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).

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, leverag-
ing 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 in the bottom 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; Yükselgönül 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 (Bontempielli 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, Bontempielli et al. (2022) propose ProtoPDebug that allows a human supervisor to provide feedback to part-prototype networks (Chen et al., 2019) (ProtoP-Nets) 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 ProtoP-Nets) and often require human annotation to identify the concepts. See Table 6 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 attributes to remove bias (Yaghoobzadeh et al., 2021; Utama et al., 2020; Dagaev et al., 2021; Zhang et al., 2022b; Nam
We consider a supervised image classification problem where we discuss the discovery of spurious concepts in Section 3.3. Empirical Risk Minimization (ERM) minimizes the empirical loss for a model trained with the ERM. This falls short of generalizing to the test distribution, i.e., \( \mathbb{P}_{te} \neq \mathbb{P}_{tr} \). Thus, the model trained with the ERM fails short of generalizing to datasets \( \mathbb{D}_{te} \sim \mathbb{P}_{te} \) 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” cannot 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 \( f = g \circ h \), where \( g \) is an encoder, and \( h : \mathbb{R}^d \rightarrow \mathbb{R}^{[d]} \) 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. 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.

**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 (Inconsistency 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 with the label with spurious concepts, 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

$$G_{j} = \bigcup_{y=1}^{|\mathcal{Y}|} G_{j}^{(y)}, \quad j = 1, ..., k,$$

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

$$M_{j} = \nabla_{\omega} \mathbb{E}_{(x,y) \sim G_{j}} \ell(f(x), y) = \frac{\partial \mathbb{E}_{(x,y) \sim G_{j}} \ell(f(x), y)}{\partial \omega},$$

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 $v_{c_{i}}$ of concept $c_{i}$. Then, we
compute $v_i \cdot M_j^T \in \mathbb{R}^{|Y|}$, 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

$$S_i = \text{Var}(\{ (v_i \cdot M_j^T)_{y'} | j = 1, \ldots, k \}), \quad (4)$$

where $y' = \text{argmax}_y (\sum_{j=1}^{k} v_i \cdot M_j^T)_y$.

Here $y'$ is the class dominated by or strongly associated with the concept $c_i$. Var is the variance operator. For convenience, we refer to $(v_i \cdot M_j^T)_{y'}$ 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 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

$$X' = \lambda X + (1 - \lambda) X_c^{(\beta_1, \beta_2, \lambda)}, \quad Y' = Y, \quad (5)$$

where $\lambda \sim \text{Beta}(\beta_1, \beta_2)$. $X_c^{(\beta_1, \beta_2)}$ denotes concept images sampled with probability $P^{(y)}$ from concept bank $C$. $(X, Y)$ are drawn from the subset where each $Y_i \in Y \setminus \{y\}$. 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.

**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.
Discover and Cure: Concept-aware Mitigation of Spurious Correlation

