Label-Efficient Group Robustness via Out-of-Distribution Concept Curation

Yiwei Yang[1] Anthony Z. Liu[2] Robert Wolfe[1] Aylin Caliskan[1] Bill Howe[1]
[1]University of Washington, [2]University of Michigan
{yanyiwei, rwolfe3, aylin, billhowe}@uw.edu anthliu@umich.edu

Abstract

Deep neural networks are prone to capture correlations between spurious attributes and class labels, leading to low accuracy on some combinations of class labels and spurious attribute values. When a spurious attribute represents a protected class, these low-accuracy groups can manifest discriminatory bias. Existing methods attempting to improve worst-group accuracy assume the training data, validation data, or both are reliably labeled by the spurious attribute. But a model may be perceived to be biased towards a concept that is not represented by pre-existing labels on the training data. In these situations, the spurious attribute must be defined with external information. We propose Concept Correction, a framework that represents a concept as a curated set of images from any source, then labels each training sample by its similarity to the concept set to control spurious correlations. For example, concept sets representing gender can be used to measure and control gender bias even without explicit labels. We demonstrate and evaluate an instance of the framework as Concept DRO, which uses concept sets to estimate group labels, then uses these labels to train with a state of the art distributively robust optimization objective. We show that Concept DRO outperforms existing methods that do not require labels of spurious attributes by up to 33.1% on three image classification datasets and is competitive with the best methods that assume access to labels. We consider how the size and quality of the concept set influences performance and find that even smaller, manually curated sets of noisy AI-generated images are effective at controlling spurious correlations, suggesting that high-quality, reusable concept sets are easy to create and effective in reducing bias.

1. Introduction

Deep learning models are prone to capture spurious correlations, patterns that are predictive of the target class in training data but are inherently irrelevant to the classification task. For example, consider an image classification task to distinguish doctors from nurses. Suppose that in the training data, 95% of the doctors present as men, and 95% of the nurses present as women. A model trained with the standard empirical risk minimization (ERM) tends to-classify images of men as doctors and images of women as nurses by learning a spurious correlation between gender-expression features (e.g., hairstyle, clothing) and occupation. While the model may achieve high average accuracy, it performs poorly on the minority groups (in this example, female-presenting doctors and male-presenting nurses), thus exhibiting low worst-group accuracy. Such spurious correlations happen in a wide variety of real-world tasks, including facial recognition, medical imaging, and language tasks [3, 8, 11, 17, 18, 23].

Existing approaches typically consider the setting in which each training (or validation) sample is explicitly assigned a *group*, represented as a pair (y, a) where y is a class label and a is a value from the (potentially unknown) spurious attribute. For instance, Group DRO assumes the availability of group labels during training, and directly minimizes the worst-group loss [17]. Recognizing that obtaining labels requires expensive human annotation, several recent approaches avoid relying on group labels during training [7, 9, 19, 24]. These approaches train an ERM model to infer group labels, then a second model is trained to optimize for worst-group loss with the inferred labels. However, these approaches require validation labels for crucial hyperparameter tuning [9], since their effectiveness depends on the accuracy of the group inference. One exception is GEORGE [19], which assumes that the feature space of a model trained for the predictive task is easily separable by groups and learns group assignment via clustering. GEORGE then optimizes for worst-cluster accuracy without requiring labels. However, while GEORGE is effective when the input features are simple and cluster well (e.g. colored digits), its performance drops significantly when the input features are more complex and clusters become less well-defined (e.g., for human faces) [19]

We propose Concept Correction (figure 1), a framework that instead uses a set of example images to represent a concept involved in the spurious correlation. These examples can be out-of-distribution and task-agnostic, as long as

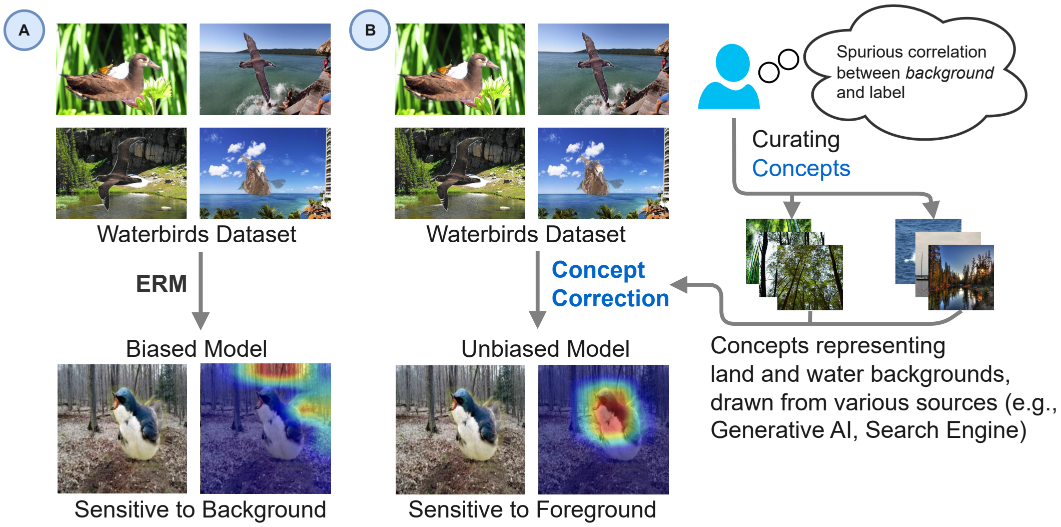
A
Waterbirds Dataset
ERM
Biased Model
Sensitive to Background
B
Waterbirds Dataset
Concept Correction
Unbiased Model
Sensitive to Foreground
Spurious correlation between *background* and label
Curating Concepts
Concepts representing land and water backgrounds, drawn from various sources (e.g., Generative AI, Search Engine)

