{\eta}_i$ are the remainders and then setting $\widehat{\boldsymbol{x}}_{i+1} = \boldsymbol{\epsilon}_i$ and $\boldsymbol{z}_{\boldsymbol{x},i+1} = \boldsymbol{\eta}_i$.

By iteratively repeating the optimization and the update n time, we decompose class input embeddings $\widehat{\boldsymbol{x}}$ as $\widehat{\boldsymbol{x}} = \sum_{i=1}^{n} u_i \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{\boldsymbol{n+1}}$ and obtain a set of discovered subgroup vectors $\{\boldsymbol{w}_i\}_1^n$, where n is a hyperparameter.

Identifying Biased Subgroups. Following discovering subgroup directions, the next step is to pinpoint further which of them are biased. We achieve this by computing pseudo-subgroup labels for images in a held-out validation set and evaluating the model's performance on each discovered subgroup. For each validation image, we compute the subgroup score $\boldsymbol{u} := (u_1, ..., u_n)^T \in \mathbb{R}^n$, assigning the subgroup with the highest score as the pseudo-subgroup label. In the soft-label case, we directly use the subgroup score $\boldsymbol{u}$. Subsequently, we compute the model's accuracy on images with each pseudo-subgroup label and identify the k subgroups with the top-k worst accuracy as k biased subgroups.

[1]While the number of subgroups G might vary across classes, for simplicity, we consider an equal number of subgroups in each class.

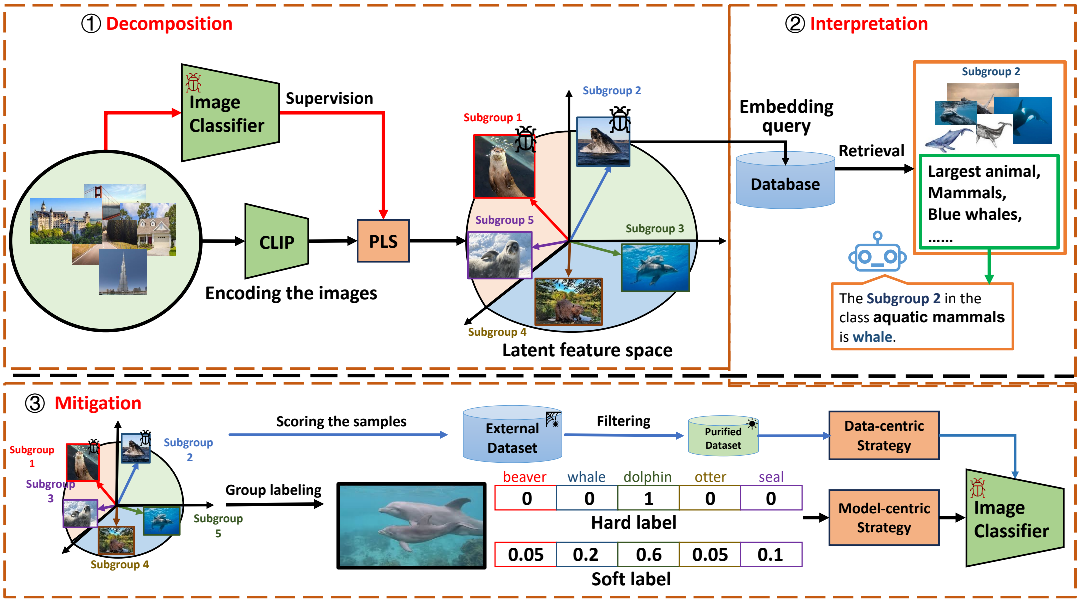

Figure 2. Overview of DIM method. DIM consists of three stages: Decomposition, Interpretation, and Mitigation. At the decomposition stage, we decompose the image features of "aquatic mammals" into the embedding directions of multiple subgroups. Then, we interpret the discovered subgroup embeddings with text descriptions, *e.g.*, the subgroup "whale" in class "aquatic mammals." At the mitigation stage, we propose data-centric and model-centric strategies to mitigate the subgroup bias to improve the robustness of the image classifier.