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

<table>
<thead>
<tr>
<th></th>
<th>MetaShift</th>
<th>Waterbirds</th>
<th>FMoW</th>
<th>ISIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM</td>
<td>72.9 ± 1.4</td>
<td>62.1 ± 4.8</td>
<td>53.0 ± 0.6</td>
<td>36.4 ± 0.7</td>
</tr>
<tr>
<td>ERM+aug</td>
<td>75.5 ± 1.7</td>
<td>57.5 ± 3.3</td>
<td>55.5 ± 0.4</td>
<td>35.7 ± 0.3</td>
</tr>
<tr>
<td>UW</td>
<td>72.1 ± 0.9</td>
<td>60.5 ± 3.8</td>
<td>52.5 ± 0.5</td>
<td>30.7 ± 1.5</td>
</tr>
<tr>
<td>IRM</td>
<td>73.9 ± 0.8</td>
<td>64.7 ± 2.1</td>
<td>50.8 ± 0.1</td>
<td>30.0 ± 1.4</td>
</tr>
<tr>
<td>IB-IRM</td>
<td>74.8 ± 0.2</td>
<td>65.6 ± 1.1</td>
<td>49.5 ± 0.5</td>
<td>28.4 ± 0.9</td>
</tr>
<tr>
<td>V-REx</td>
<td>72.7 ± 1.7</td>
<td>60.8 ± 5.5</td>
<td>48.0 ± 0.6</td>
<td>27.2 ± 0.8</td>
</tr>
<tr>
<td>CORAL</td>
<td>73.6 ± 0.4</td>
<td>62.8 ± 2.7</td>
<td>50.5 ± 0.4</td>
<td>31.7 ± 1.2</td>
</tr>
<tr>
<td>Fish</td>
<td>64.4 ± 2.0</td>
<td>53.2 ± 4.5</td>
<td>51.8 ± 0.3</td>
<td>34.6 ± 0.2</td>
</tr>
<tr>
<td>GroupDRO</td>
<td>73.6 ± 2.1</td>
<td>66.0 ± 3.8</td>
<td>51.8 ± 0.3</td>
<td>34.6 ± 0.2</td>
</tr>
<tr>
<td>JTT</td>
<td>74.4 ± 0.6</td>
<td>64.6 ± 2.3</td>
<td>52.5 ± 0.3</td>
<td>33.4 ± 0.9</td>
</tr>
<tr>
<td>DM-ADA</td>
<td>74.0 ± 0.8</td>
<td>65.7 ± 1.4</td>
<td>51.6 ± 0.2</td>
<td>34.2 ± 0.8</td>
</tr>
<tr>
<td>LISA</td>
<td>70.0 ± 0.7</td>
<td>59.8 ± 2.3</td>
<td>52.8 ± 0.9</td>
<td>35.5 ± 0.7</td>
</tr>
<tr>
<td><strong>DISC</strong></td>
<td>75.1 ± 1.1</td>
<td>73.5 ± 1.4</td>
<td>53.9 ± 0.4</td>
<td>36.1 ± 1.8</td>
</tr>
</tbody>
</table>

4. Experiments

We aim 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 Bisso (Bisso 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 4. 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 Bisso 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
whether the concept sensitivity faithfully reflects the spurious correlation are among the top spurious attributes. Interestingly, 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 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) → (fireplace, bedrooms, paisley) on MetaShift and (bamboo, forests, flowerpot) → (canopy, ground, plant) on Waterbirds. Interestingly, we find that the interpretations before and after the removal have some conceptual overlappings (e.g., “bed” → “bedrooms” on MetaShift, “forests” → “canopy” and “flowerpot” → “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...
Discover and Cure: Concept-aware Mitigation of Spurious Correlation

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 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 \(\exp\{-\sum_{i=1}^{m} P_{y_i} \cdot \max\{0, \frac{g(x_j)^{T} v_i}{|g(x_j)||v_i|}\}\}\), 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 7 (Appendix G), we report the ablation results. 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 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.
Figure 6. Visualization on the clustering and group assignments.

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 is 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, $\mu \in \mathbb{R}^p_i$ represents the signal of the causal concept with number of dimensions being $p_1$, and $\epsilon_{inv}$ is the noise term that is Gaussian with mean 0 and covariance matrix $\Sigma_1$. We further assume that the classes are balanced, i.e., $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$. The distributions of the causal and test environments are assumed to be invariant across training and test environments. In addition, the spurious concepts are modeled as $x_{spu} = \gamma_i^{(i)} + \epsilon_{spu}$, where $\gamma_i^{(i)} \in \mathbb{R}^p_i$ 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 $P_{te}$. Similar to $\epsilon_{inv}$, $\epsilon_{spu}$ is the noise term and respects a Gaussian distribution with mean 0 and covariance matrix $I$. As each coordinate of $\mu$ and $\gamma$ represents a concept, we assume that the values of $\gamma_i^{(i)}$ 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 $\ell_2$ 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 classifier being constructed as $\hat{C}_{ERM}(x) = \text{sgn}(\hat{\mu}_{ERM}^\top x_{inv} + \hat{\gamma}_{ERM}^\top x_{spu})$. Similarly, we define the classifier produced by DISC as $\hat{C}_{DISC}(x) = \text{sgn}(\hat{\mu}_{DISC}^\top x_{inv} + \hat{\gamma}_{DISC}^\top x_{spu})$, where $\hat{\mu}_{DISC}$ and $\hat{\gamma}_{DISC}$ are the solution produced by Algorithm 1.