Figure 1. We show our main idea *Concept Correction*, a framework for correcting spurious attributes in biased models using concepts, set of examples representing some pattern. We show an example from the Waterbirds dataset [17], where (a) directly training an ERM results in a model that spuriously correlates background type with bird type. A practitioner can correct this by (b) finding images with land and water backgrounds and using concept correction. Notably, these example images can be obtained outside of the original Waterbirds datset, such as using generative AI, and using search engines.

they contain the features representing the spurious attribute. For example, concept sets designed to counteract a spurious attribute of gender presentation may include a set of images of female-presenting people, but without any reference to the task-specific class label of doctor or nurse. Since the examples need not be selected from the training distribution, they are reusable across models and tasks, and practitioners can retrieve them from various sources such as a search engines, datasets on the Web, or image generation models, all of which require significantly less cognitive and technical effort than manually labeling in-distribution data. In fact, by using images generated with only two lines of prompts to the text-to-image generator Stable Diffusion [16], we show that an implementation of Concept Correction can achieve competitive results to state-of-the-art methods that require manual labeling of tens of thousands of data points [9, 17, 24].

We demonstrate an instance of Concept Correction via Concept DRO (CDRO), an approach that leverages concepts to infer group labels, then trains with a distributively robust optimization objective using the inferred labels. Similar to the existing two-stage methods, CDRO first trains a neural network with ERM. Then, exploiting the fact that the ERM model learns spurious attribute labels[7], we use the model to extract features from the concept sets, the training data, and the validation data. We then train a classifier to discriminate concepts then use the weights of this classifier to determine the extent to which each training sample exhibits that concept. These concept similarity scores are then used to infer a spurious attribute label for every data point in the training and validation sets, providing the information needed by state-of-the-art methods to improve worst-group accuracy. In our example, CDRO trains a classifier on two sets of images representing the concepts of presenting-female and presenting-male, then uses the coefficients of this classifier to estimate the extent to which each training sample exhibits ”maleness” vs. ”femaleness.”

We evaluate CDRO on three image classification datasets with spurious correlations: CMNIST [1], Waterbirds [17], and CelebA [10]. Among methods that do not assume spurious attribute labels are provided, CDRO improves worst-group accuracy by up to 12.6% on Waterbirds and 33.1% on CelebA. CDRO also achieves competitive performance to methods that *do* assume access to spurious attribute labels. Finally, we show that CDRO is robust to the quality of concepts in two aspects: size of concept and distribution distance from training data.

Our contributions are summarized as follows:

- We propose Concept Correction, a label-efficient framework that uses out-of-distribution sets of images to correct spurious correlations.
- We present CDRO, an instance of the Concept Correction framework that uses concepts to infer labels of spurious attributes and then trains with a distributively robust optimization objective.
- We show CDRO significantly outperforms ERM and GEORGE across three benchmarks while achieving competitive results to methods that require group labels during validation.

2. Preliminaries

In this section, we first present the problem setting, then describe concepts and the Concept Activation Vector (CAV) introduced by Kim *et al*. [6].

2.1. Problem Setting

We consider a classification setting, where given an input $x \in \mathcal{X}$, our goal is to predict $y \in \mathcal{Y}$. Let $X = \{x_1, ..., x_n\}$, $Y = \{y_1, ..., y_n\}$ be our training set of size n. Each datapoint x_i has a *spurious attribute* a_i, which is unobserved during training. Each x_i is assumed to belong to a group g_i, where g_i is a pair (y_i, a_i), the combination of a class label y_i and a spurious attribute label a_i. The spurious correlation problem occurs when accuracy across groups varies significantly; i.e., a model trained with ERM may achieve a high average accuracy, yet still performs poorly on some minority groups. We therefore evaluate classifiers on the accuracy of the worst-performing group.

We are interested in the setting where group information is not available for either the training set nor the validation set. Instead, we assume that users curate a set of images representing the concept without regard to any particular task. The human effort to curate the concept depends on the user's domain knowledge. For example, in Waterbirds, to curate a concept for 'landbird background,' one can generate images with the prompt 'a photo of land background'. However, if the user knows that land backgrounds of birds tend to be forests, the prompt can be adjusted accordingly.

