Discover and Cure: Concept-aware Mitigation of Spurious Correlation

Shirley Wu [1] **Mert Yuksekgonul** [1] **Linjun Zhang** [2] **James Zou** [1]

Abstract

Deep neural networks often rely on spurious correlations to make predictions, which hinders generalization beyond training environments. For instance, models that associate cats with bed backgrounds can fail to predict the existence of cats in other environments without beds. Mitigating spurious correlations is crucial in building trustworthy models. However, the existing works lack transparency to offer insights into the mitigation process. In this work, we propose an interpretable framework, Discover and Cure (DISC), to tackle the issue. With human-interpretable concepts, DISC iteratively 1) discovers unstable concepts across different environments as spurious attributes, then 2) intervenes on the training data using the discovered concepts to reduce spurious correlation. Across systematic experiments, DISC provides superior generalization ability and interpretability than the existing approaches. Specifically, it outperforms the state-of-the-art methods on an object recognition task and a skin-lesion classification task by 7.5% and 9.6%, respectively. Additionally, we offer theoretical analysis and guarantees to understand the benefits of models trained by DISC. Code and data are available at https://github.com/Wuyxin/DISC.

1. Introduction

Spurious correlations are common in real-world data analysis. Spurious attributes are typically associated with the class label but are non-generalizable (Kaushik et al., 2020; Sagawa et al., 2020). For example, as shown in Figure 1, neural networks that mistakenly associate cats with beds are prone to fail in different settings, *e.g.,* dog-on-bed or cat-on-desk, where the spurious correlation no longer holds. This lack of reliability is a central issue in critical applications, *e.g.,* medical diagnosis (Bissoto et al., 2020).

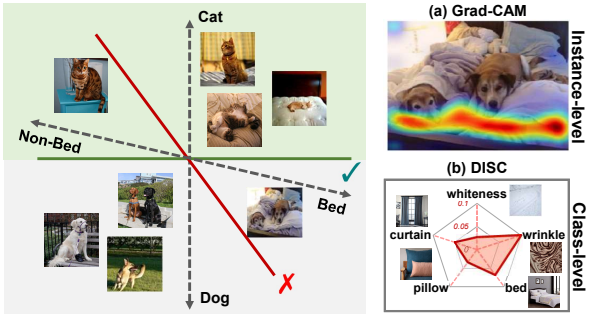


Figure 1. Left: The dog/cat classifiers that rely (red) or do not rely (green) on spurious correlations; Right: Spuriousness discovery results of Grad-CAM (Selvaraju et al., 2017) and DISC, where we propose a class-level metric to indicate the degree of spurious correlation between concepts and the "cat" class.

Existing works have developed methods to mitigate the spurious correlations inside deep models. For instance, invariant learning (Arjovsky et al., 2019; Ahuja et al., 2021) learns a stable representation across environments to avoid varying factors, including spurious attributes. Leveraging the vulnerability of Empirical Risk Minimization (ERM) models towards spurious attributes, some works upweights over-confident (Nam et al., 2020) or misclassified (Liu et al., 2021a) instances from a trained ERM model to counteract the spurious correlation.

However, these works lack interpretability into what are the information the model is learning or ignoring, which hinders human understanding and model auditing. While post-hoc explainability methods (Selvaraju et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017) offer visualization that could contain spurious regions, it is still ambiguous to understand. For example, in Figure 1 (a), the highlighted region shows the attributes that contribute most to the prediction, explaining why the image is mistakenly predicted as "cat" but "dog". Nevertheless, it is not clear which of the attributes (*e.g.,* whiteness, wrinkle texture, or items like pillow) mostly contributes to the spurious correlation. Moreover, such instance-level interpretations are not informative about the overall spurious correlations existing in the class.

In this work, we adopt concepts that align with human understandings to discover class-level spurious attributes,

[1]Department of Computer Science, Stanford University. [2]Department of Statistics, Rutgers University. Correspondence to: Shirley Wu <shirwu@cs.stanford.edu>, James Zou <jamesz@stanford.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

leveraging a concept bank as an auxiliary knowledge base. We show that the invariant concepts, *e.g.,* the shape of cats, remain stable across data environments, while the spurious concepts, *e.g.,* "bed", have inconstant existence across the instances within the class. Inspired by this property, we propose a class-level metric, *concept sensitivity*, to quantify a concept's instability across the data environments. For example, in Figure 1 (b), we identify both "wrinkle" and "bed" as highly spurious concepts of "cat" based on the large magnitude of concept sensitivity, and we refer to this step as **the discovery step**.

Upon discovering the spurious concepts, we propose an intervention step, namely *concept-aware intervention*, to reduce the models' reliance on spurious concepts. The high-level idea is to intervene on the selected classes with spurious concepts to maintain a balanced distribution of the spurious concepts. For instance, after identifying that "wrinkle" and "bed" are spurious concepts correlated with the "cat" class, we use concept images of them to intervene in the "dog" class, as shown on the right of Figure 2. With a balanced distribution of spurious concepts across different classes, we prevent the model from taking advantage of spurious concepts to make predictions, thus canceling the spurious bias. We refer to this process as **the cure step**.

Discover and Cure. Finally, our algorithm, DISC, iteratively conducts the discovery and cure steps during training. In each iteration, it discovers the spurious concepts for the current model. Then based on the discovered concepts, it intervenes on the training datasets to remove the spurious correlations, on top of which the model is updated. Here we focus on image classification tasks. Empirically, DISC discovers spurious concepts that align with ground truth spurious attributes and outperforms the state-of-the-art baselines averagely with a large margin. Our **contributions** are:

- We develop a novel and interpretable framework to discover spurious concepts and effectively mitigate spurious correlations for model generalization.
- We empirically validate our method's effectiveness on diverse datasets and reveal insights into how models overcome spurious correlations.
- We theoretically guarantee the convergence and generalization ability of the models trained by DISC.

2. Related Work

Our work, built on human-interpretable concepts, involves discovering and curing spurious correlations. Here we discuss three classes of related works:

Concepts. Concepts, *e.g., blueness* or *stripes*, are human-interpretable semantics. Concepts have been used to build interpretable models (Lampert et al., 2009; Kumar et al., 2009; Koh et al., 2020; Yüksekgönül et al., 2022), or used in a post-hoc manner (Bau et al., 2017; Kim et al., 2018) to interpret the predictions of deep neural networks (DNNs). Specifically, Kim et al. (2018) introduce Concept Activation Vectors (CAVs), where a CAV represents the direction in the hidden space of a DNN that corresponds to the existence of a concept, helping align the internal state of DNNs with human expectations.

Discovering Spurious Correlations. Previous works study spurious correlations in settings like image texture and backgrounds (Geirhos et al., 2019; Sagawa et al., 2020), domain shifts (Koh et al., 2021; Gulrajani & Lopez-Paz, 2021; Santurkar et al., 2021; Ye et al., 2023), and causally unstable attributes (Arjovsky et al., 2019; Wu et al., 2022). Detecting spurious correlations reveals model biases that are harmful to generalization. Some works obtain spurious attributes using domain knowledge (Clark et al., 2019; Kaushik et al., 2020; Nauta et al., 2021), however, spurious attributes could go beyond domain knowledge. For instance, Creager et al. (2021) infer spurious attributes by learning environments. Sohoni et al. (2020); Seo et al. (2022) cluster a model's embeddings and use the clusters to reveal spurious attributes.

Recent works (Plumb et al., 2021; Hagos et al., 2022; Abid et al., 2022) also use explainability techniques to find spurious attributes and require human inspection. Unlike instance-level auditing, we propose a class-level metric which offers high-level interpretability that is more reliable and user-friendly. Moreover, concept-level and interactive debugging methods (Bontempelli et al., 2022; Bahadori & Heckerman, 2021; Teso & Kersting, 2019) leverage concepts or human feedback to perform debugging. See Teso et al. (2023) for an overview. For example, Bontempelli et al. (2022) propose ProtoPDebug that allows a human supervisor to provide feedback to part-prototype networks (Chen et al., 2019) (ProtoPNets) on the model's explanations. In contrast to our method, they generally work with a restricted class of models (*e.g.,* CBMs (Koh et al., 2020) or ProtoPNets) and often require human annotation to identify the concepts. See Table 7 for the comparison between the selected works and our method.

Curing Spurious Bias. Learning spurious attributes makes models over-sensitive to spurious factors and their distribution shifts, which is related to invariant and robust learning.

- **Invariant Learning**. Arjovsky et al. (2019) propose learning an invariant encoder such that the downstream classifiers are optimal in different environments. Other works target invariance via correlation alignment (Sun & Saenko, 2016), variance penalty (Krueger et al., 2021; Teney et al., 2020), and gradient alignment (Shi et al., 2021) across domains. However, these are not interpretable, which provides little insight into the data bias.
- **Instance Reweighting**. Instance reweighting puts high importance on examples that unlikely contain spurious at-

tributes to remove bias (Yaghoobzadeh et al., 2021; Utama et al., 2020; Dagaev et al., 2021; Zhang et al., 2022b; Nam et al., 2020; Li et al., 2022). Despite its simplicity, such instances could be rare when models perfectly fit the training data, which limits the effectiveness. Distributionally Robust Optimizaton (DRO) (Ben-Tal et al., 2013; Oren et al., 2019; Sagawa et al., 2020; Zhang et al., 2021a) is a special case that puts more weights on observations with high loss (Namkoong & Duchi, 2016; Hu et al., 2018; Levy et al., 2020). Yet, the impact of instance reweighting on over-parameterized DNNs could diminish over epochs (Byrd & Lipton, 2019), leading to overfitting eventually.

- **Data Augmentation**. Other works use data augmentations like adversarial mixup (Xu et al., 2020), selective augmentation (Yao et al., 2022), and uncertainty-aware mixup (Han et al.) to reduce the reliance on the spurious correlation (Zhang et al., 2021b). Moreover, Pinto et al. (2022) propose that mixup (Zhang et al., 2018) as a regularizer can further improves out-of-distribution robustness. However, these augmentations do not explicitly consider multiple and coexistent spurious attributes, which is common in real-world applications. With a concept bank generated from a text-to-image generator, DISC detects the spurious concepts and adaptively mixes up concept images with instances in selected classes. Concurrent work (Jain et al., 2022) uses a captioning model to capture the failure mode and generate synthetic images for fine-tuning. Nevertheless, the generated data relies on the captioning model, which can be out-of-distribution and infeasible for hard-to-describe datasets like skin lesion images. DISC is a more flexible solution using concept images to do the intervention.

3. Method

Here, we describe the problem setup and our method. We formalize our problem setup in Section 3.1 and introduce the construction of the concept bank in Section 3.2. Then, we discuss the discovery of spurious concepts in Section 3.3 and the removal of spurious correlations in Section 3.4. For clarity, we summarize the main notation in Appendix A.

3.1. Problem Formulation

We consider a supervised image classification problem. Specifically, we are given a training dataset $\mathcal{D}_{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. We define $\mathcal{Y}$ as the label space and $\mathcal{P}_{tr}$ as the distribution of the training dataset.