Theorem 1. Assuming that (1), supp($\gamma_i^{(i)}$)'s are disjoint for different $y$'s, and $\text{Var}(\{\gamma_{y,i}^{(i)}\}_{i=1}^k) > K_0$ for $j \in [p_2]$ and some constant $K_0 > 0$, (2). $\|\mu\|_\infty \to 0$ when $p_1 \to \infty$, and $K_1 \leq \lambda_{\min}(\Sigma_1) \leq \lambda_{\max}(\Sigma_1) \leq K_2$ 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

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

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.

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).

Acknowledgement

We thank Serina Chang, Zhi Huang, Lingjiao Chen, Boyang Deng, Ruocheng Wang, Yang Zheng at Stanford University who gave great suggestions to improve our work. We also thank the anonymous reviewers at ICML2023 conference for their insightful comments. 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.
Discover and Cure: Concept-aware Mitigation of Spurious Correlation

References


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


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


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.


A. Notation Table

Table 2. Main notations used in the method section. Click here to return to the main paper.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{tr}/D_{te}$</td>
<td>Training/Testing dataset</td>
</tr>
<tr>
<td>$P_{tr}/P_{te}$</td>
<td>Training/Testing distribution</td>
</tr>
<tr>
<td>$\mathcal{Y}$</td>
<td>Label space</td>
</tr>
<tr>
<td>$f$</td>
<td>A deep model</td>
</tr>
<tr>
<td>$g/h$</td>
<td>The encoder/last linear layer of $f$</td>
</tr>
<tr>
<td>$\phi/\omega$</td>
<td>Parameters of $f/h$</td>
</tr>
<tr>
<td>$k$</td>
<td>Number of clusters per class</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of concepts in the concept bank</td>
</tr>
<tr>
<td>$d$</td>
<td>Number of hidden dimensions</td>
</tr>
<tr>
<td>$C$</td>
<td>A concept bank</td>
</tr>
<tr>
<td>$c_i$</td>
<td>The $i$-th concept in the concept bank $C$</td>
</tr>
<tr>
<td>$v_i$</td>
<td>The concept activation vector of concept $c_i$</td>
</tr>
<tr>
<td>$y_i'$</td>
<td>The dominant class of concept $c_i$</td>
</tr>
<tr>
<td>$\mathcal{P}_{c_i}$</td>
<td>The distribution of the images from the $i$-th concept</td>
</tr>
<tr>
<td>$S_i$</td>
<td>The concept sensitivity of concept $c_i$</td>
</tr>
<tr>
<td>$I_{p_i}/I_{n_i}$</td>
<td>The positive/negative image set for concept $c_i$</td>
</tr>
<tr>
<td>$N_p/N_n$</td>
<td>The number of images in positive/negative image set</td>
</tr>
<tr>
<td>$G_j(y)$</td>
<td>The $j$-th cluster in class $y$</td>
</tr>
<tr>
<td>$G_j$</td>
<td>The $j$-th environment</td>
</tr>
<tr>
<td>$M_j$</td>
<td>The Environment Gradient Matrix corresponding to $G_j$</td>
</tr>
<tr>
<td>$H(y)$</td>
<td>A boolean mask of concepts for the dominant class $y$</td>
</tr>
<tr>
<td>$P(y)$</td>
<td>Concept sampling distribution for class $y$</td>
</tr>
<tr>
<td>$X(C,P(y))$</td>
<td>Images sampled from concept bank $C$ with probability $P(y)$</td>
</tr>
</tbody>
</table>

B. Theoretical Analysis

Theorem (Restatement of Theorem 1). Assuming that \(\text{supp}(\gamma_i^{(1)})\)’s are disjoint for different $y$’s, and $\text{Var}(\{\gamma_i^{(1)}\}_{i=1}^k) > K_0$ for $j \in [p_2]$ and some constant $K_0 > 0$, (2). $\|\mu\|_\infty \to 0$ when $p_1 \to \infty$, and $K_1 \leq \lambda_{\min}(\Sigma_1) \leq \lambda_{\max}(\Sigma_1) \leq K_2$ 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