4.2. Interpretation

Interpreting the discovered subgroups is a critical step to bridge the gap between abstract representations and meaningful insights [6]. The ultimate goal is to generate natural language descriptions of biased subgroups. However, the discovered subgroup embeddings are not directly interpretable. Thus, we leverage the retrieval approach to interpret the discovered subgroup, as shown in the second part of Fig. 2. For each subgroup, the retrieval results are pairs of images and texts (metadata, including descriptions) from the LAION-5B dataset [41], providing visual illustration and contextual information. To further refine the understanding, we collect all retrieved descriptions and utilize a large language model (LLM) for summarization to explain the subgroup.

4.3. Mitigation

In DIM, we propose data-centric and model-centric strategies using the discovered subgroups for bias mitigation.

The **data-centric strategy** is applied to scenarios where access to a substantial external dataset is available. The goal is to enhance the model's performance on biased subgroups by adding a limited number of samples to the training set (due to computational constraints). To achieve this, we leverage the discovered subgroup embeddings to filter extra data to identify high-quality samples that show a strong correlation with the bias subgroups. For each image in the external pool, we first compute its subgroup scores, specifically on those bias subgroups. We then select images with the highest subgroup scores for each biased subgroup. By deliberately increasing the representation of these subgroups in the training set, we aim to systematically mitigate the multiple unknown subgroup biases in the image classifier.

The **model-centric strategy** is proposed to leverage the discovered subgroups to annotate images in the training set with pseudo-subgroup labels and integrate those labels into existing supervised bias mitigation methods. While the hard label, obtained by taking the argmax of subgroup score $\boldsymbol{u} = (u_1, ..., u_n)^T$, can be directly applied to supervised mitigation methods, we also capitalize on the soft label to improve generalization. The motivation is that one image may contain multiple biases, making it belong to multiple subgroups. Thus, some supervised mitigation methods, such as the groupDRO [39] and DI [44], are relaxed to the soft-label version. We provide an example of adapting DI to soft-label (Soft-DI) as follows. More details for the soft-label version of mitigation methods can be found in Appendix B.

Case study of model-centric strategy with Soft-DI. We consider Domain Independent (DI) [44] method with G domains (*i.e.*, number of subgroups), which contains G classification heads sharing features extracted by the backbone. In the original DI, when training on the data $\boldsymbol{x}$ with known hard group label $g \in \{1, 2, ..., G\}$, the model's output is $\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}_g$, where $\hat{p}_g$ is the g-th classifier's output. For the soft-label case, where $\boldsymbol{g} = (g_1, ..., g_G) \in \mathbb{R}^G$, we define the training output as $\hat{\boldsymbol{p}} = \sum_{i=1}^{G} g_i \hat{\boldsymbol{p}}_i$. When performing inference on the test set without group annotation, soft-DI maintains the original method so that $\hat{\boldsymbol{p}} = \frac{1}{G} \sum_{i=1}^{G} \boldsymbol{p}_i$.

5. Experiments

5.1. Setup

Datasets. Our experiment uses three datasets, including CIFAR-100 [17], Breeds [40], and Hard ImageNet [30].

- The CIFAR-100 [17] dataset contains 100 fine-classes, each with 500 images for training and 100 images for testing. These fine-classes are organized into 20 superclasses, with every 5 fine-classes constituting one superclass. To clarify, from now on, we treat superclasses as classes and fine-classes as the ground-truth subgroups.
- Breeds [40], a subset of the ImageNet-1K, is composed of 130 fine-classes. Every 10 fine-classes is grouped into one class, resulting in a total of 13 classes.
- Hard ImageNet [30], another subset of the ImageNet-1K, consist of 15 classes, without further subdivision into finer classes. This dataset is particularly challenging due to strong spurious correlations. The absence of ground truth for discovering multiple unknown subgroups offers a valuable testbed for our method. We apply DIM to mine multiple unknown biases and delineate models' failure.