For an arbitrary loss function ℓ, Empirical Risk Minimization (ERM) minimizes the empirical loss for a model f:

$$\arg\min_{\phi} \mathbb{E}_{(x,y)\sim\mathcal{D}_{tr}} \left[\ell\left(f_\phi(x), y\right)\right] \tag{1}$$

where f is parameterized by ϕ. Due to the unstable nature of spurious attributes, the test distribution $\mathcal{P}_{te}$ is often different from the training distribution, *i.e.,* $\mathcal{P}_{te} \neq \mathcal{P}_{tr}$. Thus, the model trained with the ERM falls short of generalizing to datasets $\mathcal{D}_{te} \sim \mathcal{P}_{te}$ where the spurious correlations shift or no longer hold. Thus, our goal is to overcome the model's bias in the presence of spurious correlations. From a causality perspective, spurious attributes are defined as the attributes F that are not causally related to the truth label Y, but are correlated with the truth label Y in the training data due to data sampling bias or imbalance. For example, "bed" can not determine the image being labeled as "cat", but may be correlated to the label "cat" if the cat images are mostly taken in bedrooms. For our purpose, we consider spurious attributes to be attributes whose presence is correlated with the label in some environments but not others.

3.2. Concept Bank

To describe the spurious attributes, we consider them as concepts in a human-understandable fashion instead of pixel-level patterns. We build a comprehensive concept bank that widely covers potential spurious concept candidates. Formally, we have

Definition 1 (**Concept bank**). *A concept bank with m concepts can be expressed as $\mathcal{C} := \{(c_i, \mathcal{P}_{c_i}) \mid i = 1, \ldots, m\}$, where each c_i denotes a concept, $\mathcal{P}_{c_i}$ is the distribution of the images with the concept.*

We show examples in Figure 7 (Appendix D), where we utilize text-to-image generative models, *e.g.,* Stable Diffusion (Rombach et al., 2022) to generate concept images that represent $\mathcal{P}_{c_i}$, using the concept names as prompts.

Moreover, the demand for interpretability calls into aligning concepts with the inner state of deep models. Without loss of generality, we denote a deep model as $f = h \circ g$, where g is an encoder, and $h : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ is a linear layer. Thus, given the model f and a concept c_i, we define a query operation to extract the high-dimensional concept representation v_i. Concretely, we construct the positive set I_i^p by sampling N^p images from $\mathcal{P}_{c_i}$, and the negative set I_i^n by sampling N^n images randomly from $\{\mathcal{P}_{c_j} \mid j \neq i\}$. Following Kim et al. (2018), we learn a linear SVM that finds a hyperplane in the hidden space $\mathbb{R}^d$ to best separate the positive images from the negative ones. Then, we compute the vector v_i orthogonal to the hyperplane as the Concept Activation Vector (CAV). Intuitively, a CAV is the direction in the hidden space representing the existence of a concept, as shown in Figure 2 (middle). Thus, the concept bank serves as an auxiliary knowledge base to discover and mitigate the spurious correlation in the subsequent steps.

3.3. On Discovering Spurious Concepts

With the concept bank, we aim to identify the spurious concepts from the concept candidates.

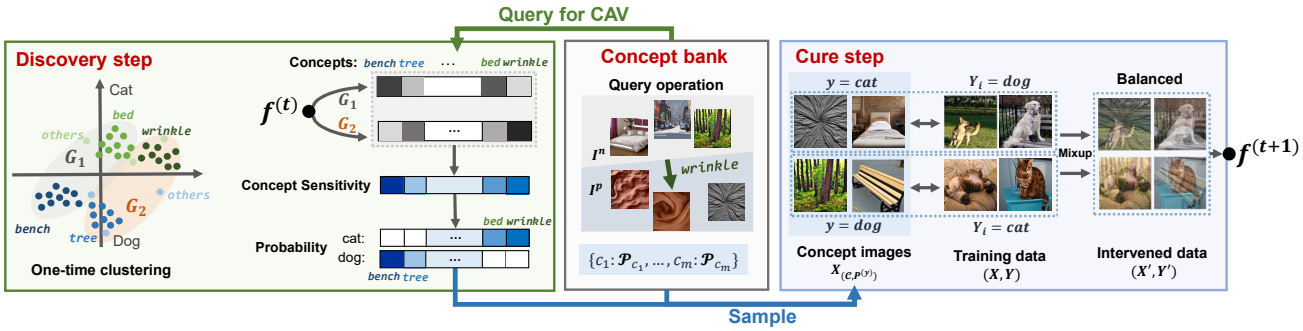


Figure 2. **DISC Framework on MetaShift dataset.** DISC starts with clustering to construct data environments $[G_1, G_2]$ that separates spurious attributes (we use 2 clusters per class to demonstrate). At t-th iteration, DISC computes the concept sensitivity and discovers *(wrinkle, bed)* and *(bench, tree)* as the spurious concepts of "cat" and "dog", respectively. In the cure step, it mixes up the selected subset, *e.g.,* dog images, with images of spurious concepts, *e.g.,* "bed", to remove the spurious correlation with an augmented dataset.

Assumption on Training Distribution. To guarantee the distinguishability of spurious concepts, we assume the training distribution is representative of the overall distribution. For example, if "bed" always coexists with "cat" in the training dataset, then there is no way we can distinguish that "cat" is the invariant concept while "bed" is not. Thus, the training distribution should reflect the essential characteristics as the overall distribution, which is implicitly assumed by the previous works (Liu et al., 2021a; Nam et al., 2020). We formalize this assumption in Theorem 1. With the representativeness assumption, we further propose the following observation about an inconstancy property of spurious concepts:

Observation 1 (Inconsistancy Property). *Spurious concepts tend to be present in heterogeneous subsets of the data and their correlations with the label are also heterogeneous.*

For example, "bed" is not a common characteristic possessed by all the "cat" instances, it might be present in specific subsets of cat images and how much the presence of bed correlates with the cat label also differs across different subsets. We want to leverage this property and insight to identify spurious concepts. Specifically, as a biased model correlates the label with spurious concepts, the distribution of spurious concepts has a large impact on the model's decision boundary. More importantly, this impact is often inconsistent across different environments. For example, imagine we cluster the "cat" instances based on whether it's an indoor or outdoor photo. The models trained under these two scenarios will exhibit distinct preferences for using the "bed" concept to make predictions. Thus, the role of spurious concepts is highly fragile and sensitive in different environments, where the distribution of spurious concepts can change dramatically. In contrast, *invariant concepts*, *e.g.,* animal shape, are more uniform and homogeneous conditioned on any data environments.

Guided by this property, we propose a class-level metric, **Concept Sensitivity**, to indicate if a concept is spurious to a specific class. Here we introduce its computation steps:

Step 1 (Clustering). We first seek good stratification on the training dataset to construct data environments. As explored by the previous works (Sohoni et al., 2020; Seo et al., 2022), even a biased model can well distinguish different features. Thus, we leverage a model trained with ERM to generate both representations and predictions on the training instances, which, similar to Eyuboglu et al. (2022), takes error type into account by including model predictions. Based the generated vectors, we conduct one-time clustering of the instances within each class. For each class $y \in \mathcal{Y}$, we obtain $G^{(y)} = \{G_j^{(y)}\}_{j=1}^k$, where $G_j^{(y)}$ is the j-th cluster in class y and k is the number of clusters. Finally, we construct the environments as

$$G_j = \bigcup_{y=1}^{|\mathcal{Y}|} G_j^{(y)}, \quad j = 1, \ldots, k, \tag{2}$$

where the combination of clusters from each class is to avoid missing labels in the individual environments. Note that the combinations could be different according to the order of cluster indices, thus we randomize the cluster indices to update the environments for more robust training.

Step 2 (Compute concept sensitivity). For each G_j, we define the Environment Gradient Matrix (EGM) as

$$M_j = \nabla_\omega [\mathbb{E}_{(x,y)\sim G_j} \ell(f(x), y)] = \frac{\partial [\mathbb{E}_{(x,y)\sim G_j} \ell(f(x), y)]}{\partial \omega} \tag{3}$$

where $\omega \in \mathbb{R}^{|\mathcal{Y}| \times d}$ is the parameter of h. Intuitively, M_j represents the change in the parameter manifold given the observations in G_j solely. While computing Equation 3 using all the training instances is expensive, we sample mini-batch data to approximate M_j. Further, to align the current model with the concept space, we query the concept bank to extract the concept representation v_i of concept c_i. Then, we compute $v_i \cdot M_j^T \in \mathbb{R}^{|\mathcal{Y}|}$, which indicates how concept c_i is preferred by each class under the environment

G_j. For example, if "bed" is strongly correlated with "cat" in the environment G_j, then the updated decision boundary reflected by M_j will align with the CAV of the "bed" concept with the logits output of being "cat". On top of this, we further define the concept sensitivity S_i of concept c_i as

$$S_i = \text{Var}(\{(v_i \cdot M_j^T)_{y_i'} \mid j = 1, \dots, k\}), \quad (4)$$
$$\text{where } y_i' = \arg\max_y (\sum_{j=1}^{k} v_i \cdot M_j^T)_y$$

Here y_i' is the class dominated by or strongly associated with the concept c_i. Var is the variance operater. For convenience, we refer to $(v_i \cdot M_j^T)_{y_i'}$ as Concept Tendency Score (CTS) of concept c_i under the environment G_j. Thus, the concept sensitivity essentially evaluates the inconsistency of CTS in different environments. A large variance of CTS indicates that the concept is unstable and its contribution to the final prediction varies across different environments. As the causal concept would exhibit invariance for environments (Arjovsky et al., 2019; Bühlmann, 2020), a large concept sensitivity can be interpreted as evidence of a concept being spurious and misleading in the model training. In our previous example where "bed" correlates with the "cat" class, the CTS's of "bed" in different environments show a large variance since "bed" has an inconstant existence with "cat" in the training data. Also, note that the concept sensitivity is class-wise, this offers high-level interpretability to understand the spurious correlations in a certain class.

3.4. Curing Bias via Concept-aware Intervention

Concretely, we blame the model bias for the imbalanced distribution of spurious concepts among classes. For example, in Waterbirds (Sagawa et al., 2020), 95% of instances in the *landbird* class have *land* backgrounds while only 4% of instances in the *waterbird* class involve with *land*. Thus, such extreme imbalance encourages models to take advantage of spurious correlations as shortcuts to make predictions.

However, simply removing spurious attributes from the training dataset could introduce more noise and make the model overfit (Khani & Liang, 2021). Instead, we maintain the distribution balance of spurious concepts in different classes by data augmentation, to cancel the correlation.

Step 3 (Concept-aware Intervention). We denote each $H^{(y)} \in \mathbb{R}^m$ as a boolean vector where $H_i^{(y)} = 1$ if $y_i' = y$, and $H_i^{(y)} = 0$ otherwise. Then we compute the concept probability on each class by normalizing the masked concept sensitivity, *i.e.,* $P^{(y)} = S \cdot H^{(y)} / \sum[S \cdot H^{(y)}]$. Intuitively, the concept probability answers both "what are the concepts correlated with the class y" and "how strong are their spurious correlations". To maintain the balance of spurious concepts, we sample concept images with probability $P^{(y)}$, and mix up them with the instances in their non-dominant classes. Formally, we have

$$X' = \lambda X + (1 - \lambda) X_{(\mathcal{C}, P^{(y)})}, \quad Y' = Y, \quad (5)$$

where $\lambda \sim \text{Beta}(\beta_1, \beta_2)$. $X_{(\mathcal{C}, P^{(y)})}$ denotes concept images sampled with probability $P^{(y)}$ from concept bank $\mathcal{C}$. (X, Y) are drawn from the subset where each $Y_i \in \mathcal{Y}/\{y\}$. Intuitively, more sensitive concepts indicate a larger degree of imbalance between the dominant class and the other classes. And we devise such leave-one-out augmentation to compensate the imbalance, where concept images with high sensitivity are more frequently sampled to be mixed up, achieving the concept distribution balance. In this way, the intervened dataset removes the spurious correlations involving multiple spurious concepts from the training dataset.