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

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

- **(1)** The support operation $\text{supp}(\cdot)$ 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 $\Sigma_1$ 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.
As a result, with probability at least $\frac{1}{2}$. Also, by assumption, we have $\|\gamma_k\|_\infty \leq 1$. Then, as we now consider the binary classification setting, the $\gamma_k$ is the basis vector for the $k$-th feature, and 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

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

The corresponding parts for the causal and spurious features are respectively

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

and

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

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

$$M_{j,y; \text{inv}} = \mathbb{E}[X^T_{\text{inv}}y\mathbf{1}_B/B - X^T_{\text{inv}}X_{\text{inv}}\mu_t/B - X^T_{\text{inv}}X_{\text{spu}}\gamma_t/B] + O\left(\frac{p_1}{n}\right) = \mu - (\mu^T + \Sigma_1)\mu_t - \gamma_t^T \gamma_{y_i}. \mu + O\left(\frac{p_1}{n}\right),$$

and

$$M_{j,y; \text{spu}} = \mathbb{E}[X^T_{\text{spu}}y\mathbf{1}_B/B - X^T_{\text{spu}}X_{\text{inv}}\mu_t/B - X^T_{\text{spu}}X_{\text{spu}}\gamma_t/B] + O\left(\frac{p_2}{n}\right) = y^T \gamma_i - \mu^T \mu_t - \gamma_{y_i}^T \gamma_{y_i} - (\gamma_i^T \gamma_{y_i}^T + I)\gamma_t + O\left(\frac{p_2}{n}\right).$$

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

$$S^{(y)}_{i,j} = \text{Var} \left( \{ \gamma_{y_i} \} \right).$$

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

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

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

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

As a result, with probability at least $1 - o(1)$, the sampling according to $P^{(y)}$ will always draw from the spurious concepts from $U_{i=1}^k \text{supp} (\gamma_{y_i})$. 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 $c_j$, with the $j$-th entry equal to 1, indicating the presence of this concept.

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

$$S^{(y)}_{\text{spu}} = \sum_{y \in \{-1,1\}} y^T (\tilde{\gamma}_{y_i} + c_{y, \text{spu}}) - \mu^T \mu_t \cdot (\sum_{y \in \{-1,1\}} \tilde{\gamma}_{y_i} + c_{y, \text{spu}}) - (\sum_{y \in \{-1,1\}} (\tilde{\gamma}_{y_i})^T + c_{y, \text{spu}})^T) + I \gamma_t + O\left(\frac{p_2}{n}\right).$$

---

**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. Under our model set-up, at iteration $t$, we have for $j \in [k]$, the $M_j$ in (3) equals to
Note that we assume \( y \in \{-1,1\} \). Using the fact that the supports of \( \hat{\gamma}_y \) and \( c_{y,spu} \) are disjoint, we have that

\[
S^t_{spu} = \sum_{y \in \{-1,1\}} y \cdot (\hat{\gamma}_y + c_{y,spu}) - \mu^\top \mu_t \cdot \left( \sum_{y \in \{-1,1\}} \hat{\gamma}_y + c_{y,spu} \right) - \left( \sum_{y \in \{-1,1\}} (\hat{\gamma}_y + c_{y,spu}) (\hat{\gamma}_y + c_{y,spu})^\top + I \right) \gamma_t + O\left( \frac{p_2}{n} \right).
\]

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

\[
S^t_{inv} = \mu - (\mu^\top + \Sigma c) \mu_t - \gamma^\top \gamma^{(i)} \cdot \mu + O\left( \frac{p_2}{n} \right).
\]

As a result, the update in each iteration \( t \) of is equivalent to running gradient descent on minimizing the loss function \( \ell(\hat{\mu}; \hat{\gamma}) = (\hat{\mu}; \hat{\gamma})^\top \left( \sum_{y \in \{-1,1\}} (\hat{\gamma}_y + c_{y,spu}) (\hat{\gamma}_y + c_{y,spu})^\top + I \right) (\hat{\mu}; \hat{\gamma}) + \mu \cdot \gamma_t \cdot \mu + O\left( \frac{p_2}{n} \right) \).