Baselines. In our experiments, we evaluate three tasks: the discovery of multiple unknown subgroups, the identification of biased subgroups, and the mitigation of pinpointed biases. For discovery and identification tasks, our method, Decomposition, Interpretation, and Mitigation (DIM), is compared with Jain *et al.* [12], Domino [7]. We also test our method by replacing PLS with PCA (DIM-PCA) as an ablation study attesting to the importance of supervision during the decomposition phase. For the mitigation task, we annotate the samples with pseudo labels generated by Jain *et al.* [12] and ours. These labels are then applied to various mitigation methods to verify the effectiveness of discovered subgroups on bias reduction. We implement 1) unsupervised model-centric methods: JTT [26], SubY [11], LfF [32], EIIL [4]; 2) supervised model-centric methods: groupDRO (gDRO) [39], soft-label groupDRO (Soft-gDRO), DI [44], and soft-label DI (Soft-DI), and 3) the data-centric strategy by intervention through filtering extra data [12], with the decision value of Jain *et al.* [12] and ours.

The selection of supervision. Supervision plays a crucial role in guiding the decomposition within the latent feature space in our methodological design. We adopt 3 distinct training dynamics as supervision: 1) correctness: whether the model accurately classifies a given image; 2) logit: the logit output of the model on each image; and 3) loss: the value of loss function for each image. Besides, we also utilize the ground-truth subgroup features as an additional form of supervision for a comprehensive evaluation.

Implementation details. We adopt a consistent approach where models are trained at the class level, and the fine-classes serve as the ground truth for the subgroup discovery task. We train ResNet-18 on CIFAR-100 and ResNet-34 on Breeds from scratch and use ResNet-34 with pre-trained weights for experiments on Hard ImageNet. All models undergo full training until convergence. We leverage CLIP [37] as the foundation model to map images or text to unified representations in the latent feature space. For interpretation, we employ ChatGPT [33] to summarize the concept of grouped description from collected images. More implementation details of our method can be found in Appendix A.

Table 1. The overall cosine similarity score ↑ (averaged over classes) between the subgroups discovered by different methods and the ground-truth subgroups on the CIFAR-100 and Breeds dataset. T.D.: training dynamics; G.T.: ground truths.

Dataset	Jain *et al.* [12]	Domino [7]	DIM (Ours)		DIM-PCA
Supervision	Correctness	Probability	**T.D.**	G.T.	-
CIFAR-100	3.06	12.20	**22.97**	22.03	10.76
Breeds	3.45	10.3	**24.15**	25.26	7.56

Evaluation metrics. For the discovery task, we perform a quantitative assessment to provide a robust and transparent comparison. For each class, we assume to identify G subgroup embeddings[2] $\{\boldsymbol{w}_i\}_1^G$. Then, we compute the representations of the ground-truth subgroups involved in this class via the CLIP text encoder $f_{\text{text}}\colon \mathcal{T} \rightarrow \mathbb{R}^d$ to ensure semantic interpretability. We adopt the text embedding $\boldsymbol{w}_i^t := f_{\text{text}}(\boldsymbol{T}_{\text{prompt}})$ generated from the prompt "a photo of {subgroup}" to represent the ground-truth subgroup in the same latent space. Building upon this, we enumerate all combinations of subgroup embeddings $\{\boldsymbol{w}_i\}_1^G$ and text embeddings $\{\boldsymbol{w}_i^t\}_1^G$, aiming to maximize the overall similarity score. This can be framed as the search for a bijective function $\sigma\colon \{1, 2, \dots, G\} \rightarrow \{1, 2, \dots, G\}$ that maximizes the sum of matched absolute cosine similarities,

$$\max_{\sigma} \sum_{i=1}^{G} |< \boldsymbol{w}_i^g, \boldsymbol{w}_{\sigma(i)}^t >|. \tag{3}$$

For the mitigation task, we evaluate the classification accuracy of multiple underperforming subgroups (averaged across classes). Classification accuracy on the whole dataset is also provided to show the overall performance. More experimental details can be found in Appendix C.

5.2. Evaluation on CIFAR-100

We begin by validating our framework on the CIFAR-100 dataset with known partitions on the classes.

Multiple unknown subgroups discovery. We respectively apply Jain *et al.* [12], Domino [7], our DIM, and DIM-PCA to subgroups discovery for comparative ablation analysis on supervision use. We report the maximum matching similarity between discovered subgroup directions and ground-truth subgroup text embeddings within the latent space in Tab. 1. DIM (ours) achieves superior performance in terms

[2] Though the number of ground-truth subgroups is often unknown, we assume it's available for simplicity.