2.2. Concepts and Concept Activation Vectors

A concept is defined using a set of images. For example, a set of ocean and lake images can represent the concept *water background*. The concept examples need not be drawn from the training data and need not be relevant to the classification task.

Given a concept C and a layer l of the neural network, a concept activation vector (CAV) $\mathbf{v}_C^l \in R^m$ [6] is defined as the vector normal to the hyperplane separating samples with a concept (the concept set P_C) and examples without a concept (the contrastive set N) in the model's activations. Given layer activations of both the concept set and the contrastive set, CAV can be obtained by training a linear classifier distinguishing the two sets of activations.

3. Concept Correction

We propose Concept Correction, a framework that uses concepts to correct spurious correlations, without relying on manual labeling of in-distribution data. Instead, Concept Correction takes a set of a small number of examples as input and leverages the examples to infer spurious attribute information. Then, a robust model is trained with the inferred group of spurious attribute information using any worst-group loss minimization algorithms.

Concept Correction expects the concept examples to represent the spurious attribute. As described in the introduction, these images can be out-of-distribution and irrelevant to the prediction task, de-coupling concept curation from the model setting. In fact, using generative models, practitioners can easily generate many examples with just a few lines of prompting. In contrast to existing methods that require manual labeling of either training or validation samples, concept curation does not require access to the training data. Practitioners can reason about potential biases the model may exhibit then curate concept sets accordingly.

Concept Correction is compatible with any method that requires spurious attribute or group information, such as Group DRO [17], Just-Train-Twice [9], Correct-n-Contrast [24], and more. For example, a simple way of applying Concept Correction to Correct-n-Contrast and Just-Train-Twice is simply replacing the validation group labels with inferred labels using concepts, alleviating the need of manual labeling of a large amount of data.

3.1. Concept DRO

We present Concept DRO (CDRO), an implementation of Concept Correction on Group DRO. As illustrated in Figure 2, in stage 1, we first train an ERM model and then train a linear classifier on concepts representing the spurious attribute to infer group labels. In stage 2, we use Group DRO to train a robust model with the inferred group labels. We include further details on both stages below, and give pseudocode for CDRO in the Appendix.

3.1.1 Stage 1: Inferring Labels of Spurious Attributes

We train a neural network with ERM on the training data. Given a spurious attribute $\mathcal{A} = \{a_1, ..., a_m\}$, we curate two sets of examples for each concept a_i: a concept set $C = \{c_1, ..., c_k\}$, which represents the presence of the concept, and a contrastive set $N = \{n_1, ..., n_k\}$, which represents the absence of the concept. We then use the ERM model as a feature extractor, by taking the first to the penultimate layer of the model, denoted as f, to extract features from the two sets. We then train a linear classifier on the ERM features of the concept sets, $\{f(c_1), ..., f(c_k)\}$ and $\{f(n_1), ..., f(n_k)\}$. The resulting coefficient of the linear classifier gives us a CAV pointing in the direction of the concept a_i. Then, for both training and validation data, denoted as $X = \{x_1, ..., x_n\}$, we use f to extract representations, $\{f(x_1), ..., f(x_n)\}$.

We then compute the cosine similarity between the representations and the CAV. We use these similarity scores to separate the data into groups — examples with the same attributes are likely to activate similar similarity scores. To separate points by similarity scores, we use a naive approach: we train a Gaussian Mixture Model (GMM) on the similarity scores, where the number of mixtures is set to m, the size of the domain of the spurious attribute. Typically (but not exclusively) $m = 2$, since many formulations of bias posit a binary distinction between minority and majority populations. The mixture with the highest mean similarity to the concept is labeled as $\mathcal{A} = a_i$. Intuitively, data with high similarity scores with respect to the concept set tend to exhibit concept a_i, and thus belong to groups with $\mathcal{A} = a_i$ (see figure 3). We repeat the above steps for all $a_i \in \{a_1, ..., a_m\}$. Lastly, for the data points with missing labels, we randomly assign one from $\{a_1, ..., a_m\}$.

In the case that the spurious attribute is binary, *i.e.*$\mathcal{A} = \{a_1, a_2\}$, we only need to curate concepts for a_1 or a_2 and train one CAV. Suppose we train a CAV for a_1, and label data with the CAV, yielding $X_{a_1} = \{x \in X | \mathcal{A}(x) = a_1\}$. We can then use the negative of the CAV to label a_2, which can be considered the complement, $X_{a_1}^C$.

Since Group DRO takes group labels as input, where a group label g is the pair (y, a) defined by the label y and spurious attribute a. We generate inferred group labels $\hat{g}$ using ground truth labels y and inferred spurious attribute labels $\hat{a}$ to use in Stage 2.

3.1.2 Stage 2: Optimize for worst-group loss with inferred labels

We use Group DRO to train a classifier with inferred group labels $\hat{g}$ from stage 1. Instead of overall loss, Group DRO optimizes for worst-group loss.