Algorithm 1 Pseudocode of DISC

Require: Training data $\mathcal{D}$, concept bank $\mathcal{C}$, a model $f = h \circ g$, learning rate α, parameters β_1, β_2 of Beta distribution

Obtain $\{\{G_j^{(y)}\}_{j=1}^k\}_{y=1}^{|\mathcal{Y}|}$ by clustering the image embeddings
while not converge **do**
  Randomize cluster indices and obtain $\{G_j\}_{j=1}^k$ (Eq. 2)
  $\{P^{(y)}\}_{y=1}^{|\mathcal{Y}|} \leftarrow$ CONCEPT_SENSITIVITY$(G, f, \mathcal{C})$
  for each class y **do**
    Sample minibatch (X, Y), where each $Y_i \neq y$
    Sample concept images $X_{(\mathcal{C}, P^{(y)})}$ with prob. $P^{(y)}$
    Conduct mixup to obtain (X', Y') (Eq. 5)
    $\phi \leftarrow \phi - \alpha \cdot \partial \ell[\mathbb{E}(f(X'), Y')] / \partial \phi$

function CONCEPT_SENSITIVITY$(G, f, \mathcal{C})$
Query for the CAV matrix $V = [v_1^T, \dots, v_m^T]$
for $j = 1, \dots, k$ **do**
  Sample minibatch $(X, Y) \sim G_j$
  Compute the EGM M_j (Eq. 3)
for each concept c_i **do**
  Compute the dominant class y_i' and sensitivity S_i (Eq. 4)
  $H_i^{(y)} \leftarrow \mathbb{I}(y_i' = y), \quad y = 1, \dots, |\mathcal{Y}|$
return $\{P^{(y)}\}_{y=1}^{|\mathcal{Y}|}$

Adaptive Mitigation. However, the model can learn various spurious correlations at different stages of training. It is necessary for the concept sensitivity to be adjusted accordingly to thoroughly correct the model's decision boundary. Therefore, as shown in Figure 2, we propose an adaptive framework DISC, which iteratively conducts spurious concept discovery (Step 2) and concept-aware intervention (Step 3). At each epoch, DISC computes concept sensitivity which guides the concept-aware intervention. Then, the model evolves on the newly intervened dataset, where the spurious correlations are canceled. Thus, DISC can gradually mitigate the spurious correlations learnt by the model in the previous training. In Theorem 1 of Section 5, we provide guarantees for the convergence of concept sensitivity and the final model. By iteratively mixing up images as in Equation 5, DISC reduces the contribution of spurious concepts to the final model and improves model generalization.

Table 1. Overall experimental results. The best results are **bold** and the second best results are underlined.

	MetaShift		Waterbirds		FMoW		ISIC
	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.	Avg. AUROC
ERM	72.9 ± 1.4%	62.1 ± 4.8%	97.0 ± 0.2%	63.7 ± 1.9%	53.0 ± 0.6%	32.3 ± 1.3%	36.4 ± 0.7%
ERM+aug	75.5 ± 1.7%	65.7 ± 3.3%	87.4 ± 0.5%	76.4 ± 2.0%	55.5 ± 0.4%	35.7 ± 0.3%	38.9 ± 1.5%
UW	72.1 ± 0.9%	60.5 ± 3.8%	96.3 ± 0.3%	76.2 ± 1.4%	52.5 ± 0.5%	30.7 ± 1.5%	39.2 ± 0.6%
IRM	73.9 ± 0.8%	64.7 ± 2.1%	87.5 ± 0.7%	75.6 ± 3.1%	50.8 ± 0.1%	30.0 ± 1.4%	45.5 ± 3.6%
IB-IRM	74.8 ± 0.2%	65.6 ± 1.1%	88.5 ± 0.9%	76.5 ± 1.2%	49.5 ± 0.5%	28.4 ± 0.9%	38.6 ± 1.5%
V-REx	72.7 ± 1.7%	60.8 ± 5.5%	88.0 ± 1.4%	73.6 ± 0.2%	48.0 ± 0.6%	27.2 ± 0.8%	24.5 ± 6.4%
CORAL	73.6 ± 0.4%	62.8 ± 2.7%	90.3 ± 1.1%	79.8 ± 1.8%	50.5 ± 0.4%	31.7 ± 1.2%	37.9 ± 0.7%
Fish	64.4 ± 2.0%	53.2 ± 4.5%	85.6 ± 0.4%	64.0 ± 0.3%	51.8 ± 0.3%	34.6 ± 0.2%	42.0 ± 0.8%
GroupDRO	73.6 ± 2.1%	66.0 ± 3.8%	91.8 ± 0.3%	**90.6 ± 1.1%**	52.1 ± 0.5%	30.8 ± 0.8%	36.4 ± 0.9%
JTT	74.4 ± 0.6%	64.6 ± 2.3%	93.3 ± 0.3%	86.7 ± 1.5%	52.5 ± 0.3%	33.4 ± 0.9%	33.8 ± 0.0%
DM-ADA	74.0 ± 0.8%	65.7 ± 1.4%	76.4 ± 0.3%	53.0 ± 1.3%	51.6 ± 0.2%	34.2 ± 0.8%	35.8 ± 1.0%
LISA	70.0 ± 0.7%	59.8 ± 2.3%	91.8 ± 0.3%	88.5 ± 0.8%	52.8 ± 0.9%	35.5 ± 0.7%	38.0 ± 1.3%
DISC	75.5 ± 1.1%	**73.5 ± 1.4%**	93.8 ± 0.7%	88.7 ± 0.4%	53.9 ± 0.4%	**36.1 ± 1.8%**	**55.1 ± 2.3%**

4. Experiments

We conduct comprehensive experiments to answer the following research questions:

- **RQ1**: How effective is DISC on tasks with spurious correlations, compared to state-of-the-art baselines?
- **RQ2**: What are the training dynamics and insights of DISC that are beneficial for future works?
- **RQ3**: How does each component affect DISC 's performance and contribute to its improvements?

4.1. Settings

Datasets. We summarize the datasets in Appendix C. We consider image classification tasks with various types of spurious correlations. Specifically, Waterbirds (Sagawa et al., 2020) associates each class with water or land backgrounds, and MetaShift (Liang & Zou, 2022) constructs disjoint spurious attributes for each class. We also use FMoW from Wilds Benchmark (Koh et al., 2021) where satellite images are collected from different geographical regions that contribute to potential spurious correlations. Moreover, we consider a challenging task, ISIC (Codella et al., 2019), which classifies dermoscopic images of skin lesions into benign or melanoma. We use the train-test splits in Bissoto et al. (2020), where each training split amplifies the correlations with 7 spurious attributes. This task is difficult because multiple spurious attributes, *e.g.,* hairs and gel bubbles, could co-exist and cover the skin lesion region.

Baselines. We compare DISC with Empirical Risk Minimization (ERM) with and without data augmentations; Upweighting (UW) which upweights the instances of minority groups; Invariant Learning algorithms: IRM (Arjovsky et al., 2019), IB-IRM (Ahuja et al., 2021); Domain generalization/adaptation methods: V-REx (Krueger et al., 2021), CORAL (Sun & Saenko, 2016), and Fish (Shi et al., 2021); Instance reweighting methods: GroupDRO (Sagawa et al., 2020), JTT (Liu et al., 2021a); Data augmentation methods: DM-ADA (Xu et al., 2020), LISA (Yao et al., 2022).

Concept Bank. Inspired by previous works (Cimpoi et al., 2014; Fong & Vedaldi, 2018), we build a concept bank with 224 concepts under 6 categories. We generate 200 images per concept from a pre-trained text-to-image generation model, Stable Diffusion (Rombach et al., 2022). To avoid unrealistic interventions, we select the concept categories for each dataset as shown in Table 2. The details of concept bank construction and concept category selection are described in Appendix D.

Model Training. We summarize the hyperparameters in Appendix E and use Gaussian Mixture Model (GMM) as the clustering algorithm. In Waterbirds, due to the extreme imbalance of majority and minority groups, we upweight the minority group for more stable results. While we do not require group information on the other datasets in training.

Evaluation. For ISIC, since the group size is 2^7 considering combinations of spurious attributes, which results in many small groups, we compute the average AUROC score across the train-test splits, as standard in Bissoto et al. (2020). For other datasets, we evaluate the average and worst-group performance. All the experiments are repeated three times.

4.2. Overall Results (RQ1)

Analysis on the baselines. In Table 1, we summarize the overall performance of DISC and the baselines. We observe that ERM with data augmentations constantly surpasses ERM, showing the effect of simple data augmentations in preventing the model from overfitting. Also, we found that GroupDRO performs well in MetaShift and Waterbirds datasets. Yet, its performances are close to or worse than ERM in FMoW and ISIC datasets, which is in line with the observation in Gulrajani & Lopez-Paz (2021); Koh et al. (2021) that GroupDRO generally fails to improve over ERM in the wild. Moreover, the models trained under invariant learning are suboptimal given the insufficient performance.

Furthermore, we focus on the ISIC dataset. One interesting

observation is that JTT fails – the average AUROC is 3% less than ERM – which also justifies our assumption that the effectiveness of instance reweighting is largely limited when the minority instances are rare. Under this setting, the results of data mixup strategies are also unsatisfactory. Specifically, intra-label mixup proposed by LISA can cause even stronger spurious correlations, as it increases the population of majority groups. In conclusion, the limitations of baselines prevent them from obtaining steady success.

The effectiveness of DISC. Overall, DISC outperforms most of the baselines for improving the worst environment accuracy. In particular, we obtain large performance gains over the best baselines in MetaShift and ISIC datasets by 7.5% and 9.6%, respectively. This evidence suggests that DISC is able to mitigate bias when combinations of spurious attributes exist. In FMoW, DISC achieves the state-of-the-art result on both average accuracy and worst-group accuracy, showing that our method can also perform well in wild image datasets. In Waterbirds, DISC improves over JTT by 2.7% while it underperforms GroupDRO. A potential explanation is that the images in the concept bank do not exactly cover the spurious attributes, which hinders the strength of mitigation. Nevertheless, DISC provides interpretability to understand model bias, which is discussed in Section 4.3.

4.3. Interpretability and Training Dynamics (RQ2)

Besides the excellent task performance, our algorithm provides high transparency thanks to concept sensitivity. Here we validate that concept sensitivity correctly reveals spurious correlation and enables user-oriented understanding of the spurious correlations during training.

Validation on the interpretations of DISC. We investigate whether the concept sensitivity faithfully reflects the spurious correlations in the training data. For each concept, we compute the cumulative concept sensitivity over the training epochs to indicate the degree of overall spurious correlation. The top 3 concepts with the largest cumulative sensitivity are **"dotted"** (0.056), **"stripes"** (0.032), and **"stained"** (0.030). Meanwhile, we borrow Cramér's V (Cramér, 2016) to measure the spuriousness for the 7 ground truth spurious attributes, where **gel bubbles** (0.184), **ruler** (0.411), **ink** (0.215) are among the top spurious attributes. Interestingly, in Figure 3, we found strong alignments between the spurious concepts and the spurious attributes. For example, the rulers and "stripes" have a large feature-level similarity.