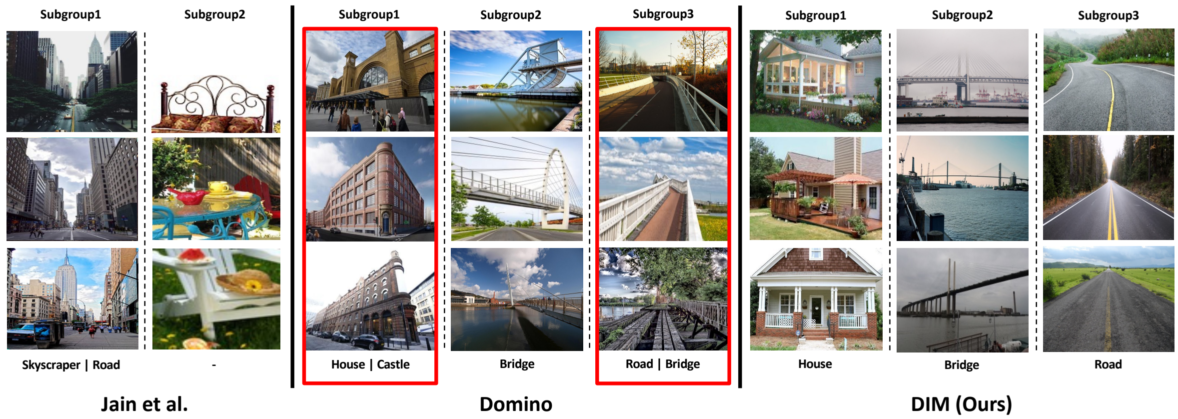

Figure 3. The CLIP-Retrieval results of discovered biased subgroup embeddings in the class "large man-made outdoor things." Images in each column come from the same identified subgroup. Jain *et al.* [12] inherently discovers two subgroups, positive and negative. Although it successfully detected "Skyscraper | Road," it failed to detect the low-performance subgroup. In Domino [7], retrieved images from the first subgroup are a mix of "House" and "Castle." Similarly, images from the third subgroup confuse the "Bridge" and "Road."

Table 2. The success rate of bias subgroup detection on CIFAR-100 and Breeds datasets. T.D.: training dynamics; G.T.: ground truths.

Method	Jain *et al.* [12]	Domino [7]	**DIM (Ours)**		DIM-PCA
Supervision	Correctness	Probability	**T.D.**	G.T.	-
CIFAR-100	45.0	35.0	**57.5**	57.5	52.5
Breeds	46.1	38.4	**61.5**	61.5	53.8

of Eq. (3). Notably, when incorporating training dynamics (T.D.), DIM surpass Jain *et al.* [12] with a significant margin of 10.77. Furthermore, without supervision, DIM-PCA can only achieve a low similarity of 10.76, indicating the crucial role of model supervision in the decomposition.

Biased subgroup detection. We evaluate the success rate of the biased subgroup detection on the CIFAR-100 dataset. Specifically, we scrutinize the correspondence between the detected bias subgroups using different methods and the ground-truth bias subgroups. As shown in Tab. 2, DIM (ours) accurately identified 57.5% under-represented subgroups, outperforming Domino [7], which achieves a 35.0% success rate. It's worth noting that though Jain *et al.* [12] achieved a fairly high success rate of 45.0%, its similarity between the negative direction and the matched low-performance subgroup is critically low, nearing zero. Quantitatively, the negative direction contributes a mere 0.82 out of the 3.06 similarity score achieved by Jain *et al.* [12] in Tab. 1. This indicates that Jain *et al.* [12] aligns with high-performance subgroups through the identified positive direction, yet its negative direction doesn't effectively represent any biased subgroups. This deficiency further leads to the negative direction not being well interpreted in the next stage.