Group DRO uses ground truth group labels to select the best model checkpoint based on validation worst-group accuracy. We instead use the inferred labels to compute the surrogate worst-group accuracy at validation.

4. Experimental Results

In our experiments, we first show that CDRO substantially outperforms ERM and GEORGE, and achieves comparable worst-group accuracy to existing methods that require group labels. To explain the effectiveness of CDRO, we then demonstrate that CDRO separates data into corresponding groups with concepts, and infers labels of spurious attributes with high accuracy. Lastly, we show that CDRO is robust to the quality of concepts in two aspects: concept size and distribution distance from training data.

4.1. Setup

We describe three image datasets, corresponding concept sets, and baseline methods used in our experiments.

4.1.1 Datasets

We briefly describe the three image classification benchmarks used in our evaluation. Full dataset and training details are included in the Appendix.

CMNIST [1]: The task is to classify digits into one of the five classes: $\mathcal{Y} = \{(0,1), (2,3), (4,5), (6,7), (8,9)\}$. Each class is correlated with one of five colors: $\mathcal{A} = \{\texttt{red}, \texttt{yellow}, \texttt{green}, \texttt{blue}, \texttt{purple}\}$, respectively, each represented by its hex code (see figure 3).

Waterbirds [17]: The task is to classify birds into one of the 2 classes: $\mathcal{Y} = \{\text{waterbird}, \text{landbird}\}$. Each class is correlated with the backgrounds: $\mathcal{A} = \{\text{water background}, \text{land background}\}$, respectively.

CelebA [10]: The task is to classifiy celebrities' hair color $\mathcal{Y} = \{\text{blond}, \text{not blond}\}$, with the spurious attribute $\mathcal{A} = \{\text{female}, \text{male}\}$. The class blond correlates with female, and the class not blond correlates with male.

4.1.2 Baselines

We consider seven baselines that train models under different assumptions regarding the availability of group labels.

Group labels available during training - Group DRO [17] is a state-of-the-art method, which optimizes for the worst-group loss using group labels.

Group labels available during validation - Just Train Twice (JTT) [9] trains a model to detect minority group examples, then upweights those examples during training. Correct-n-Contrast (CNC) [24] infers group information similar to JTT, then uses contrastive learning to learn representations robust to spurious correlations. Automatic Feature Reweighting (AFR) [15] finetunes the last layer of the ERM model on a reweighted training set. Spread Spurious Attribute (SSA) [14] uses semi-supervised learning to train a predictor to infer pseudo-group label, then trains a robust model with worst-group loss minimization.

Group labels available during neither training nor validation - Empirical Risk Minimization (ERM) is the standard training procedure that minimizes the average loss. Similar to JTT, GEORGE also follows a two-step procedure by first estimating the group labels and then using these estimates to train a robust classifier. Unlike JTT, after training a standard model, the feature space of each class is split into groups via unsupervised clustering. During validation, GEORGE then optimizes for worst-cluster accuracy without requiring group labels.

4.1.3 Concepts

In our experiments we generated 50 images per concept set. We generated synthetic data as concepts for CMNIST, and used a popular off-the-shelf generative AI model Stable Diffusion [16] to generate concepts for Waterbirds and CelebA.

CMNIST - Since the 5 target classes are correlated with 5 colors, we curate concept sets for each color. For each color, we sample 50 images from the original MNIST dataset [2], which consists of white digits 0-9 on black background, and we paint the digits with the corresponding color. We then curate the contrastive set by concatenating the concept sets of the 4 remaining colors, and trained a CAV pointing towards the color.

