Decompose-and-Compose: A Compositional Approach to Mitigating Spurious Correlation

Fahimeh Hosseini Noohdani
fhosseini@ce.sharif.edu

Parsa Hosseini[*]
parsa.hosseini@sharif.edu

Aryan Yazdan Parast[*]
arian.yazdanparast@sharif.edu

Hamidreza Yaghoubi Araghi
hamidreza.yaghoubiaraghi@sharif.edu

Mahdieh Soleymani Baghshah
soleymani@sharif.edu

Sharif University of Technology
Tehran, Iran

Abstract

While standard Empirical Risk Minimization (ERM) training is proven effective for image classification on in-distribution data, it fails to perform well on out-of-distribution samples. One of the main sources of distribution shift for image classification is the compositional nature of images. Specifically, in addition to the main object or component(s) determining the label, some other image components usually exist, which may lead to the shift of input distribution between train and test environments. More importantly, these components may have spurious correlations with the label. To address this issue, we propose Decompose-and-Compose (DaC), which improves robustness to correlation shift by a compositional approach based on combining elements of images. Based on our observations, models trained with ERM usually highly attend to either the causal components or the components having a high spurious correlation with the label (especially in datapoints on which models have a high confidence). In fact, according to the amount of spurious correlation and the easiness of classification based on the causal or non-causal components, the model usually attends to one of these more (on samples with high confidence). Following this, we first try to identify the causal components of images using class activation maps of models trained with ERM. Afterwards, we intervene on images by combining them and retraining the model on the augmented data, including the counterfactual ones. This work proposes a group-balancing method by intervening on images without requiring group labels or information regarding the spurious features during training. The method has an overall better worst group accuracy compared to previous methods with the same amount of supervision on the group labels in correlation shift. Our code is available at https://github.com/fhn98/DaC.

1. Introduction

While deep neural networks are capable of superhuman performance, they still fail when faced with out-of-distribution (OOD) data [2, 4, 24]. Studies have shown that these models tend to make their predictions according to simple features that have a high correlation with the label, although these correlations are unstable across data distributions [2, 24, 29]. Relying on these spurious correlations instead of the stable ones causes the model to overfit on the training data and fail on OOD samples, for which those previous correlations do not hold.

To tackle this challenge, methods based on invariant learning focus on learning representations that can be used to make invariant predictions across different environments, to make the trained model more robust to distribution shifts [1, 2, 9, 23].

Another line of work approaches this problem by balancing minority/majority groups of data [6, 8, 25] to remove spurious correlation. Among these works, [8] proposes DFR, which retrains the last layer of a model previously trained by ERM, with group-balanced data to make it robust to spurious correlation. Nonetheless, DFR requires group labels. On the other hand, methods like [13, 17, 20] try to learn a robust model by upweighting samples that are less likely to contain spurious attributes, without access to group labels during training. Relying on the assumption that datapoints on which the model has a high loss are most probably from minority groups, most of these methods aim to place more emphasis on these samples. However, in these samples, the obscure core object may be the source of high

[*] Equal contribution.