Subgroup interpretation. Both our method and Jain *et al.* [12] capitalize on the latent space of CLIP, allowing us to decode discovered subgroups to images using CLIP-Retrieval naturally. We present the retrieval outcomes for the discovered subgroups within the class "large man-made outdoor things" using Jain *et al.* [12], Domino [7], and our DIM in Fig. 3. As displayed, Jain *et al.* [12] was partial to the well-performance group but failed to illustrate the biased one. Due to space constraints, we selectively exhibit three subgroups for Domino [7] and our DIM. We can see that DIM distinctly and accurately embodied two low-performance subgroups, "bridge" and "road." In contrast, though Domino [7] discovered multiple unknown subgroups, it confused multiple concepts. This finding corroborates our argument that using proxies induces misinterpretation. More results on CIFAR-100 can be found in Appendix D.

Table 3. The classification accuracy of ResNet-18 on the CIFAR-100 test set. We present results on the worst two subgroups to study the mitigation performance of multiple biased subgroups. The overall accuracy is also provided for a comprehensive evaluation.

Type	Method	Worst subgroup accuracy 1st	2nd	Acc.
-	ERM	24.8	33.6	44.4
Model-centric (Unsupervised)	JTT [26]	26.9	34.6	48.5
	SubY [11]	25.1	33.8	45.6
	LfF [31]	25.0	33.8	44.3
	EIIL [4]	25.9	34.8	47.2
Model-centric (labeled by Jain *et al.* [12])	gDRO [39]	26.7	34.5	46.9
	DI [44]	24.3	34.2	47.5
Model-centric (labeled by Domino [7])	gDRO [39]	25.9	34.8	47.2
	DI [44]	25.6	35.3	47.1
Model-centric (labeled by **Ours**)	gDRO [39]	27.2	35.8	48.3
	Soft-gDRO	**27.2**	**38.4**	**49.8**
	DI [44]	26.5	36.7	48.7
	Soft-DI	26.8	37.1	48.9
Data-centric	Jain *et al.* [12]	33.7	41.9	53.1
	DIM (Ours)	**35.6**	**45.1**	**54.7**

Bias mitigation. We apply the model- and data-centric strategies to bias mitigation. For model-centric methods (groupDRO [39] and DI [44]), we use the hard label generated from Jain *et al.* [12] and Domino [7] for supervised training. Our method provides both hard and soft labels for mitigation. For the data-centric strategy, we select the top

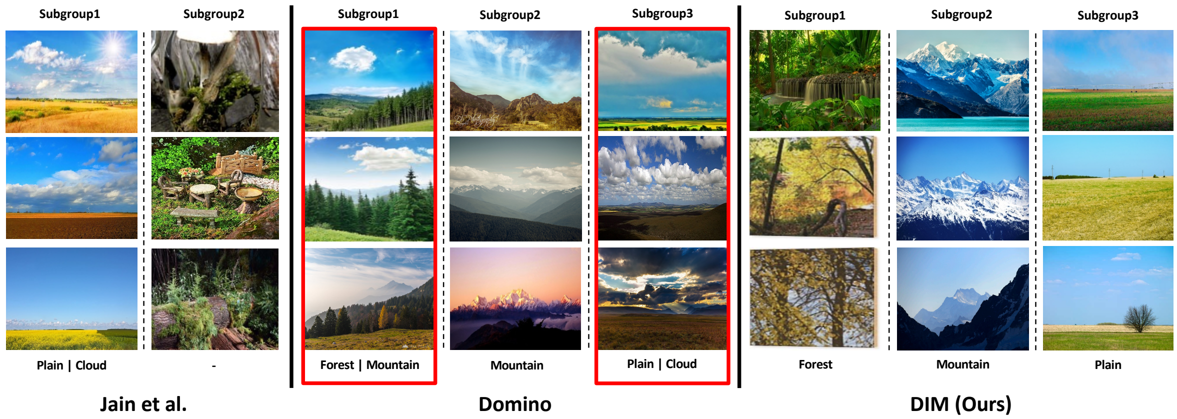

Figure 4. The CLIP-Retrieval results of discovered biased subgroup embeddings in the class "large natural outdoor scenes." Images in each column come from the same identified subgroup. Jain *et al.* [12] inherently discovers two subgroups, positive and negative. Although it successfully detected "Plain | Cloud," it failed to detect the low-performance subgroup. In Domino [7], retrieved images from the first subgroup are a mix of "Forest" and "Mountain." Similarly, images from the third subgroup confuse the "Plain" and "Cloud."