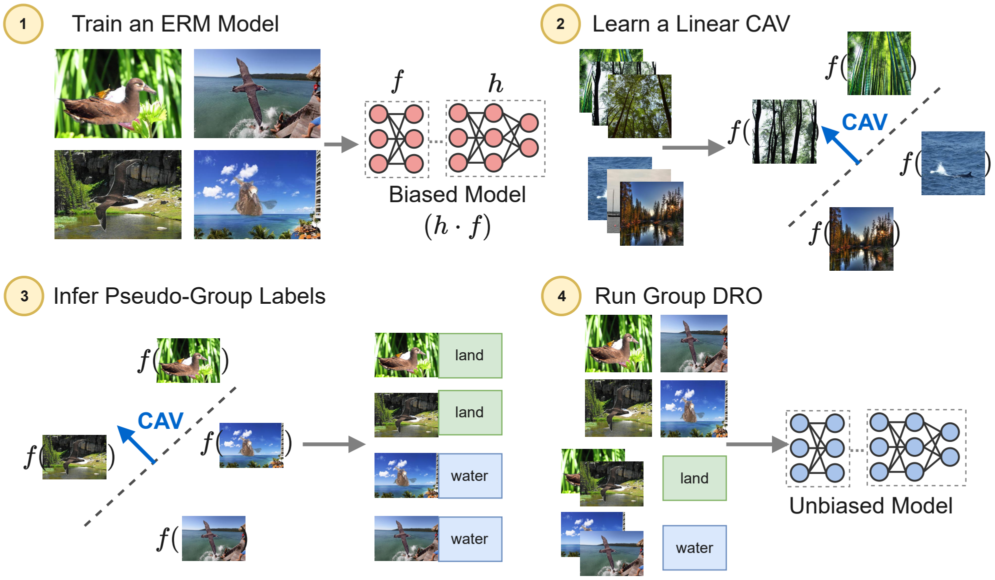

Figure 2. We show Concept DRO (CDRO), a method of training an unbiased model using a concept set. (1) First, we train a neural network model ($h \circ f$) using ERM. (2) We train a concept activation vector (CAV) using the given concept set and the model [6]. A CAV is a linear decision boundary between the model representations of the concepts ($f(\cdot)$). (3) We use the inferred CAV to infer pseudo-group labels over the training data. (4) Finally, we run Group DRO [17] the training data using the pseudo-labels.

Waterbirds - Since land background is correlated with landbirds, and water background is correlated with waterbirds, we curate sets of examples to represent the concepts "land background" and "water background." With the knowledge that land background in the dataset are either bamboo or broadleaf trees, we used prompts "A photo of bamboo trees" and "A photo of broadleaf trees" to generate a total of 50 (25 each) images to represent the "land background" concept. Similarly, knowing that the water background in the dataset are either ocean or lake, we used prompts "A photo of ocean" and "A photo of lake" to generate a total of 50 (25 each) images to represent the "water background" concept. Since the spurious attribute is binary, we trained a CAV in the direction of "land background," using the "water background" concept as the contrastive set.

CelebA - Since female is correlated with blond and male is correlated with not blond, we curate sets of examples to represent the concepts *female* and *male*. With the knowledge that the dataset consists of celebrities' faces, we used prompts "A photo of a female celebrity face" and "A photo of a male celebrity face" to generate a total of 100 images (50 each) for the *female* and *male* concepts, respectively. We trained a CAV with *female* as the concept set and *male* as the contrastive set, such that positive similarity can be interpreted as exhibiting female features.

4.2. Main Results

Table 1 reports the average and worst-group accuracies of all approaches. CDRO outperforms ERM across all tasks. Comparing to GEORGE, CDRO achieves higher worst-group accuracy on Waterbirds and CelebA. The performance of GEORGE relies on how well the learned features can be clustered. In CMNIST, where the features (color and shape) are easily distinguishable, GEORGE easily finds clusters that correspond to groups and therefore yields a high worst-group accuracy. However, in Waterbirds and CelebA, where the features are more complex, clustering is less effective and the worst-group accuracy is lower. Further, with only 100 out-of-distribution images, CDRO achieves similar performance to methods that require validation labels, which may require large-scale manual labeling: CelebA, for example, has 20K validation images. Specifically, CDRO outperforms JTT by 5% on Waterbirds, and 6.5% on CelebA and approaches the performance of Group DRO itself. CDRO outperforms CNC by 0.3% on Waterbirds, and falls short by 0.8% on CelebA. CDRO outperforms SSA with 5% validation labels by 1.7% on Waterbirds, and 1.3% on CelebA. Lastly, CDRO outperforms AFR with 5% validation labels by 2.2% on Waterbirds, and 10.5% on CelebA.

4.3. Ablation Studies

4.3.1 CDRO effectively infers group labels

To examine the effectiveness of CDRO in inferring group labels, we first visualize the distribution of both train and validation data by their cosine similarity scores to the corresponding CAV for all three datasets. As shown in Figure 3, for both train and validation data of the three datasets, data representations with high similarity scores tend to exhibit the concept that the CAV directed to.For example, Figure 3 shows that, given a CAV in the direction of the concept *red*,

Method	Group Info	CMNIST [1]		Waterbirds [17]		CelebA [10]	
		Worst(%)	Mean(%)	Worst(%)	Mean(%)	Worst(%)	Mean(%)
Group DRO [17]	Train & Val	74.7 (1.0) [1]	90.6 (0.1)	89.9 (0.6)	92.0 (0.6)	88.9 (1.3)	93.9 (0.1)
JTT [9]	Val	74.5 (2.4)	90.2 (0.8)	83.8 (1.2)	89.3 (0.7)	81.5 (1.7)	88.1 (0.3)
CNC [24]	Val	77.4 (3.0)	90.9 (0.6)	88.5 (0.3)	90.9 (0.1)	88.8 (0.9)	89.9 (0.5)
SSA [14]	Val	-	-	89.0 (0.6)	92.2 (0.9)	89.8 (1.3)	92.8 (0.1)
	5% Val	-	-	87.1 (0.7)	92.6 (0.2)	86.7 (1.1)	92.8 (0.3)
AFR [15]	Val	-	-	90.4 (1.1)	94.2 (1.2)	82.0 (0.5)	91.3 (0.3)
	5% Val	-	-	86.6 (4.7)	-	77.5 (1.3)	-
GEORGE [19]	None	76.4 (2.3)	89.5 (0.3)	76.2 (2.0)	95.7 (0.5)	54.9 (1.9)	94.6 (0.2)
ERM	None	0.0 (0.0)	20.1 (0.2)	62.6 (0.3)	97.3 (1.0)	47.7 (2.1)	94.9 (0.3)
CDRO (ours)	Concepts	69.3 (2.7)	85.0 (1.0)	88.8 (0.2)	90.5 (0.1)	88.0 (0.7)	90.8 (0.3)