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

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

Since \( \langle \hat{\mu}_{DISC}; \hat{\gamma}_{DISC} \rangle \) minimizes \( \ell \), we can write out its analytical solution as

\[
\hat{\mu}_{DISC} = (\Sigma c + \mu^\top)^\top \mu + O\left( \frac{p_2}{n} \right),
\]

and

\[
\hat{\gamma}_{DISC} = \left( \sum_{y \in \{-1,1\}} (\hat{\gamma}_y + c_{y,spu}) (\hat{\gamma}_y + c_{y,spu})^\top + I \right)^{-1} \sum_{y \in \{-1,1\}} y \cdot (\hat{\gamma}_y + c_{y,spu})
\]

Similarly, we have

\[
\hat{\mu}_{ERM} = (\Sigma c + \mu^\top)^\top \mu + O\left( \frac{p_2}{n} \right),
\]

and

\[
\hat{\gamma}_{ERM} = \left( \sum_{y \in \{-1,1\}} (\hat{\gamma}_y) (\hat{\gamma}_y)^\top + I \right)^{-1} \sum_{y \in \{-1,1\}} y \cdot \hat{\gamma}_y.
\]

Since all the entries of \( \hat{\gamma}_y \) are either 0 or 1, all the entries of \( c_{y,spu} \) are between 0 and 1, and the support of \( \hat{\gamma}_y \) 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

\[
P_{D_{\text{te}}}(\text{sgn}(\hat{\mu}^\top x_1 + \hat{\gamma}^\top x_2) \neq y) = \frac{1}{2} P_{D_{\text{te}}}((\hat{\mu}^\top \mu_1 + \hat{\mu}^\top \epsilon_1 + \hat{\gamma}^\top \epsilon_2 > 0) + \frac{1}{2} P_{D_{\text{te}}}((\hat{\mu}^\top \mu_1 + \hat{\mu}^\top \epsilon_1 + \hat{\gamma}^\top \epsilon_2 < 0)
\]

\[
= \frac{1}{2} \mathbb{E}[P(\hat{\mu}^\top \mu_1 + \hat{\mu}^\top \epsilon_1 + \hat{\gamma}^\top \epsilon_2 > 0 | \epsilon_1)] + \frac{1}{2} \mathbb{E}[P(\hat{\mu}^\top \mu_1 + \hat{\mu}^\top \epsilon_1 + \hat{\gamma}^\top \epsilon_2 < 0 | \epsilon_1)]
\]

\[
= \mathbb{E}[\Phi(- \frac{\hat{\mu}^\top \mu + \hat{\gamma}^\top \epsilon}{\sqrt{\|\mu\|^2 + \|\gamma\|^2}})],
\]

where \( \Phi \) 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

\[
P_{D_{\text{te}}}(\hat{\gamma}_{DISC}(x) \neq y) < P_{D_{\text{te}}}(\hat{\gamma}_{ERM}(x) \neq y).
\]

\[
\square
\]
C. Datasets

Table 2. (a) Metashift Dataset.

<table>
<thead>
<tr>
<th>Image:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group $g$:</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Target $y \in {0, 1}$:</td>
<td>0 (cat)</td>
<td>0 (cat)</td>
<td>1 (dog)</td>
<td>1 (dog)</td>
<td>0 (cat)</td>
</tr>
<tr>
<td>Spurious $s$:</td>
<td>0 (sofa)</td>
<td>1 (bed)</td>
<td>2 (bench)</td>
<td>3 (bike)</td>
<td>4 (shelf)</td>
</tr>
<tr>
<td># Train data:</td>
<td>231</td>
<td>380</td>
<td>145</td>
<td>367</td>
<td>-</td>
</tr>
<tr>
<td># Val data (OOD):</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>34</td>
</tr>
<tr>
<td># Test data:</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>201</td>
</tr>
</tbody>
</table>

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

Table 2. (b) Waterbirds Dataset.

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

Target: bird type; Spurious feature: background type.

Table 2. (c) FMoW Dataset.

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

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).
Discover and Cure: Concept-aware Mitigation of Spurious Correlation

Table 2. (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.