Table 4. The classification accuracy of ResNet-34 on the Breeds test set. We present results on the worst four subgroups to study the mitigation performance of multiple biased subgroups. The overall accuracy is also provided for a comprehensive evaluation.

Type	Method	Worst subgroup accuracy					Acc.
		1st	2nd	3rd	4th	5th	
-	ERM	52.4	58.4	64.4	69.0	71.4	72.7
Model-centric (Unsupervised)	JTT [26]	64.7	70.0	73.2	75.9	79.2	80.0
	SubY [11]	60.8	62.9	67.5	70.3	73.5	74.1
	LfF [31]	63.3	64.1	69.5	72.1	76.9	76.9
	EIIL [4]	62.8	64.5	70.6	73.2	75.3	77.0
Model-centric (labeled by Jain *et al.* [12])	gDRO [39]	63.2	69.6	72.8	74.1	77.9	76.9
	DI [44]	65.9	71.5	76.2	83.6	85.2	84.4
Model-centric (labeled by Domino [7])	gDRO [39]	65.2	71.3	74.2	76.5	78.2	78.3
	DI [44]	66.1	72.3	77.9	84.2	85.3	85.0
Model-centric (labeled by **Ours**)	gDRO [39]	66.8	72.5	75.7	78.2	82.5	80.7
	Soft-gDRO	**69.4**	**75.3**	79.7	83.0	84.1	82.5
	DI [44]	67.3	74.2	79.8	87.6	88.5	86.0
	Soft-DI	68.2	**75.3**	**81.4**	**88.3**	**89.6**	**86.7**
Data-centric	Jain *et al.* [12]	71.6	79.5	82.9	89.3	89.9	89.7
	DIM (Ours)	**72.5**	**82.1**	**84.6**	**91.3**	**91.5**	**89.7**

20% of images with the highest scores on the two bias subgroups identified by Jain *et al.* [12] and our DIM. The results are compiled in Tab. 3. Compared with unsupervised mitigation methods, discovered subgroups in supervised methods boost the mitigation performance with a clear gap of 2.2% on average. Our DIM provides a more precise subgroup discovery for mitigation, achieving a better mitigation performance improvement with up to 1.8% compared with Jain *et al.* [12] and Domino [7]. By filtering multiple under-represented subgroups, DIM has a significant improvement of 3.2% over Jain *et al.* [12], which not only filters a single subgroup but also falls short in precisely representing it. The statistical significance is discussed in Appendix C.2.

5.3. Evaluation on Breeds

Our evaluation extends to Breeds [40] dataset, where we test multiple unknown subgroup discovery and bias mitigation. Breeds, as outlined in Sec. 5.1, mount a more serious challenge in scalability with its fine-grained partition into 130 ground-truth subgroups across 13 classes.

We begin by discovering multiple unknown subgroups. As shown in Tab. 1, compared to baselines, our DIM achieves the highest similarity score with a significant lead of 13.85 for the discovery of unknown subgroups.

Subsequently, we leverage identified subgroups to label data in preparation for different model-centric mitigation methods. We report the classification accuracy of the worst 4 subgroup to evaluate the mitigation performance. The results, as depicted in Tab. 4, suggest that our proposed DIM provides accurate subgroup supervision for mitigation methods, improving the worst subgroup accuracy of 17.3% on average compared with the baseline (ERM). This constitutes an improvement margin of 4.33% versus other subgroup discovery methods, Jain *et al.* [12] and Domino [7]. More results on Breeds can be found in Appendix E.

5.4. Interpreting Hard ImageNet

Unlike CIFAR-100 and Breeds, which have known well-defined class partitions into different subgroups, Hard ImageNet hasn't been extensively analyzed for the existence of multiple biases, presenting unique challenges. Previous studies primarily focus on using counterfactual images to explain the failure modes of models from the single-image level but lack analysis on the subgroup level. We apply our method to discover implicit subgroups within Hard ImageNet.