The conclusion here is two-fold. First, the interpretations align well with the ground-truth spurious attributes, showing their trustworthiness. We also provide a qualitative comparison of the interpretations of DISC and the existing interpretability methods in Appendix F to further show the high quality of DISC interpretations. Second, we demonstrate our method's applicability when certain concepts are absent

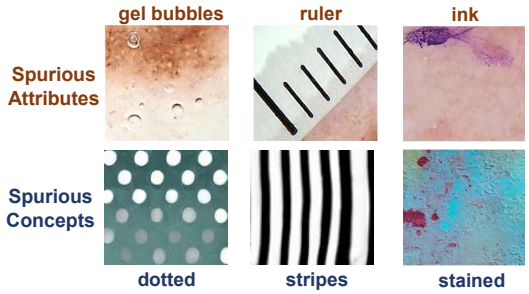


Figure 3. Alignment of spurious attributes and spurious concepts.

from the concept bank, *e.g.,* ruler, which are substituted by other concepts preserving the same essential attributes. Such global and unambiguous interpretation clearly reveals the spurious correlations.

Robustness of interpretations and mitigation under absent ground truth spurious concepts. The previous example shows that DISC finds highly similar concepts even when the ground truth concepts are not included in the concept bank. Thus, we aim to investigate whether such a pattern is consistent. We designed two experiments: (1) Removing "sofa" and "bed" concepts (correlated with "cat") on MetaShift, and (2) removing "bamboo" and "forest" concepts (correlated with "landbird") on Waterbirds. We run the DISC algorithm under concept removal for each case and observe the interpretations on the corresponding class before and after the removal.

The top 3 interpretations before and after removal are **(wrinkle, bed, curtain)** $\rightarrow$ **(fireplace, bedrooms, paisley)** on MetaShift and **(bamboo, forests, flowerpot)**$\rightarrow$**(canopy, ground, plant)** on Waterbirds. Interestingly, we find that the interpretations before and after the removal have some conceptual overlappings (*e.g.,* **"bed"**$\rightarrow$ **"bedrooms"** on MetaShift, **"forests"** $\rightarrow$**"canopy"** and **"flowerpot"**$\rightarrow$ **"plant"** on Waterbirds). We further study the effect of concept removal on the worst group performance. Concretely, the performance decreases by 0.2% on MetaShift and 0.9% on Waterbirds. The absent concepts have a minor effect on MetaShift. While the performance on Waterbirds dataset is more sensitive to the absent concepts, the performance after removal still outperforms most of the baselines.

DISC can discover spurious correlations when the CAV of the absent spurious concept is similar to the CAVs of other concepts. The intuition is that the spurious concept (*e.g.,* forests) and other concepts (*e.g.,* canopy) may share part of the essential attributes (*e.g.,* leaves or greenness) that partially cause the spurious correlation, which results in the similar CAVs in the embedding space. Thus, the interpretations and performance of DISC are robust when ground truth is missing, as supported by our experiments.

Training dynamics as reducing concept sensitivity. The concept sensitivity reflects the extent of a model being affected by spurious bias, which helps probe the current model

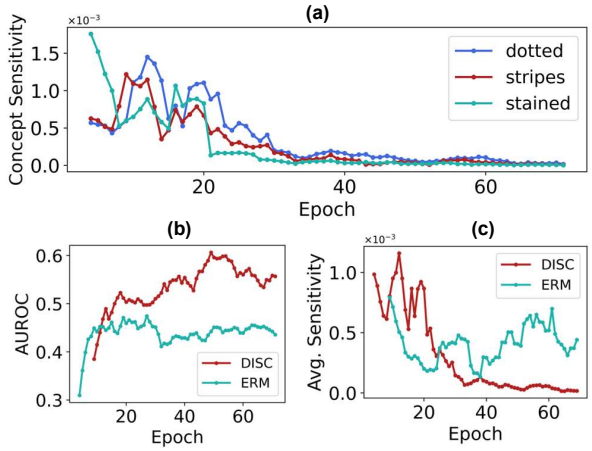


Figure 4. Training dynamics on ISIC. (a) Individual concept sensitivity *vs.* epoch. (b) Test AUROC and (c) The average concept sensitive of DISC and ERM during training.

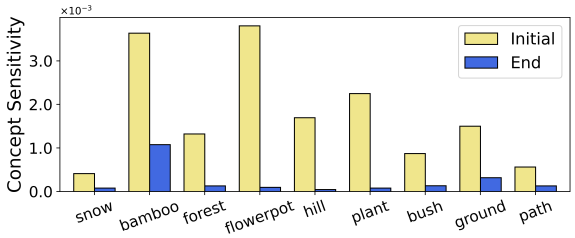


Figure 5. The concept sensitivity of spurious concepts on *landbird* class on Waterbirds at the beginning and the end of training.

state. As shown in Figure 4 (a), we observe the individual concept sensitivity during training. With randomly initialized weights, the model tends to learn from the spurious attributes at the early stage. Correspondingly, the sensitivities of the top 3 concepts are relatively large at the beginning. Fortunately, we maintain the balance of spurious concepts among environments by concept-aware intervention, which gradually decreases the average sensitivity to almost zero.

Moreover, we compare the average sensitivity of the three concepts and task performance for DISC and ERM. In Figure 4 (b) and (c), we found the average concept sensitivity of ERM has increased and remains high at the end. We believe the spurious concepts that are falsely associated with labels and remembered by the model result in the poor performance of ERM. In contrast, DISC reached a low concept sensitivity at the end. This pattern is consistent in the used datasets. For another example, in Figure 5, we show the comparison of the concept sensitivity before and after the training on Waterbirds. The reduction of average sensitivity indicates that the model weight has reached a "sweet spot" where the model is not affected by spurious concepts.

4.4. Ablation and Sensitivity Study (RQ3)

Here we empirically dissect the contribution of (1) concept sensitivity, (2) concept-aware Intervention , and (3) adaptive mitigation in our algorithm. We proposed three ablations

Table 2. Experimental results of the ablation models.

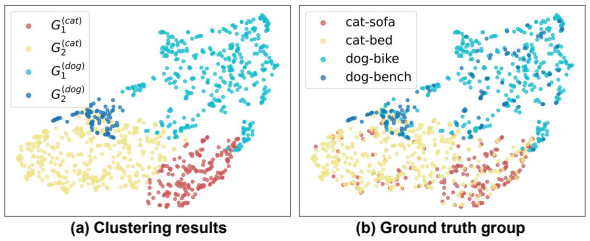
	FMoW		ISIC
	Avg. Acc.	Worst Acc.	Avg. Acc.
DISC-Randint	53.0%	32.1%	49.3%
DISC-Reweight	51.0%	32.0%	35.9%
DISC-Inadaptive	51.9%	31.8%	47.1%
DISC	53.9%	36.1%	55.1%

models respectively:

- **DISC-Randint**, which discards the concept sensitivity and randomly samples concept images for the cure step.
- **DISC-Reweight**, which replaces the cure step with reweighting instances unlikely to contain spurious concepts. Formally, the weight of an instance (x_j, y_j) is $\exp\{-\sum_{i=1}^{m} P_i^{(y_j)} \cdot \max\{0, \frac{g(x_j)^T v_i}{|g(x_j)|\cdot|v_i|}\}\}$, which is negatively proportional to the alignment between its representation $g(x_j)$ and the CAVs of sensitive concepts.
- **DISC-Inadaptive**, which, instead of updating the concept sensitivity every epoch, generates the concept sensitivity based on the pre-trained ERM model and fixes it to conduct the intervention during training.

Ablation results. In Table 2, we report the ablation results on FMoW and ISIC, and include the results of other datasets in Appendix G. By comparing DISC-Randint and DISC, we discover that it is not just intervention, but the *proper intervention* that can effectively reduce spurious correlation. By "*proper*", we mean both "what concept images should be chosen" and "what portion of training data should be intervened by a specific concept", as fulfilled by concept sensitivity. Further, the comparison between DISC-Reweight and DISC implies that the concept-aware intervention promotes the balance of spurious concepts and further mitigates spurious bias, which is a key to DISC's success. DISC-Inadaptive consistently underperforms DISC, and also underperforms DISC-Randint on FMoW and ISIC. Specifically, on these two datasets, we found while using fixed concept sensitivity scores removes the spurious correlations of the concepts with large sensitivity, the severity of the spurious correlations on the other concepts could increase, showing the importance of our adaptive mechanism in removing the spurious correlations more thoroughly. Overall, these three ablation models justify the efficacy of our framework.

Unsupervised clustering. In Step 1, we conduct unsupervised clustering on the training instances to find good stratification on the spurious attributes. Here we are interested in DISC's reliance on the clustering results. To search for the number of clusters k, we adopt Silhouette score as a heuristic following Sohoni et al. (2020). Due to space limit, we include the results in Table 9 (Appendix G). We show that, on most of the datasets, DISC outperforms the best baseline within a wide range of the cluster number. We further visualize the clustering results on MetaShift in



(a) Clustering results (b) Ground truth group

Figure 6. Visualization on the clustering and group assignments.

Figure 6, where the clusters well match the ground truth group assignments. More visualizations are included in Appendix G. Thus, we validate the clustering by a trained ERM is informative to construct data environments.

5. Theoretical Analysis

In this section, we provide theoretical insights to understand the benefits of DISC in removing spurious correlation.

General Assumptions. We consider a Gaussian mixture model as the data-generating mechanism, which has been widely adopted in the machine learning theory to shed light upon understanding complex phenomenon (Montanari et al., 2019; Liu et al., 2021b; Ji et al., 2021; Deng et al., 2021; Zhang et al., 2022a). We consider the setting where the concepts are all well-learned, and build model on the concept level. Specifically, the causal (invariant) concepts are modeled as $x_{inv} = y \cdot \mu + \epsilon_{inv}$, where $y \in \{-1, 1\}$ denotes the class index, $\mu \in \mathbb{R}^{p_1}$ represents the signal of the causal concept with number of dimensions being p_1, and ϵ_{inv} is the noise term that is Gaussian with mean $\mathbf{0}$ and covariance matrix Σ_1. We further assume that the classes are balanced[1], *i.e.*, $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$. The distributions of the causal concepts are assumed to be invariant across training and test environments. In addition, the spurious concepts are modeled as $x_{spu} = \gamma_y^{(i)} + \epsilon_{spu}$, where $\gamma_y^{(i)} \in \mathbb{R}^{p_2}$ controls the spurious correlation and would vary according to different environments $i \in [k]$. As x_{spu} is spurious, we have $\gamma = 0$ in the test distribution $\mathcal{P}_{te}$. Similar to ϵ_{inv}, ϵ_{spu} is the noise term and respects a Gaussian distribution with mean $\mathbf{0}$ and covariance matrix I. As each coordinate of μ and γ represents a concept, in this data generative model, we assume that the values of $\gamma_y^{(i)}$ are either 0 or 1, with 1 indicating the presence of the corresponding concept in the i-th environment of class y.