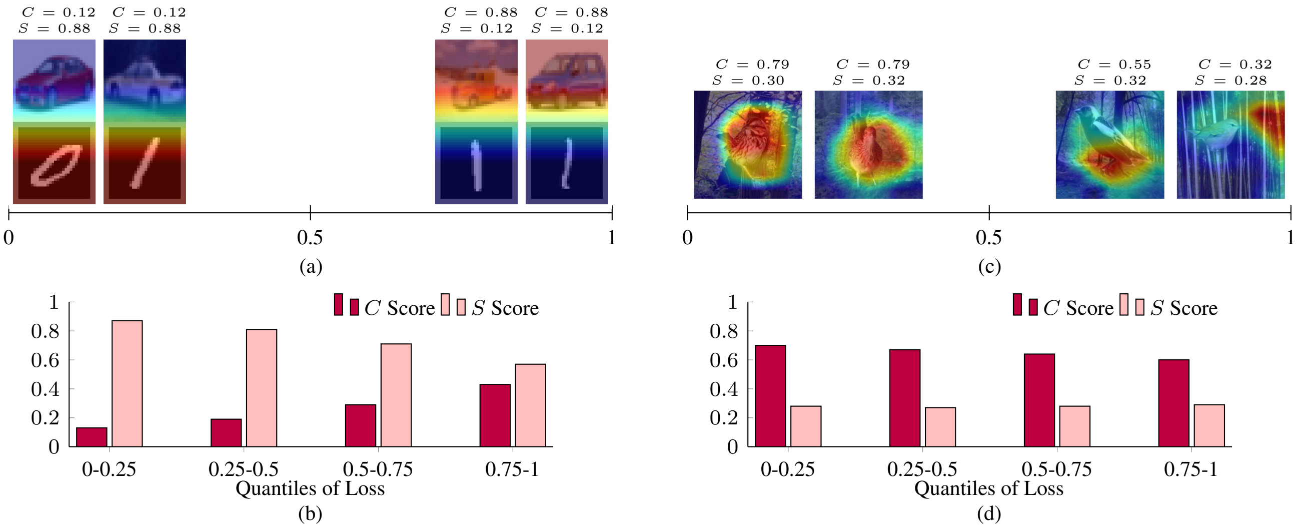

Figure 1. Behaviour of a model trained with standard ERM in different datasets. Based on the easiness of inferring the label from the causal or non-causal parts across the whole dataset, the model attends more to one of them, this behaviour is more evident in samples on which the model has a low loss. (a), (b) Average xGradCAM score of Cifar10 (causal) and MNIST (non-causal) pixels in four loss quantiles of the Dominoes training set. The model generally attends more to the *non-causal* parts, and as the loss decreases, the *non-causal* attention increases. (c), (d) Average xGradCAM score of foreground (causal) and background (non-causal) pixels in four loss quantiles of the Waterbirds training set. The model generally attends to the *causal* parts, and as the loss decreases, the *causal* attention increases.

loss (i.e., the target object itself cannot be easily classified), and overrating these samples may have some side effects. To be more precise, we need to discover and analyze parts of images to make a more accurate decision.

Due to the compositional nature of images, the problem of correlation shift can be viewed through the lens of compositionality, as models fall into the trap of spurious correlation because they make their predictions based on non-causal components of images. This fine-grained perspective could lead to a more precise approach compared to methods that consider images as a whole. While the viewpoint of compositionality is essential to OOD generalization, especially when facing correlation shift, only a limited number of works have explored this problem from this perspective. As a recent work, [36] combines parts of different images and uses them for model distillation on the representation level. Nonetheless, they cannot label the combined images, as they could not determine whether the parts taken from images for combining are causal or non-causal parts. Additionally, they do not offer any evaluations on correlation shift benchmarks. Masktune [3] takes a step further and hypothesizes that in datasets exhibiting spurious correlation, the parts of an image with high attribution scores according to a model trained with ERM, are non-causal and misleading, and based on this assumption, masks these parts for finetuning the model.

Inspired by this compositional viewpoint, we propose Decompose-and-Compose (DaC), a method for balancing groups by intervening on non-causal components of images and creating new ones. The same idea of intervening on images or using synthetic data as a means of group-balancing has been previously studied in a few works [21, 35]. However, unlike our method, which does not require any external aid during training, both these studies have a knowledge of the possible spurious attributes, and based on this knowledge, they create concept banks [35] or intervene in images using generative models [21].

In this paper, we first analyze the behaviour of a model trained with ERM and utilize its attribution map on images to decompose them into causal and non-causal parts; then, based on the performance of the model trained with our method, on the validation set, we identify the causal parts.

More precisely, as opposed to MaskTune [3], which assumes that for a given model trained with standard ERM, the regions with high attribution scores are spurious ones, we show that a model trained with ERM usually focuses on either causal or non-causal parts of images, based on the easiness of predicting the label from them. Gaining this knowledge about the causal parts enables us to intervene on datapoints from the majority groups to create new ones from under-represented groups as a means of group balancing.