Table 1. Average and worst-group accuracies of models trained with CDRO and baselines on the benchmarks. CDRO outperforms approaches that do not use group information (GEORGE, ERM) and approaches that use 5% of Validation (SSA, AFR). CDRO also has comparable performance to approaches that use validation group labels (JTT, CNC, AFR, SSA).

	Precision(%)	Recall(%)
CMNIST	87.0	96.9
Waterbirds	76.3	83.9
CelebA	78.0	78.5

Table 2. We show the average precision and recall of the pseudo-group labels inferred by CDRO compared to the ground truth. We show CDRO effectively infers group labels.

data points with high similarity scores tend to be the red digits. Similarly, given a CAV in the direction of the concept *land background*, data points with high similarity scores tend to be images with land background. These results justify the decision to label all data points in the highest-mean mixture with the group associated with the concept.

Further, Table 2 shows that CDRO estimates group labels with high accuracy in all three datasets.

4.3.2 CDRO is robust to quality of concepts

From a practical perspective, since CDRO requires practitioners to curate the concepts, the quality of concepts depends on how much time and effort the practitioners spend on curating the concepts. We therefore evaluate the effectiveness of CDRO when the quality of concepts varies on Waterbirds and CelebA. To examine this, we ask the following two questions: 1) Does CDRO require that the concept images be drawn from the training distribution? 2) Does CDRO require a large number of concept images?

We first fix the size of concepts to be 100 (50 for the concept set and 50 for the contrastive set), and examine the effect of distribution distance between concepts and training data on CDRO. We vary the distance between training distribution and concept distribution by varying the prompts we use for Stable Diffusion.

We design the prompts based on how much effort a hypothetical practitioner would spend on curating the concepts. A practitioner may not have specific information about the dataset and decide to use very general prompts, potentially resulting in a low-quality concept set. For the Waterbirds dataset, we use the following prompts, varying the level of effort and domain knowledge: “A photo of a [water/land] background” as the “distant” distribution, “A photo of a [water/land]bird habitat” as the “somewhat near” distribution, and “A photo of a [water/land] background” as the “near” distribution. For the CelebA dataset, we use the following prompts: “A photo of a [female/male] person” as the “distant” distribution, “A photo of a [female/male] celebrity” as the “somewhat near” distribution, and “A photo of a [female/male] celebrity face” as the “near” distribution. We include our reasoning and a discussion regarding these prompts in the Appendix.

Table 3 shows that CDRO is effective even if the concept images are out of distribution. In both Waterbirds and CelebA, CDRO achieves similar worst group accuracy to the in-distribution set even when the concept distribution is only *somewhat near*. This result suggests that practitioners need not meticulously curate the concept sets for CDRO to be effective.

We then fix the concept distribution to be *near training distribution*, and vary the size from the list (4, 8, 40, 100, 200), splitting equally between concept set and contrastive set. For example, if size is 4, then the concept and contrastive sets each have two images. Figure 4 shows that CDRO is robust to the size of the concepts. Even when there are only four images, CDRO achieves higher worst-group accuracy than GEORGE on both Waterbirds and CelebA. Further, with 40 concept images, CDRO achieves competitive results to methods that require validation group labels, *e.g.*, CNC. A small concept set is sufficient in part since the