We consider minimizing the ℓ_2 loss for the classification problem, which has been commonly used in the deep learning theory community (Ma & Belkin, 2017; Shankar et al., 2020; Liang & Recht, 2021). Here we compare the proposed DISC method with the standard ERM method

$$(\hat{\mu}_{ERM}, \hat{\gamma}_{ERM}) = \arg\min_{\mu,\gamma} \sum_{i=1}^{n} (y_i - \mu^\top x_{inv,i} - \gamma^\top x_{spu,i})^2,$$

with the classifer being constructed as $\hat{C}_{ERM}(x) = sgn(\hat{\mu}_{ERM}^\top x_{inv} + \hat{\gamma}_{ERM}^\top x_{spu})$. Similarly, we define the classifier produced by DISC as $\hat{C}_{DISC}(x) = sgn(\hat{\mu}_{DISC}^\top x_{inv} + \hat{\gamma}_{DISC}^\top x_{spu})$, where $\hat{\mu}_{DISC}$ and $\hat{\gamma}_{DISC}$ are the solution produced by Algorithm 1.

Theorem 1. *Assuming that (1). $supp(\gamma_y^{(i)})$'s are disjoint for different y's, and $\mathrm{Var}(\{\gamma_{y,j}^{(i)}\}_{i=1}^k) > K_0$ for $j \in [p_2]$ and some constant $K_0 > 0$, (2). $\|\mu\|_\infty \to 0$ when $p_1 \to \infty$, and $K_1 \le \lambda_{\min}(\Sigma_1) \le \lambda_{\max}(\Sigma_1) \le K_2$ for some constants $K_1, K_2 > 0$, (3). $p_1/n \to 0$ and p_2 is fixed. Then when training size n is sufficiently large, Algorithm 1 converges exponentially fast. Moreover, with probability at least $1 - o(1)$, the solution $(\hat{\mu}_{DISC}, \hat{\gamma}_{DISC})$ satisfies*

$$\mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{DISC}(x) \neq y] < \mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{ERM}(x) \neq y].$$

We clarify the assumptions and include the detailed proof in Appendix B. This theorem implies that by iteratively discovering and intervening, DISC mitigates the variation of the contribution of spurious concepts to the final model. Thus, DISC reduces the spurious correlations in the final model and outperforms ERM.

6. Conclusion and Discussions

We propose DISC as a principled method to discover spurious correlations in a user-friendly way and then mitigate these correlations with data augmentation. DISC is guided by the empirical observation that in many cases, spurious attributes are heterogeneous across different subsets of the data. Our systematic experiments demonstrate that DISC significantly improves model generalizability. Moreover, it provides useful insights into which concepts are sensitive and how this sensitivity changes over training. DISC also leverages the fact that concept images are easy to obtain, especially using generative models like Stable Diffusion.

Limitations. While CAVs connect embedding space with concept space, the learning of the CAVs requires additional computation during training. Another limitation is that the concept bank using a generative model may have its own biases, which may limit the effectiveness of mitigation.

Future Works. Interestingly, we conducted experiments on CIFAR-10-C and found DISC outperforms ERM by 13.1% averaged across four types of corruptions, showing the potential of DISC in **OOD generalization**. Moreover, future works can also build **better concept bank and tools for automatic concept category selection**, as discussed in Appendix D. One can also extend the **applicability of DISC** to multi-object vision datasets and NLP tasks or adopt DISC to analyze the concepts generated by techniques like SENN (Alvarez-Melis & Jaakkola, 2018).

[1]This assumption is considered for the simplicity of presentation, and our analysis can be directly extended to the imbalance case as long as $\min\{\mathbb{P}(Y = 1), \mathbb{P}(Y = -1)\} > c$ for some universal constant $c > 0$.

Acknowlegement

The research of Linjun Zhang is partially supported by NSF DMS-2015378. The research of James Zou is partially supported by funding from NSF CAREER and the Sloan Fellowship. We would like to thank the following people at Stanford University who gave great suggestions in improving our manuscript: Serina Chang, Zhi Huang, Lingjiao Chen, Boyang Deng, Ruocheng Wang, Yang Zheng. We also thank the anonymous reviewers for their insightful comments.

References

Abid, A., Yüksekgönül, M., and Zou, J. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *ICML*, 2022.

Ahmed, F., Bengio, Y., van Seijen, H., and Courville, A. C. Systematic generalisation with group invariant predictions. In *ICLR*, 2021.

Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. 2021.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, 2018.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bahadori, M. T. and Heckerman, D. Debiasing concept-based explanations with causal analysis. In *ICLR*, 2021.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

Ben-Tal, A., den Hertog, D., Waegenaere, A. D., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.*, 2013.

Bissoto, A., Valle, E., and Avila, S. Debiasing skin lesion datasets and models? not so fast. In *CVPR Workshops*, 2020.

Bontempelli, A., Teso, S., Giunchiglia, F., and Passerini, A. Concept-level debugging of part-prototype networks. *ICLR*, 2022.

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Byrd, J. and Lipton, Z. C. What is the effect of importance weighting in deep learning? In *ICML*, 2019.

Cai, T. T. and Zhang, L. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):675–705, 2019.

Cai, T. T. and Zhang, L. A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *The Annals of Statistics*, 49(3):1537–1568, 2021.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5): 2825–2850, 2021.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *CVPR*, 2014.

Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *EMNLP*, 2019.

Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. A., Kittler, H., and Halpern, A. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). 2019.

Cramér, H. *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton university press, 2016.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., and Love, B. C. A too-good-to-be-true prior to reduce shortcut reliance. 2021.

Deng, Z., Zhang, L., Vodrahalli, K., Kawaguchi, K., and Zou, J. Y. Adversarial training helps transfer learning via better representations. *Advances in Neural Information Processing Systems*, 34:25179–25191, 2021.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J., Lee-Messer, C., Dunnmon, J., Zou, J., and Ré, C. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR*, 2022.

Fong, R. and Vedaldi, A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *CVPR*, 2018.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *ICLR*, 2021.

Hagos, M. T., Curran, K. M., and Namee, B. M. Identifying spurious correlations and correcting them with an explanation-based learning. 2022.

Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., and Yao, J. UMIX: improving importance weighting for subpopulation shift via uncertainty-aware mixup.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *ICML*, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jain, S., Lawrence, H., Moitra, A., and Madry, A. Distilling model failures as directions in latent space. 2022.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021.

Kaushik, D., Hovy, E. H., and Lipton, Z. C. Learning the difference that makes A difference with counterfactually-augmented data. In *ICLR*, 2020.

Khani, F. and Liang, P. Removing spurious features can hurt accuracy and affect groups disproportionately. In Elish, M. C., Isaac, W., and Zemel, R. S. (eds.), *FAccT*, 2021.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *ICML*, 2020.

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W. T., Isola, P., Globerson, A., Irani, M., and Mosseri, I. Explaining in style: Training a GAN to explain a classifier in stylespace. In *ICCV*, 2021.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In *NeurIPS*, 2020.

Li, Z. and Xu, C. Discover the unknown biased attribute of an image classifier. In *ICCV*, 2021.

Li, Z., Hoogs, A., and Xu, C. Discover and mitigate unknown biases with debiasing alternate networks. In *ECCV*, 2022.

Liang, T. and Recht, B. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *ICML*, pp. 6781–6792. PMLR, 2021a.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021b.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *NeurIPS*. 2017.

Ma, S. and Belkin, M. Diving into the shallows: a computational perspective on large-scale shallow learning. *Advances in neural information processing systems*, 30, 2017.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Nakada, R., Gulluk, H. I., Deng, Z., Ji, W., Zou, J., and Zhang, L. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4348–4380. PMLR, 2023.

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

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NeurIPS*, 2016.

Nauta, M., Walsh, R., Dubowski, A., and Seifert, C. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 2021.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. In *EMNLP-IJCNLP*, 2019.

Pinto, F., Yang, H., Lim, S., Torr, P. H. S., and Dokania, P. K. Regmixup: Mixup as a regularizer can surprisingly improve accuracy and out distribution robustness. 2022.

Plumb, G., Ribeiro, M. T., and Talwalkar, A. Finding and fixing spurious patterns with explanations. 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pp. 1135–1144, 2016.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Santurkar, S., Tsipras, D., and Madry, A. BREEDS: benchmarks for subpopulation shift. In *ICLR*, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Seo, S., Lee, J., and Han, B. Unsupervised learning of debiased representations with pseudo-attributes. In *CVPR*, 2022.

Shankar, V., Fang, A., Guo, W., Fridovich-Keil, S., Ragan-Kelley, J., Schmidt, L., and Recht, B. Neural kernels without tangents. In *International Conference on Machine Learning*, pp. 8614–8623. PMLR, 2020.

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Singla, S. and Feizi, S. Salient imagenet: How to discover spurious features in deep learning? In *ICLR*, 2022.

Singla, S., Moayeri, M., and Feizi, S. Core risk minimization using salient imagenet. abs/2203.15566, 2022.

Sohoni, N. S., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS 2020*, 2020.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Teney, D., Abbasnejad, E., and van den Hengel, A. Unshuffling data for improved generalization. *arXiv*, 2002.11894, 2020.

Teso, S. and Kersting, K. Explanatory interactive machine learning. In *AIES*, 2019.

Teso, S., Alkan, Ö., Stammer, W., and Daly, E. Leveraging explanations in interactive machine learning: An overview. *Frontiers Artif. Intell.*, 2023.

Utama, P. A., Moosavi, N. S., and Gurevych, I. Towards debiasing NLU models from unknown biases. In *EMNLP*, 2020.

Wu, Y., Wang, X., Zhang, A., He, X., and Chua, T. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022.

Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., and Zhang, W. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6502–6509, 2020.

Yaghoobzadeh, Y., Mehri, S., des Combes, R. T., Hazen, T. J., and Sordoni, A. Increasing robustness to spurious correlations using forgettable examples. In *EACL*, 2021.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *ICML*, 2022.

Ye, H., Zou, J., and Zhang, L. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 8968–8990. PMLR, 2023.

Yüksekgönül, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. *CoRR*, 2022.

Zhang, H., Cissé, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Zhang, J., Menon, A., Veit, A., Bhojanapalli, S., Kumar, S., and Sra, S. Coping with label shift via distributionally robust optimisation. In *ICLR*, 2021a.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? In *ICLR*, 2021b.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. When and how mixup improves calibration. In *International Conference on Machine Learning*, pp. 26135–26160. PMLR, 2022a.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022b.

A. Notation Table

Table 3. Main notations used in the method section. Click here to return to the main paper.