Following the same setting in [30], we study the ResNet-50 model first pre-trained on the full ImageNet dataset and then fine-tuned on the Hard ImageNet dataset. In our DIM framework, We set the number of subgroups to be discovered in each class to 5 for latent space decomposition. The ablation accuracy is used to evaluate the model's performance on Hard ImageNet, which is the performance drop introduced by masking the target object. A higher ablation accuracy

Figure 5. Example of subgroup interpretation on Hard ImageNet. The first two rows are the retrieval images of identified biased subgroups and corresponding summary descriptions by ChatGPT based on metadata. The last row is from the high-performance subgroup.

Table 5. Ablation study on the n, number of components, in discovering the unknown subgroups. We report different methods' maximum matching similarity ↑(%).

# of comp. (n)	2	3	4	5	6	7	8	9	10
Jain *et al.* [12]	3.1	—	—	—	—	—	—	—	—
Domino [7]	6.1	8.4	10.4	12.2	12.7	12.9	13.3	13.4	13.5
DIM (Ours)	**9.9**	**14.9**	**19.4**	**22.9**	**25.2**	**26.6**	**27.8**	**28.7**	**29.9**

means that the model relies more on spurious features, indicating the existence of implicit bias. We use the ablation accuracy of the validation set to identify biased subgroups.

Using our proposed DIM, we discover 5 subgroups for each class. To explain the failure modes of the studied model, we retrieve images with descriptions (metadata) belonging to the worst 2 subgroups. We then use ChatGPT to summarize the aggregated descriptions into coherent concepts. For comparison, we also analyze images from the best-performing subgroup. We present examples of the class "balance beam" and "dog sled" in Fig. 5. Our findings reveal that the model is vulnerable to spurious correlations, *i.e.*, the "population" and "horizontal bar" in the class "balance beam." In the class "dog sled," our retrieval images and interpretation show that the model may make bad decisions based on necessary but insufficient objects like "snow" and "dogs." These results surface the failure modes, *i.e.*, the presence of multiple spurious features, significantly impairing model's robustness. More results on Hard ImageNet can be found in Appendix F.

5.5. Ablation study

On the number of subgroups to be discovered. On the CIFAR-100 dataset, we vary the number n of iteration, corresponding to the subgroups for discovery, from 2 to 10, match the discovered embeddings to the ground-truth subgroups using the same strategy as Eq. (3), and report the maximum similarity score. As displayed in Tab. 5, while Jain *et al.* [12] fails to identify multiple subgroups and Domino [7] lacks the capacity of fine-grained partitions due to the dataset-level clustering, our DIM demonstrates a better performance when varying n, showing great stability and scalability in the hyperparameter selection by using dimension reduction.

On the use of supervision. In our DIM, the model supervision is managed to guide the latent space decomposition. Without supervision, the decomposition stage may fail to reflect the subgroups learned by the model. To support our argument, we remove supervision at the decomposition stage and use DIM-PCA to discover unknown subgroup directions. The performance degradation in numerical results ($14.4\% \downarrow$ for subgroup discovery in Tab. 1 and $6.25\% \downarrow$ for biased subgroup detection in Tab. 2 on average) sufficiently demonstrates the crucial role of supervision. Supportive retrieval results can be found in Appendix D.2.

6. Conclusion

In this work, we present a novel approach to the problem of multiple unknown biased subgroups in image classifiers. The proposed DIM includes three stages: decomposition, interpretation, and mitigation. We employ model supervision to guide the latent space decomposition and reveal distinct subgroup directions. Then, we identify the biased subgroups and interpret the model failures. The discovered subgroups can be further integrated into the downstream mitigation stage. Our methodology, validated through experiments on CIFAR-100, Breeds, and notably Hard ImageNet, effectively detects subgroups and uncovers new model failure modes related to spurious correlations. The use of model training dynamics as supervision is pivotal in this process, yet selecting optimal supervisory signals for enhanced subgroup representation remains an open area for future research. This study contributes to the field by providing a methodological framework for understanding and improving image classifiers' subgroup robustness.

Acknowledgments. This work was supported by NSF under grant 2202124 and the Center of Excellence in Data Science, an Empire State Development-designated Center of Excellence. The content of the information does not necessarily reflect the position of the Government, and no official endorsement should be inferred.