<table>
<thead>
<tr>
<th>Image:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Target $y \in {0,1}$:</td>
<td>0 (benign)</td>
<td>0 (benign)</td>
<td>1 (malignant)</td>
</tr>
<tr>
<td>Spurious s:</td>
<td>patch, gel border</td>
<td>ink, hair</td>
<td>dark corner, gel bubble</td>
</tr>
<tr>
<td># Train data:</td>
<td>1,826</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Val data:</td>
<td>154</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Test data:</td>
<td>618</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Spurious features:** dark corners, hair, gel borders, gel bubbles, ruler, ink markings/staining, patches.

**Target:** benign / melanoma skin lesions

**D. Concept Bank**

**Concept categories.** In Table 3, 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>
<thead>
<tr>
<th>Concept category</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>[blackness, blueness, greenness, redness, whiteness]</td>
</tr>
<tr>
<td>Texture</td>
<td>[concrete, granite, leather, laminate, metal, blotchy, blurriness, stripes, polka dots, knitted, cracked, frilly, waffled, scaly, lace-like, 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]</td>
</tr>
<tr>
<td>Nature</td>
<td>[bamboo, beach, bush, canopy, earth, field, flower, flowerpot, fluorescent, forest, grass, ground, harbor, hill, lake, mountain, muzzle, path, plant, river, sand, sea, snow, tree, water]</td>
</tr>
<tr>
<td>City</td>
<td>[awning, base, bench, building, earth, fence, field, ground, house, manhole, path, snow, streets]</td>
</tr>
<tr>
<td>Household</td>
<td>[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, diningrooms, 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]</td>
</tr>
<tr>
<td>Others</td>
<td>[bird, cat, cow, dog, horse, mouse, paw, arm, baby, 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, heat, 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]</td>
</tr>
</tbody>
</table>

**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.
Discover and Cure: Concept-aware Mitigation of Spurious Correlation

**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.

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

### Table 4. Selected concept categories for each dataset

<table>
<thead>
<tr>
<th>Concept categories</th>
<th>MetaShift</th>
<th>Waterbirds</th>
<th>FMoW</th>
<th>ISIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Texture</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Nature</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>City</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Household</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Others</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Concept category selection and filtering.** As shown in Table 4, 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 CAs 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 3.

**Learning CAs.** To learn the CAs, we use $N^p = N^u = 150$ 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 5. 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 5. 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 CAs to generate interpretations. Moreover, DISC offers more dynamic inspection during model training.
Table 6. Comparison between interpretations of DISC and the existing methods.

<table>
<thead>
<tr>
<th>Explanation types</th>
<th>Class/group-wise</th>
<th>Concept/caption-based</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singla &amp; Feizi (2022)</td>
<td>✓ (partial)</td>
<td>✓ (partial)</td>
<td>✗</td>
</tr>
<tr>
<td>Selvaraju et al. (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singla et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sohoni et al. (2020)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seo et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selvaraju et al. (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singla et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creager et al. (2021)</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Liu et al. (2021a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ahmed et al. (2021)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abid et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bontempelli et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jain et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyuboglu et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. (2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ahmed et al. (2021)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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 7. All experimental results of the ablation of model design choices.

<table>
<thead>
<tr>
<th></th>
<th>MetaShift</th>
<th>Waterbirds</th>
<th>FMoW</th>
<th>ISIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISC-Randint</td>
<td>71.7%</td>
<td>64.5%</td>
<td>91.0%</td>
<td>85.9%</td>
</tr>
<tr>
<td>DISC-Reweight</td>
<td>72.8%</td>
<td>62.5%</td>
<td>88.9%</td>
<td>81.4%</td>
</tr>
<tr>
<td>DISC-Inadaptive</td>
<td>73.0%</td>
<td>68.3%</td>
<td>89.6%</td>
<td>86.5%</td>
</tr>
<tr>
<td>DISC</td>
<td>75.4%</td>
<td>72.6%</td>
<td>93.8%</td>
<td>88.7%</td>
</tr>
</tbody>
</table>

Ablation Results. In Table 7, 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.

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.