Notation	Meaning
$\mathcal{D}_{tr}/\mathcal{D}_{te}$	Training/Testing dataset
$\mathcal{P}_{tr}/\mathcal{P}_{te}$	Training/Testing distribution
$\mathcal{Y}$	Label space
f	A deep model
g/h	The encoder/last linear layer of f
ϕ/ω	Parameters of f/h
k	Number of clusters per class
m	Number of concepts in the concept bank
d	Number of hidden dimensions
$\mathcal{C}$	A concept bank
c_i	The i-th concept in the concept bank $\mathcal{C}$
v_i	The concept activation vector of concept c_i
y_i'	The dominant class of concept c_i
$\mathcal{P}_{c_i}$	The distribution of the images from the i-th concept
S_i	The concept sensitivity of concept c_i
I_i^p/I_i^n	The positive/negative image set for concept c_i
N^p/N^n	The number of images in positive/negative image set
$G_j^{(y)}$	The j-th cluster in class y
G_j	The j-th environment
M_j	The Environment Gradient Matrix corresponding to G_j
$H^{(y)}$	A boolean mask of concepts for the dominant class y
$P^{(y)}$	Concept sampling distribution for class y
$X_{(\mathcal{C},P^{(y)})}$	Images sampled from concept bank $\mathcal{C}$ with probability $P^{(y)}$

B. Theoretical Analysis

Theorem (Restatement of Theorem 1). *Assuming that (1). $supp(\gamma_y^{(i)})$'s are disjoint for different y's, and $\mathrm{Var}(\{\gamma_{y,j}^{(i)}\}_{i=1}^k) > K_0$ for $j \in [p_2]$ and some constant $K_0 > 0$, (2). $\|\mu\|_\infty \to 0$ when $p_1 \to \infty$, and $K_1 \le \lambda_{\min}(\Sigma_1) \le \lambda_{\max}(\Sigma_1) \le K_2$ for some constants $K_1, K_2 > 0$, (3). $p_1/n \to 0$ and p_2 is fixed. Then when training size n is sufficiently large, Algorithm 1 converges exponentially fast. Moreover, with probability at least $1 - o(1)$, the solution $(\hat{\mu}_{DISC}, \hat{\gamma}_{DISC})$ satisfies*

$$\mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{DISC}(x) \neq y] < \mathbb{P}_{\mathcal{P}_{te}}[\hat{C}_{ERM}(x) \neq y].$$

Clarification on the assumptions. We provide intuitive clarification on each of the assumptions as follows:

- **(1)** The support operation $supp(\cdot)$ is a set consisting of all indices corresponding to nonzero entries in the input vector. The condition of the disjoint supports assumes that the spurious concepts are disjoint in different classes, which is supported by our observations in the experiments, *e.g.,* Figure 6 in Section 4. The condition of the variance assumes that the strength of the spurious correlation (measured by the variance of the contribution of the spurious concepts) is not too small. This condition is necessary to detect spurious concepts by assuming a certain level of distinguishability.
- **(2)** $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and largest eigenvalues of a matrix. The condition on $\|\mu\|_\infty$ assumes that the contribution of causal concepts should be spread out. The assumptions on the upper and lower bounds of eigenvalues of Σ_1 are standard in statistics and machine learning literature (Cai & Zhang, 2019; 2021; Cai et al., 2021; Nakada et al., 2023).
- **(3)** We assume limited number of spurious concepts. Moreover, the number of training data n needs to largely exceed the number of causal concepts so that the model can learn the invariant concepts well for the classification.

Proof. We start the proof by analyzing the two main steps: concept sensitivity computation and the gradient update via mixup in each iteration.

Denoting the mini-batch size as B. We note that we assume the concepts are all well-learned and only analyze the fitting on top of the well-learned concepts. The matrix multiplication by CAV will not show up throughout this proof. Under our model set-up, at iteration t, we have for $j \in [k]$, the M_j in (3) equals to

$$M_{j,y} = X^\top (y\mathbf{1}_B - X(\mu_t; \gamma_t))/B.$$

The corresponding parts for the causal and spurious features are respectively

$$M_{j,y;inv} = X_{inv}^\top (y\mathbf{1}_B - X(\mu_t; \gamma_t))/B = X_{inv}^\top y\mathbf{1}_B/B - X_{inv}^\top X_{inv}\mu_t/B - X_{inv}^\top X_{spu}\gamma_t/B,$$

and

$$M_{j,y;spu} = X_{spu}^\top (y\mathbf{1}_B - X(\mu_t; \gamma_t))/B = X_{spu}^\top y\mathbf{1}_B/B - X_{spu}^\top X_{inv}\mu_t/B - X_{spu}^\top X_{spu}\gamma_t/B.$$

As X are assumed to be sub-gaussian, we have that

$$\begin{aligned} M_{j,y;inv} =& \mathbb{E}[X_{inv}^\top y\mathbf{1}_B/B - X_{inv}^\top X_{inv}\mu_t/B - X_{inv}^\top X_{spu}\gamma_t/B] + O(\sqrt{\frac{p_1}{n}}) \\ =& \mu - (\mu\mu^\top + \Sigma_1)\mu_t - \gamma_t^\top \gamma_y^{(i)} \cdot \mu + O(\sqrt{\frac{p_1}{n}}), \end{aligned}$$

and

$$\begin{aligned} M_{j,y;spu} =& \mathbb{E}[X_{spu}^\top y\mathbf{1}_B/B - X_{spu}^\top X_{inv}\mu_t/B - X_{spu}^\top X_{spu}\gamma_t/B] + O(\sqrt{\frac{p_2}{n}}) \\ =& y \cdot \gamma_y^{(i)} - \mu^\top \mu_t \cdot \gamma_y^{(i)} - (\gamma_y^{(i)}(\gamma_y^{(i)})^\top + I)\gamma_t + O(\sqrt{\frac{p_2}{n}}) \\ =& y \cdot \gamma_y^{(i)} - \mu^\top \mu_t \cdot \gamma_y^{(i)} - (\gamma_y^{(i)})^\top \gamma_t \cdot \gamma_y^{(i)} + I\gamma_t + O(\sqrt{\frac{p_2}{n}}). \end{aligned}$$

Then, as we now consider the binary classification setting, the S_i in (4) now equals to

$$S_{i,j}^{(y)} = \mathrm{Var}(\{\gamma_{y,j}^{(i)}\}_{i=1}^k).$$

Now, for the invariant part, as $\|\mu\|_\infty = O(1)$, and fixed p_2 implying that $|\gamma_t^\top \gamma_y^{(i)}| = O(1)$, we have for all $j \in [p_1]$,

$$S_{i,j}^{(y)} = o(1).$$

Also, by assumption, we have $\|\gamma_y^{(i)}\| > 1$ and $\mathrm{Var}(\{\gamma_{y,j}^{(i)}\}_{i=1}^k) > K_0$ for $j \in [p_2]$, and therefore for all $j \in [p_2]$ and some constant $K_0 > 0$, we have

$$S_{i,j}^{(y)} = \Omega(1).$$

As a result, with probability at least $1 - o(1)$, the sampling according to $P^{(y)}$ will always draw from the spurious concepts from $\cup_{i=1}^k \mathrm{supp}(\gamma_y^{(i)})$. We denote such an event by E with $\mathbb{P}(E) \geq 1 - o(1)$.

Now we analyze the mixup part on the event E. According to our model setup, for $j \in [p_2]$, the concept image is modeled as the basis vector e_j, with the j-th entry equal to 1, indicating the presence of this concept.

Letting $\tilde{\gamma}_y = \frac{1}{k}\sum_{i=1}^k \gamma_y^{(i)}$. Then after mixup, for the spurious concepts, there exists a vector $c_{y,spu}$ with support belonging to $\cup_{i=1}^k \mathrm{supp}(\gamma_y^{(i)})$ and nonzero entries are in $(0, 1)$, such that the gradient update becomes

$$S_{spu}^{(i)} = \sum_{y\in\{-1,1\}} y\cdot(\tilde{\gamma}_y + c_{-y,spu}) - \mu^\top \mu_t \cdot (\sum_{y\in\{-1,1\}} \tilde{\gamma}_y + c_{-y,spu}) - (\sum_{y\in\{-1,1\}} (\tilde{\gamma}_y\tilde{\gamma}_y^\top + c_{-y,spu}c_{-y,spu}^\top) + I)\gamma_t + O(\sqrt{\frac{p_2}{n}}),$$

Note that we assume $y \in \{-1, 1\}$. Using the fact that the supports of $\tilde{\gamma}_y$ and $c_{-y,spu}$ are disjoint, we have that

$$S^t_{spu} = \sum_{y\in\{-1,1\}} y\cdot(\tilde{\gamma}_y+c_{-y,spu})-\mu^\top\mu_t\cdot(\sum_{y\in\{-1,1\}}\tilde{\gamma}_y+c_{-y,spu})-(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y+c_{-y,spu})(\tilde{\gamma}_y+c_{-y,spu})^\top+I)\gamma_t+O(\sqrt{\frac{p_2}{n}}).$$

In addition, the gradient on the invariant (causal) part

$$S^t_{inv} = \mu - (\mu\mu^\top + \Sigma_1)\mu_t - \gamma_t^\top\gamma_y^{(i)}\cdot\mu + O(\sqrt{\frac{p_1}{n}}).$$

As a result, the update in each iteration t of is equivalent to running gradient descent on minimizing the loss function $\ell(\hat{\mu},\hat{\gamma}) = (\hat{\mu};\hat{\gamma})^\top \begin{pmatrix} \Sigma_1+\mu\mu^\top & 0 \\ 0 & \sum_{y\in\{-1,1\}}(\tilde{\gamma}_y+c_{-y,spu})(\tilde{\gamma}_y+c_{-y,spu})^\top + I \end{pmatrix}(\hat{\mu};\hat{\gamma}) + (\hat{\mu};\hat{\gamma})^\top(\mu;\sum_{y\in\{-1,1\}} y\cdot(\tilde{\gamma}_y+c_{-y,spu}))$.

Since $\lambda_{\min}(\Sigma_1), \lambda_{\min}(I) > K_1$, ℓ is a strongly convex function, implying that Algorithm 1 converges exponentially fast.

At last, we compare the performance of DISC and ERM.

Since $(\hat{\mu}_{DISC}, \hat{\gamma}_{DISC})$ minimizes ℓ, we can write out its analytical solution as

$$\hat{\mu}_{DISC} = (\Sigma_1 + \mu\mu^\top)\mu + O(\sqrt{\frac{p_1}{n}}),$$

and

$$\begin{aligned}\hat{\gamma}_{DISC} =&(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y + c_{-y,spu})(\tilde{\gamma}_y + c_{-y,spu})^\top + I)^{-1}\sum_{y\in\{-1,1\}} y\cdot(\tilde{\gamma}_y + c_{-y,spu})\\ =&(\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y)(\tilde{\gamma}_y)^\top + \sum_{y\in\{-1,1\}}(c_{-y,spu})(c_{-y,spu})^\top + I)^{-1}\sum_{y\in\{-1,1\}} y\cdot(\tilde{\gamma}_y + c_{-y,spu})\end{aligned}$$

Similarly, we have

$$\hat{\mu}_{ERM} = (\Sigma_1 + \mu\mu^\top)\mu + O(\sqrt{\frac{p_1}{n}}),$$

and

$$\hat{\gamma}_{ERM} = (\sum_{y\in\{-1,1\}}(\tilde{\gamma}_y)(\tilde{\gamma}_y)^\top + I)^{-1}\sum_{y\in\{-1,1\}} y\cdot\tilde{\gamma}_y.$$