The contributions of this work are as follows:

- We provide an analysis of the behaviour of models trained with standard ERM, especially on low-loss data.
- Based on this analysis, we introduce a method for identifying the causal parts of an image.

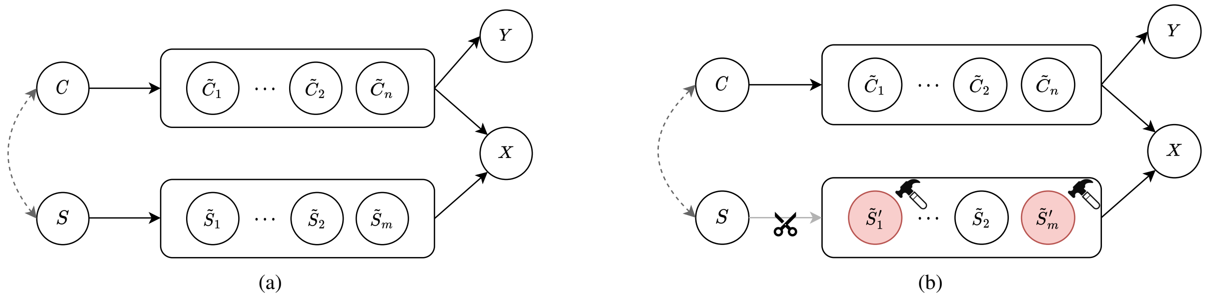

Figure 2. (a) Image as a composition of causal and non-causal components. (b) The edge between S and $\tilde{S}$ can be removed by intervention on components in $\tilde{S}$. This removes the spurious correlation between $\tilde{S}$ and Y.

- We propose a method for combining images, to create new datapoints representing the minority groups, as a means for group balancing.
- Our proposed method performs better than previous baselines on well-known benchmarks in the literature.

2. Preliminaries

Spurious correlations that are perceivable by humans can be categorized into two groups: 1) Spurious parts: spurious correlations between other objects of the image and the label (e.g., spurious correlation of the background and the label in the Waterbirds [24] dataset) and 2) Spurious attributes: spurious correlations between some non-causal attributes of the object of interest and the label (e.g., spurious correlation between the colour of the digits and their label in the CMNIST [2] dataset). In this section, we propose a method that provides robustness to the first type of spurious correlation, which is prevalent in most benchmarks. The proposed method mitigates correlation shifts by discovering non-causal parts of images and intervening on these parts.

2.1. Problem Definition

Consider a dataset $D_{tr} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ for a classification problem. Each $x \in \mathcal{X}$ has *core* features c, which are the cause of the label, and their correlation with the label is persistent across different environments. It has also *spurious* features s. Based on combinations of core and spurious features, (c, s), training samples can be partitioned into several groups. We consider the case when there is an imbalance between the size of groups in the training set. In this case, the group containing the majority of samples in a class is called the *majority* group of that class, and the others are called the *minority* groups. This imbalance induces a *spurious correlation* between the spurious features and the label, i.e. the value of the spurious features and the label corresponding to a majority group are frequently seen together in the dataset. The proportion of samples in groups could be different in the test set, causing a *correlation shift* between the training and the test sets. For instance, in the Waterbirds dataset [24], in which the task is to determine whether each image shows a waterbird or a landbird, the core and spurious features are the foreground and the background respectively. Waterbirds consists of four groups: waterbirds on water background, landbirds on land background, waterbirds on land background and landbirds on water background, with the first two being the majority groups.

Our objective is to train a classifier, denoted as f, that performs well across both the training and test distributions. This entails ensuring that f exhibits strong performance not only on the majority but also on the minority groups.