References

[1] Herve Abdi and Lynne J Williams. Partial least squares methods: partial least squares correlation and partial least square regression. *Computational Toxicology: Volume II*, pages 549–579, 2013. 2, 3

[2] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993. 1

[3] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI Conference on AI, Ethics, and Society*, pages 156–170, 2022. 1

[4] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021. 2, 5, 6, 7, 12

[5] 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. Ieee, 2009. 1

[6] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. *arXiv preprint arXiv:2312.02974*, 2023. 4

[7] Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2021. 5, 6, 7, 8, 12

[8] Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[10] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22910–22921, 2023. 1

[11] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022. 5, 6, 7, 12

[12] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022. 1, 2, 5, 6, 7, 8, 12, 13, 14

[13] Jinyang Jiang, Zeliang Zhang, Chenliang Xu, Zhaofei Yu, and Yijie Peng. One forward is enough for neural network training via likelihood ratio method. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[14] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. 3

[15] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 1

[16] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022. 1

[17] Alex Krizhevsky, Geoffrey Hinton, *et al.* Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 2, 5

[18] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, *et al.* Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. 2

[19] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023. 1

[20] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019. 2

[21] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023. 2, 3

[22] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022. 3

[23] Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14970–14979, 2021. 1, 2

[24] Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. Biasadv: Bias-adversarial augmentation for model debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3832–3841, 2023. 2

[25] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35:24529–24542, 2022. 3

[26] 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. 5, 6, 7, 12

[27] Haoyang Liu, Maheep Chaudhary, and Haohan Wang. Towards trustworthy and aligned machine learning: A data-

centric survey with causality perspectives. *arXiv preprint arXiv:2307.16851*, 2023. 3

[28] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. Zero-shot model diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11631–11640, 2023. 1

[29] Yuzhou Mao, Liu Yu, Yi Yang, Fan Zhou, and Ting Zhong. Debiasing intrinsic bias and application bias jointly via invariant risk minimization (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16280–16281, 2023. 2

[30] Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *Advances in Neural Information Processing Systems*, 35:10068–10077, 2022. 5, 7

[31] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20673–20684. Curran Associates, Inc., 2020. 2, 6, 7, 12

[32] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 3, 5

[33] OpenAI. https://chat.openai.com, 2023. 5

[34] Geon Yeong Park, Sangmin Lee, Sang Wan Lee, and Jong Chul Ye. Training debiased subnetworks with contrastive weight pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7929–7938, 2023. 3

[35] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021. 3

[36] Maan Qraitem, Kate Saenko, and Bryan A Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20320, 2023. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, *et al.* Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5

[38] Julia M Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science*, 1(1):27–42, 2018. 3

[39] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 1, 2, 4, 5, 6, 7, 12

[40] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021. 5, 7

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, *et al.* Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4

[42] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. 2

[43] Agnieszka Słowik and Léon Bottou. On distributionally robust optimization and data rebalancing. In *International Conference on Artificial Intelligence and Statistics*, pages 1283–1297. PMLR, 2022. 2

[44] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 2, 4, 5, 6, 7, 12

[45] Hongyi Wen, Xinyang Yi, Tiansheng Yao, Jiaxi Tang, Lichan Hong, and Ed H Chi. Distributionally-robust recommendations for improving worst-case user experience. In *Proceedings of the ACM Web Conference 2022*, pages 3606–3610, 2022. 2

[46] Hanna Witzgall and Weicheng Shen. Reducing co-occurrence bias to improve classifier explainability and zero-shot detection. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–8. IEEE, 2022. 2

[47] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2

[48] Zeliang Zhang, Wei Yao, Susan Liang, and Chenliang Xu. Random smooth-based certified defense against text adversarial attack. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1251–1265, 2024. 1

[49] Zeliang Zhang, Rongyi Zhu, Wei Yao, Xiaosen Wang, and Chenliang Xu. Bag of tricks to boost adversarial transferability. *arXiv preprint arXiv:2401.08734*, 2024. 1

[50] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022. 1

[51] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 2

[52] Zhiying Zhu, Weixin Liang, and James Zou. Gsclip: A framework for explaining distribution shifts in natural language. *arXiv preprint arXiv:2206.15007*, 2022. 2