Since all the entries of $\tilde{\gamma}_y$ are either 0 or 1, all the entries of $c_{y,spu}$ are between 0 and 1, and the support of $\tilde{\gamma}_y$ and $c_{-y,spu}$ are disjoint, we have that

$$\|\hat{\gamma}_{DISC}\| < \|\hat{\gamma}_{ERM}\|.$$

Now we analyze the misclassification error in the test domain. For any $\hat{\mu}$ and $\hat{\gamma}$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_{te}}(sgn(\hat{\mu}^\top x_1 + \hat{\gamma}^\top x_2) \neq y) =&\frac{1}{2}\mathbb{P}_{\mathcal{D}_{te}}(\hat{\mu}^\top\mu + \hat{\mu}^\top\epsilon_1 + \hat{\gamma}^\top\epsilon_2 > 0) + \frac{1}{2}\mathbb{P}_{\mathcal{D}_{te}}(-\hat{\mu}^\top\mu + \hat{\mu}^\top\epsilon_1 + \hat{\gamma}^\top\epsilon_2 < 0)\\ =&\frac{1}{2}\mathbb{E}[\mathbb{P}(\hat{\mu}^\top\mu + \hat{\mu}^\top\epsilon_1 + \hat{\gamma}^\top\epsilon_2 > 0 \mid \epsilon_1)] + \frac{1}{2}\mathbb{E}[\mathbb{P}(-\hat{\mu}^\top\mu + \hat{\mu}^\top\epsilon_1 + \hat{\gamma}^\top\epsilon_2 < 0 \mid \epsilon_1)]\\ =&\mathbb{E}[\Phi(-\frac{\hat{\mu}^\top\mu}{\sqrt{\|\hat{\mu}\|^2 + \|\hat{\gamma}\|^2}})],\end{aligned}$$

where Φ is the cumulative distribution function of a standard normal distribution. As a result, using $\hat{\mu}^\top\mu = c\|\mu\|^2 + O(\sqrt{p_1/n}) > 0$ and $\|\hat{\gamma}_{DISC}\| < \|\hat{\gamma}_{ERM}\|$, We have that

$$\mathbb{P}_{\mathcal{D}_{te}}[\hat{C}_{DISC}(x) \neq y] < \mathbb{P}_{\mathcal{D}_{te}}[\hat{C}_{ERM}(x) \neq y].$$

□

C. Datasets

Table 3. (a) **Metashift Dataset**.

Target: classify cat / dog.
Spurious feature: object / background; sofa, bed (cat); bench, bike (dog).

Image:						
Group g:	0	1	2	3	4	5
Target $y \in \{0,1\}$:	0 (cat)	0 (cat)	1 (dog)	1 (dog)	0 (cat)	1 (dog)
Spurious s:	0 (sofa)	1 (bed)	2 (bench)	3 (bike)	4 (shelf)	4 (shelf)
# Train data:	231	380	145	367	-	-
# Val data (OOD):	-	-	-	-	34	47
# Test data:	-	-	-	-	201	259

Table 3. (b) **Waterbirds Dataset.**

Target: bird type; **Spurious feature**: background type.

Image:				
Group g:	0	1	2	3
Target $y \in \{0,1\}$:	0 (landbird)	0 (landbird)	1 (waterbird)	1 (waterbird)
Spurious s:	0 (land)	1 (water)	0 (land)	1 (water)
# Train data:	3,498 (73%)	184 (4%)	56 (1%)	1,057 (22%)
# Val data:	467	466	133	133
# Test data:	2,255	2,255	642	642

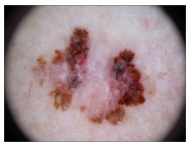
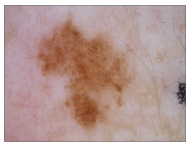
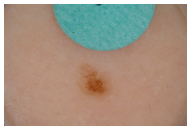
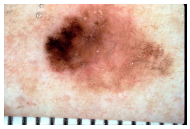
Table 3. (c) **FMoW Dataset.**

Target: one of 62 building or land use categories, e.g., park, shopping mall, dam, stadium, airport.
Spurious features: Unknown (not explicitly given by the data source).

Image:					
Group g:	Europe	Asia	Americas	Africa	Oceania
# Train data:	34,816	17,809	20,973	1,582	1,641
# Val data:	7,732	4,121	6,562	803	693
# Test data:	5,858	4,963	8,024	2,593	666

Table 3. (d) **ISIC Dataset**. For methods that require domain information, we use the existence of hairs as the domain labels. Each training split amplifies different correlations, and the corresponding testing set provides reversed correlations.

Target: benign / melanoma skin lesions
Spurious features: dark corners, hair, gel borders, gel bubbles, ruler, ink markings/staining, patches.

Image:			...		
Target $y \in \{0, 1\}$**:**	0 (benign)	0 (benign)	...	1 (malignant)	1 (malignant)
Spurious s**:**	patch, gel border	ink, hair	...	dark corner, gel bubble	ruler, dark corner
# Train data:			1,826		
# Val data:			154		
# Test data:			618		

D. Concept Bank

Concept categories. In Table 4, we list all the 224 concepts in the concept bank under 6 categories, which are *(Color, Texture, Nature, City, Household, Others)*. Note that the concept bank could be easily extended with user-defined concepts since the concept images are cheap to obtain, leveraging the text-to-image generative models.

Table 4. A comprehensive concept list of the concept bank in this work.

Concept category	Concepts
Color	[*blackness, blueness, greenness, redness, whiteness*]
Texture	[*concrete, granite, leather, laminate, metal, blotchy, blurriness, stripes, polka dots, knitted, cracked, frilly, waffled, scaly, lacelike, grooved, stratified, gauzy, marbled, flecked, stained, braided, matted, meshed, cobwebbed, spiralled, dotted, crosshatched, wrinkled, woven, potholed, crystalline, paisley, veined, fibrous, studded, bubbly, pleated, grid, perforated, porous, interlaced, smeared, honeycombed, sprinkled, chequered, lined, banded, bumpy, zigzagged, swirly, pitted, freckled*]
Nature	[*bamboo, beach, bridge, bush, canopy, earth, field, flower, flowerpot, fluorescent, forest, grass, ground, harbor, hill, lake, mountain, muzzle, palm, path, plant, river, sand, sea, snow, tree, water*]
City	[*awning, base, bench, building, earth, fence, field, ground, house, manhole, path, snow, streets*]
Household	[*air-conditioner, apron, armchair, back-pillow, balcony, bannister, bathrooms, bathtub, bed, bedclothes, bedrooms, cabinet, carpet, ceiling, chair, chandelier, chest-of-drawers, countertop, curtain, cushion, desk, dining-rooms, door, door-frame, double-door, drawer, drinking-glass, exhaust-hood, figurine, fireplace, floor, flower, flowerpot, fluorescent, ground, handle, handle-bar, headboard, headlight, house, jar, lamp, light, microwave, mirror, ottoman, oven, pillow, plate, refrigerator, sofa, stairs, toilet*]
Others	[*bird, cat, cow, dog, horse, mouse, paw, arm, back, body, ear, eye, eyebrow, female-face, leg, male-face, foot, hair, hand, head, inside-arm, knob, mouth, neck, nose, outside-arm, ashcan, airplane, bag, bus, beak, bicycle, blind, board, book, bookcase, bottle, bowl, box, brick, basket, bucket, bumper, can, candlestick, cap, car, cardboard, ceramic, chain-wheel, chimney, clock, coach, coffee-table, column, computer, counter, cup, desk, engine, fabric, fan, faucet, flag, floor, food, foot-board, frame, glass, keyboard, lid, loudspeaker, minibike, motorbike, napkin, pack, painted, painting, pane, paper, pedestal, person, pillar, pipe*]

Concept image generation and examples. All the concept images are synthetic and generated by the Stable Diffusion model with the pretrained weights "stable-diffusion-v1-4", where we use the concept name or its pluralization form as prompts. The code of generating concept bank is made public at this link.As shown in Figure 7, we present the selected concept images in the concept bank as demonstrations.

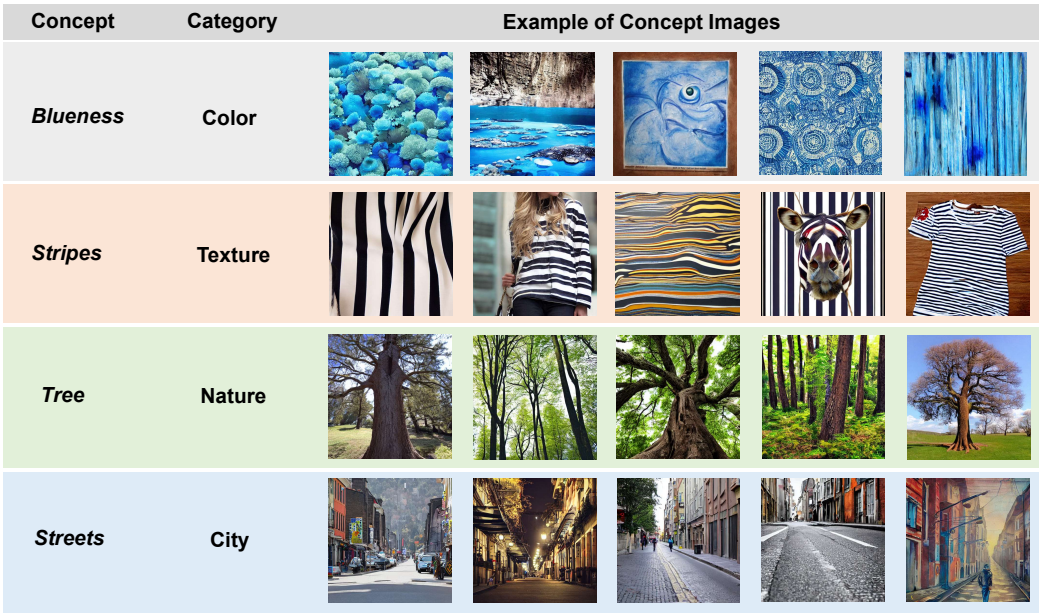


Figure 7. Examples of concept images in the concept bank.

Potential bias of concept images. Generative models may not necessarily be perfect at generating concept images. While it is true that they may have their own biases, they are trained on much larger datasets and thus are less likely to contain more severe spurious correlations for simple concepts. Empirically, we found that synthetic concept images are less noisy or biased compared to real images. For example, we observed that the concept images of "tree" in the BRODEN dataset (Fong & Vedaldi, 2018) of visual concepts highly coexist with "human" (*e.g.,* while hiking), while the synthetic images in our concept bank are much less likely to contain such bias, as shown in Figure 7. Moreover, previous work (Abid et al., 2022) also shows that learning the CAVs does not require a large number of concept images, which allows simple filtering on the concept bank to further guarantee its trustworthiness.

Table 5. Selected concept categories for each dataset

	Concept categories					
	Color	Texture	Nature	City	Household	Others
MetaShift	✔	✔	✔	✔	✔	✔
Waterbirds	✔	✔	✔			
FMoW	✔	✔	✔	✔		
ISIC	✔	✔				

Concept category selection and filtering. As shown in Table 5, we select concept categories for each dataset. The general principle of selection is including the appeared objects in the dataset, based on the prior knowledge of the dataset context. We give the following demonstrations:

- We include Color and Texture for all the datasets since these two concept categories have general existence.
- With the prior knowledge that FMoW is a satellite image dataset, we include Nature and City categories since they may appear in the dataset and thus could contain candidates of spurious concepts.
- With the prior knowledge that ISIC is a skin disease dataset, we exclude Nature, City, *etc.*, that do not exist from the concept candidates for this dataset, and only include Texture and Color concepts.