2.2. A Causality Viewpoint to Spurious Correlation

To study the problem of spurious correlation from the perspective of causality, we model the data-generating process as the Structural Causal Model (SCM) [19] shown in Fig. 2a. In this SCM, C and S indicate unobservable causal and non-causal variables, from which the observable causal and non-causal components $\tilde{C}$ and $\tilde{S}$ for an image are obtained. The final image X is the output of $\psi(\tilde{C}, \tilde{S})$, where $\psi(.,.)$ is a combining function. The label of the image is caused by $\tilde{C}$.

In the case of spurious correlation, a hidden confounder E, mostly referred to as the *environment* [2] or *group* [24] variable in the literature, would be present in the SCM such that $S \leftarrow E \rightarrow C$. This creates the path $\tilde{S} \leftarrow S \leftarrow E \rightarrow C \rightarrow \tilde{C} \rightarrow Y$, which introduces a spurious correlation between $\tilde{S}$ and Y. E is mainly sample selection bias.

Whereas most previous studies on mitigating spurious correlation did not approach this problem from a causality perspective, methods based on group balancing, such as [6, 8, 24], resolve this issue by eliminating the effect of E. While solutions based on group balancing are effective when E is observable, they are not feasible when group annotation is not provided.

Another solution that is effective even in the absence of group annotation is removing the edge $S \rightarrow \tilde{S}$ by intervening on some components in $\tilde{S}$ in order to break the path, as shown in Fig. 2b. More concretely, this solution intervenes on a subset of non-causal components of images without changing the label to reduce their correlation. Intervention on $\tilde{S}$ could be done in a more efficient manner if we could set $\tilde{S}$ to a value that has less co-occurrence with Y. Such intervention would create a new datapoint that can be assigned to a minority group. Hence, this type of intervention would be a method for upweighting datapoint from minority groups.

3. To Which Does ERM Attend More?

To intervene on the non-causal components of an image, it is essential to determine the causal and non-causal parts of it first. The attribution map of a model trained with standard ERM on an image could be utilized in distinguishing these parts. MaskTune [3] partly addressed this issue by assuming that for a model trained with standard ERM on a dataset exhibiting spurious correlations, the image parts with high attribution scores are spurious.

However, this assertion does not hold for the majority of realistic datasets. Specifically, the behaviour of a model trained with ERM might vary depending on the easiness of predicting the label from the causal and non-causal parts across different datasets.

Since non-causal parts of images that have spurious correlations with the label, such as the background in the Waterbirds dataset, are shortcuts for models, it is often presumed that a model trained with standard ERM attends more to the non-causal parts of images. However, we show that this assumption does not hold in many real-world scenarios. Across the entire dataset, as opposed to the causal components, non-causal ones do not persistently appear in accordance with the label. Consequently, the causal parts may become generally more predictive. As a result, the easiness of predicting labels from the causal and non-causal parts across the entire dataset determines the focus of ERM.

To illustrate this point, we report the average xGrad-CAM [27] score for the foreground and background pixels in the Waterbirds datasets in Fig. 1(a,b) and for the Cifar part and the MNIST part of the Dominoes dataset [18] in Fig. 1(c,d). For the Waterbirds, the average score of the foreground (causal) pixels exceeds that of the background (spurious part) and for the Dominoes, the average score of the spurious part (MNIST part) exceeds that of the causal pixels (i.e. Cifar part). Indeed, if the spurious patterns are easy to learn, the model may attend more to the spurious parts. This tendency becomes more pronounced in samples with low loss, as depicted in Fig. 1.

4. Method

Regardless of whether models trained with ERM attend more to the causal or non-causal parts of an image, their attribution map on the image remains useful for distinguishing these two parts. In the following, we introduce a method for identifying the most predictive parts of images using attribution maps.