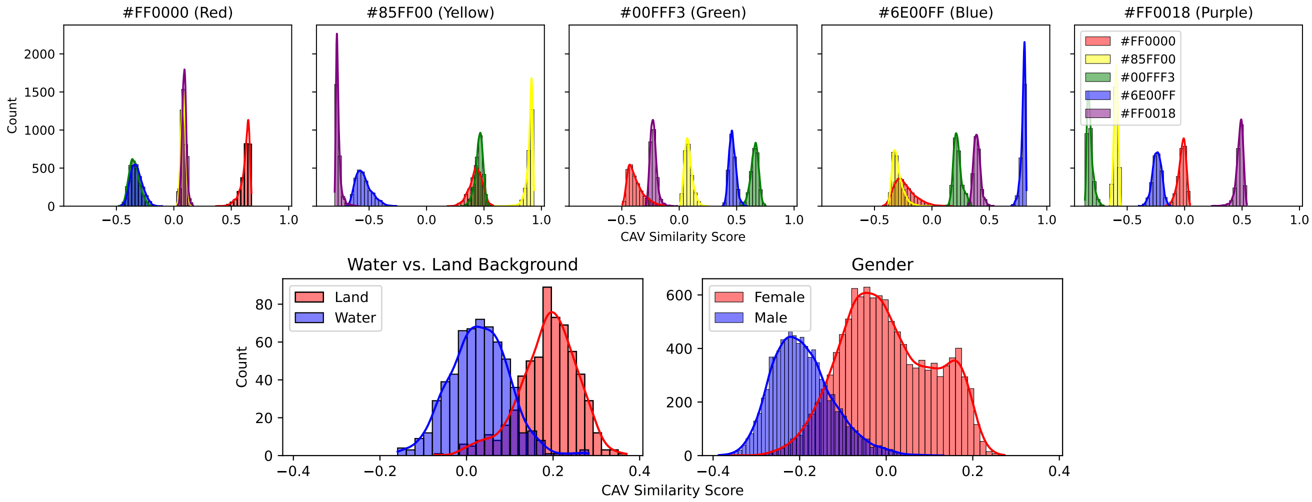

Figure 3. Distributions of similarity between the CAV trained with out-of-distribution concepts and validation samples. Each peak is colored by the ground truth group label. The peaks are separable, and the peak furthest to the right (highest similarity to the concept) corresponds to the ground truth label in all cases, indicating how we can infer spurious attribute labels. Top row: CMNIST validation samples for each of five color concepts. Bottom left: Waterbird validation samples with land as the concept set and water as the contrastive set. Bottom right: CelebA validation samples with women as the concept set and men as the contrastive set.

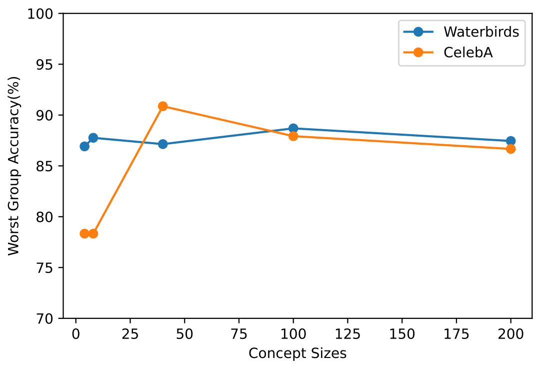

Figure 4. CDRO is robust to size of concepts. Even when there are only 4 images, CDRO achieves higher worst group accuracy than GEORGE on both Waterbirds and CelebA, since the learned embeddings from ERM make it easy to train a classifier. Further, with 40 concept images, CDRO achieves competitive results to methods that require validation group labels, *e.g.*, CNC.

	Source	Worst(%)
In Distribution	Waterbirds [17]	89.3
Near	Stable Diffusion [16]	88.5
Somewhat Near	Stable Diffusion [16]	80.1
Distant	Stable Diffusion [16]	53.1

Table 3. The worst group accuracy given different concept set qualities on Waterbirds. Concept set quality is judged by the level of perceived effort and domain knowledge needed to create.

input to the GMM is ERM-trained embeddings, such that training data features are already represented.

	Source	Worst(%)
In Distribution	CelebA [10]	87.2
Near	Stable Diffusion [16]	87.9
Somewhat Near	Stable Diffusion [16]	87.2
Distant	Stable Diffusion [16]	75.6

Table 4. The worst group accuracy given different concept set qualities on CelebA. The concept set qualities are judged similar to Waterbird concepts.

4.3.3 Concept Correction is Compatible with CNC

We demonstrate that Concept Correction is compatible with other robust training methods via CNC. Specifically, CNC consists of two stages: first, CNC trains a model with ERM and uses the predictions of the ERM model to identify majority and minority groups; then CNC uses contrastive learning to maximize the similarity of examples with same class label but different ERM predictions, and minimize the similarity examples with different class labels but same ERM predictions. We replace ERM predictions with inferred labels of spurious attributes, which we generate according to stage 1-3 of Figure 2. Table 5 shows that models trained with concept-based pseudo-labels achieve comparable worst-group accuracy to models trained with ground truth group labels on both train and validation set.

5. Related Work

We discuss related work on improving robustness to spurious correlations, both with and without group information available.

[1] Note that we ran Group DRO on CMNIST according to hyperparameters specified in [24], whereas the result reported by [24] was 78.5(4.5)

Method	Group Info	Waterbirds Worst(%)	CelebA Worst(%)
CNC	Train & Val	88.7 (0.4)	88.4 (0.1)
Concept CNC	Concepts	86.2 (0.7)	87.2 (0.1)

Table 5. Worst-group accuracy of Concept CNC compared to CNC. We show that we can use concept sets to achieve comparable performance to CNC, which uses both train and validation group labels in our implementation.