In real-world applications, such knowledge of dataset contexts is fundamentally required for downstream tasks, which is **generalizable** to the other datasets. Moreover, since the large concept bank is shared across datasets and the practitioner can select the categories instead of the individual concepts, which requires **little labor**.

Moreover, in our implementation, we use a filtering module to filter relevant concepts in a dataset inspired by Abid et al. (2022). The benefits of the concept category selection and filtering are (1) avoiding unrealistic interventions, *e.g.,* mixup animal images with satellite images, and (2) reducing the computational cost of computing CAVs during the training process.

Automatic concept category selection. As a future direction, to further avoid the concept category selection for an unknown downstream task, the protocol to automatically select suitable concept categories can be

- Leveraging image recognition models to identify existing objects in the datasets.
- Then, extracting concepts or concept categories from our dataset-agnostic concept bank, which is defined in Table 4.

Learning CAVs. To learn the CAVs, we use $N^p = N^n = 150$ for all the concepts. Another future direction is that we can learn more accurate concept representations by using hard negative samples. For example, we can construct the negative set for *tree* concept using concepts images that are similar to tree images, *e.g., grass* and *flowers*. For simplicity, we use random sampling to construct the negative sets in this work.

E. Model and Optimization Details

We adopt DenseNet121 (Huang et al., 2017) on FMoW and ResNet-50 (He et al., 2016) on the other datasets. The hyper-parameters are summarized in Table 6. For the Beta distribution, we use $\alpha = \mu = 2$ in all the datasets. Note that we search the number of clusters per class using Silhouette score, which is detailed in Appendix G.

Table 6. Hyper-parameters of DISC during training.

	Leaning Rate	Batch Size	Weight Decay	#Clusters per Class
MetaShift	$5e$-4	16	$1e$-4	2
Waterbirds	$1e$-4	32	$1e$-4	3
FMoW	$1e$-4	10	0.0	3
ISIC	$5e$-4	16	$1e$-5	3

F. Results of Interpretation Comparison

Here we first analyze the advantages of the interpretations generated by DISC over the existing baselines that identify spurious correlation. We study three dimensions of interpretability:

- **Class/group-wise**: Whether the explanations are concerning a class or group, which have the advantage of obtaining common insights across several instances, as opposed to instance-wise explanations.
- **Concept/caption-based**: Whether the explanations are based on captions or concepts that are more human-friendly and unambiguous instead of feature maps.
- **Adaptive**: Whether the explanations are adaptive or intrinsic during the training process, which enables dynamic inspection, as opposed to post-hoc explanations.

We consider different explanation types, including the existing saliency-based and concept-based methods. To highlight, DISC is the only method that fulfills the three advantages.

In Figure 8, we further qualitatively evaluate the interpretations of DISC and three other types of explanations: (1) Grad-CAM (saliency-based method). (2) Failure-Direction (Jain et al., 2022) (caption-based method). (3) CCE (Abid et al., 2022) (concept-based method). For Grad-CAM, similar to the previous observation, the instance-wise saliency maps could be hard to interpret and draw global insights in understanding the predictions for a class. For Failure-Direction, we compute the caption scores following the original paper and obtain the word scores by aggregation. Specifically, we found that the caption model sometimes focuses on the foreground instead of the background, making a subset of the captions uninformative for debugging. Moreover, the interpretations lack diversity due to the limitation of the captioning model. For CCE, we find that the interpretations of DISC and CCE are similar. This aligns with our expectations since both DISC and CCE leverage CAVs to generate interpretations. Moreover, DISC offers more dynamic inspection during model training.

Table 7. Comparison between interpretations of DISC and the existing methods.

	Explanation types	Class/group-wise	Concept/caption-based	Adaptive
Singla & Feizi (2022) Selvaraju et al. (2017) Singla et al. (2022)	Saliency-based	✔(partial)	✔(partial)	✗
Sohoni et al. (2020) Seo et al. (2022)	Clustering-based	✔	✗	✗
Creager et al. (2021) Liu et al. (2021a) Li et al. (2022) Ahmed et al. (2021)	Partition-based	✔	✗	✗
Abid et al. (2022) Bontempelli et al. (2022)	Concept-based	✔	✔	✗
Jain et al. (2022) Eyuboglu et al. (2022)	Caption-based	✗	✔	✗
Lang et al. (2021) Li & Xu (2021)	Generative counterfactuals	✗ ✔	✗ ✗	✗ ✗
DISC	Adaptive concept-based	✔	✔	✔

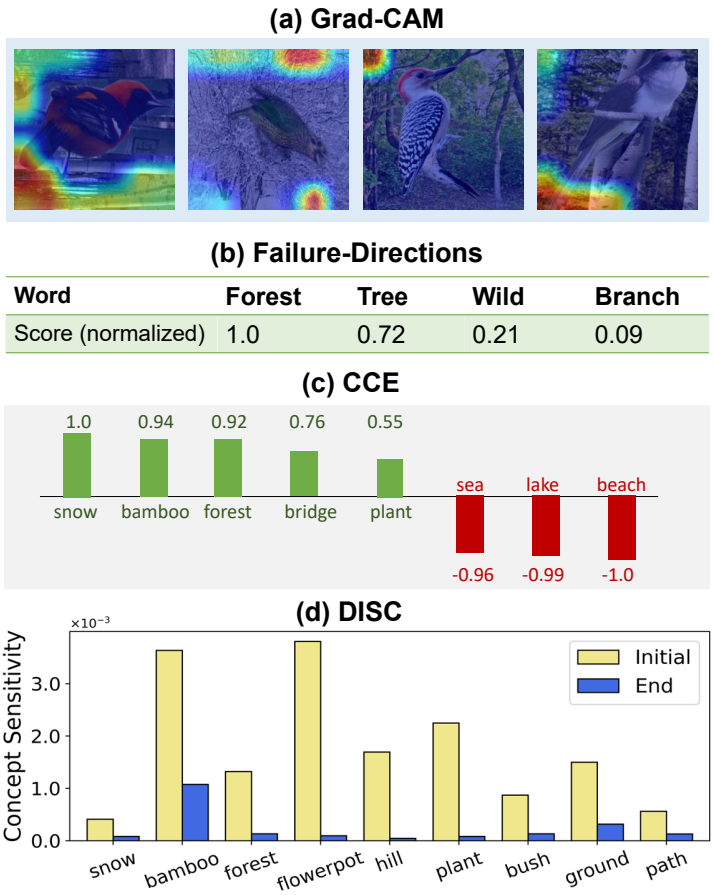


Figure 8. Different interpretations on Waterbirds explaining "**why the images are predicted as land birds**?" . (a) Grad-CAM visualization. (b) The word score generated by Jain et al. (2022). (c) The averaged concept scores when generating counterfactuals using CCE (Abid et al., 2022). (d) The concept sensitivity of spurious concepts on *landbird* class before the after the DISC training.

G. Results of Ablation and Sensitivity Study

Table 8. All experimental results of the ablation of model design choices.

	MetaShift		Waterbirds		FMoW		ISIC
	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.	Avg. Acc.	Worst Acc.	Avg. AUROC
DISC-Randint	71.7%	64.5%	91.0%	85.9%	53.0%	32.1%	49.3%
DISC-Reweight	72.8%	62.5%	88.9%	81.4%	51.0%	32.0%	35.9%
DISC-Inadaptive	73.0%	68.3%	89.6%	86.5%	51.9%	31.8%	47.1%
DISC	75.4%	72.6%	93.8%	88.7%	53.9%	36.1%	55.1%

Ablation Results. In Table 8, we report the ablation results on all the datasets. The conclusions are consistent with our statements in the main paper. Specifically, DISC outperforms the ablation models by large margins, validating our algorithm design empirically.

Unsupervised Clustering. We use the Silhouette score as a heuristic to search for the hyper-parameter of cluster number per class. As shown in Figure 9, interestingly, we found this metric well aligns with the testing performances on most datasets. Empirically, we found a small number of clusters per class, *e.g.,* 3, generally achieves the best results. One potential explanation is that when the number of clusters increases, the concept sensitivity could be passive and arbitrary by recognizing insignificant spurious concepts. We believe this is also an interesting perspective to investigate concept sensitivity or, in general, environment construction, in future works.

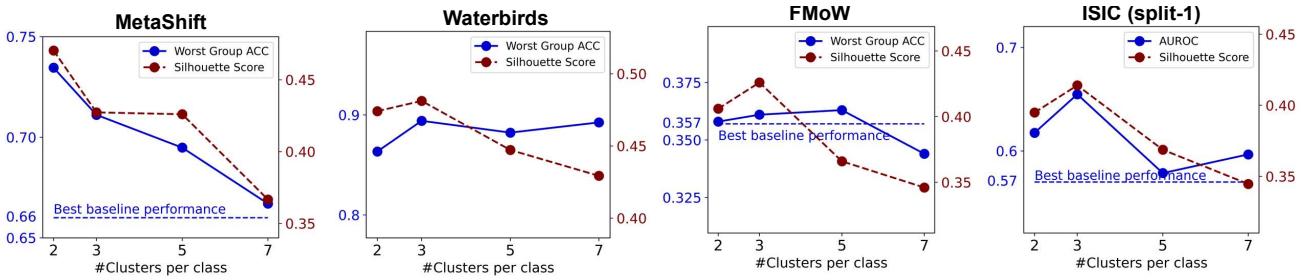


Figure 9. Worst Group Accuracy and Silhouette score *w.r.t.* number of clusters per class. For the ISIC dataset, we report the sensitivity result on one of the train-test splits.

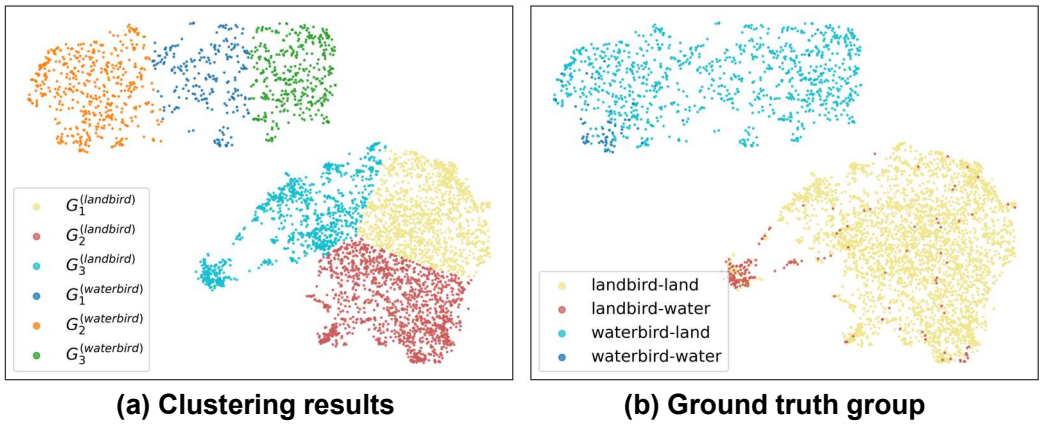


Figure 10. Comparison of clustering and group assignments on Waterbirds.

Besides the clustering results of MetaShift in Figure 6, we visualize the clustering results on Waterbirds in Figure 10. We found the clustering algorithm is able to capture part of the spurious attributes. Yet, good data environments could be difficult to find with extremely uneven groups. While DISC also outperforms most of the baselines, these results suggest that DISC is more robust even with "imprecise" environments.