Improving Robustness with Group Information Sagawa *et al.*proposed Group Distributionally Robust Optimization (DRO) [17], which assumes access to spurious attribute labels and optimizes for worst-group loss instead of average loss. The authors show that Group DRO can achieve state-of-the-art worst-group-accuracy. To demonstrate CDRO, we use the same distributionally robust objective of Group DRO but infer group labels instead of assuming they are provided.

Several proposals to improve worst group accuracy resample the training data to balance classes or groups [4, 5, 7, 12]. These approaches show that while a neural network trained via ERM may be biased at the classification layer, it can still learn the core features for the task at the penultimate layer. Therefore it is effective to adjust the weights of the features by fine-tuning on a balanced hold-out set [5, 7], or re-weighting the loss function for majority and minority groups [4], or processing the model outputs post-hoc [12]. However, these methods acknowledge the expense of obtaining spurious attribute labels for a large dataset. Further, multiple spurious correlations can occur in practice [25], complicating reweighting schemes.

Improving Robustness without Group Information Recent proposals to improve worst-group accuracy without access to group labels typically follow a two-stage paradigm: first train a model with ERM, then use the predictions of the ERM model to identify minority groups, then train another model with emphasis on the minority groups, thereby achieving high accuracy on both the majority and minority groups [5, 9, 14, 15, 19, 20, 22, 24]. For example, JTT [9] uses the incorrect predictions of the ERM classifier as a proxy for the minority group, then trains a model with an upweight on the misclassified examples. Similarly, Correct-N-Contrast (CNC) [24] uses misclassifications as pseudo spurious attribute labels, then uses contrastive learning to align representations for samples within the same class but different spurious attribute. Recognizing that retraining a model from scratch is computationally expensive, Automatic Feature Reweighting (AFR) [15] finetunes the last layer of the ERM model on a reweighted training set. However, these methods still require a validation set with group labels for hyperparameter tuning and selecting the best model checkpoint, which turns out to be crucial for achieving their results. CDRO avoids the need of validation group labels by using concepts to infer pseudo group labels.

In [14] and [20], the authors assume access to a small number of group-labeled data and use semi-supervised learning to exploit the group information for better worst group accuracy. Specifically, Spread Spurious Attribute (SSA) [14] trains a spurious attribute predictor using both group-labeled and group-unlabeled data, then uses the predicted labels for worst-group loss minimization. Unlike SSA, CDRO does not require concepts to represent task-specific groups, which encapsulate both the spurious attribute and target class. Instead, concepts can be out-of-distribution examples solely representing the spurious attribute.

GEORGE [19] requires group labels neither during training nor validation, instead inferring group labels by clustering the feature space learned by the ERM model, then running Group DRO. During validation, GEORGE uses worst-cluster accuracy to select the best hyperparameters and model checkpoint, which avoids the need of group labels. However, while GEORGE achieves high worst group accuracy on synthetic datasets where clusters are well-defined, it is less effective on real-world datasets such as CelebA where clusters are harder to distinguish [19].

Some approaches do not follow the two-stage paradigm. Learning from Failure (LfF) [13] trains two models simultaneously, one intended to be biased and one intended to be unbiased. When training the unbiased model, LfF focuses on the examples on which the biased model tends to make mistakes thereby learning to ignore the sources of bias. Further, Taghanaki *et al.*introduced CIM [21], a contrastive learning approach that learns input-space transformation of the data to preserve task-relevant information.

6. Conclusion

In this work, we present Concept Correction, a framework that uses out-of-distribution concepts to improve worst-group accuracy in the presence of spurious correlations. We demonstrate an instance of the framework via Concept DRO, which uses concepts to infer group labels, and then uses these to train with worst-group loss minimization. We show that Concept DRO achieves competitive results to existing methods that assume access to group labels.

7. Acknowledgement

This work was supported by NIST Grant 60NANB23D194, NSF Award 1934464, Cisco Grant 70618863.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 4, 6

[2] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 4

[3] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[4] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022. 8

[5] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022. 8

[6] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2, 3, 5

[7] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 1, 2, 8

[8] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. Towards hybrid human-ai workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*, pages 2432–2442, 2020. 1

[9] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 1, 2, 3, 4, 6, 8

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 4, 6, 7

[11] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019. 1

[12] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2020. 8

[13] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 8

[14] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022. 4, 6, 8

[15] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. *arXiv preprint arXiv:2306.11074*, 2023. 4, 6, 8

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 4, 7

[17] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[18] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 1

[19] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33: 19339–19352, 2020. 1, 6, 8

[20] Nimit S Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, and Christopher Ré. Barack: Partially supervised group robustness with guarantees. *arXiv preprint arXiv:2201.00072*, 2021. 8

[21] Saeid A Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, and Anirudh Goyal. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning*, pages 10043–10053. PMLR, 2021. 8

[22] Haotian Ye, James Zou, and Linjun Zhang. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics*, pages 8968–8990. PMLR, 2023. 8

[23] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. 1

[24] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 1, 2, 3, 4, 6, 7, 8

[25] Dora Zhao, Jerone Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning*, pages 42000–42017. PMLR, 2023. 8