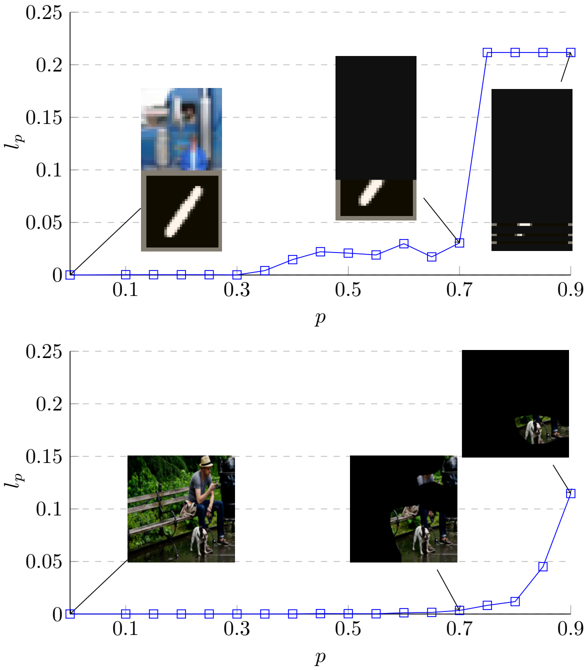

Figure 3. Adaptive masking according to the attention scores obtained from the ERM model. The loss value of the masked images for which different portions p of pixels (with the lowest attention score) has been masked is shown as l_p. (a) The loss curve for an image of the Dominoes dataset with the label 'truck' on which the ERM model has non-causal attention, and (b) The loss curve for an image of the MetaShift dataset with the label 'dog' on which the ERM model has causal attention.

4.1. Adaptive Masking by ERM

Given a model $f_\theta(.)$ that is trained with ERM, and a datapoint (x, y) on which f's loss is low, we want to find a mask that conceals pixels in x except for the most predictive ones. The assumption that the loss on (x, y) is low is necessary since we are more confident that the predictive parts in x are indeed predictive of the correct label.

First, the attribution scores of pixels of x are computed using a visual explanation method [27, 32]. Due to its efficiency, we use xGradCAM [27] for this means and denote the xGradCAM of an input x as AttributionScore(x). To have precise adaptive masking, we first define $l(f_\theta(\tilde{x}^p), y)$, in which l is the cross-entropy loss, and $\tilde{x}^p$ is the new image obtained from x by masking out the portion p of pixels with the lowest attribution scores according to AttributionScore(x). When gradually masking x, first the less predictive parts are masked out, which will not have a significant effect on the loss of the masked image. However, when a large proportion of the image, including the predictive part, is masked, the loss increases rapidly, as it is hard for the model to predict the label when the center of attention is (partially) obscured. The effects of gradually masking two images in the Metashift and Dominoes datasets are shown in Fig. 3. More precisely, $l(f_\theta(\tilde{x}^p), y)$ can be considered as a function of p whose elbow located at p^* shows the optimal amount of masking for the input x. This amount of masking is expected to conceal the non-predictive parts as much as possible while keeping the predictive parts intact. Therefore, we define a function $m : \mathbb{R}^{H \times W \times 3} \to \{0, 1\}^{H \times W}$ that returns a mask for its input through an adaptive masking. In fact, $m(x)$ provides a binary mask for x whose value is 0 for a proportion p^* of pixels of x with the lowest attribution score and is 1 for the remaining pixels (p^* denotes the optimal amount of masking found as the elbow of $l(f_\theta(\tilde{x}^p), y)$).

Based on the observation mentioned above, we propose *Adaptive Masking* algorithm shown in Algorithm 1 to find the optimal amount of masking for an image.

Algorithm 1: Adaptive Masking

```
Input: Model $f_\theta$; Image $x$; Label $y$; Loss function
       $l(.,.)$;
Output: Mask $m$;
$\delta = 0.2$
$x\_score \leftarrow$ AttributionScore($x$) ;
// Calculate $l_p$ for $p \in \{0, \delta, 2\delta, 3\delta, ..., 1\}$
for $p$ in $\{0, \delta, 2\delta, 3\delta, ..., 1\}$ do
    $\tilde{x}^p \leftarrow$ MaskWithProportion($x\_score$, $p$);
    $l_p \leftarrow l(f_\theta(\tilde{x}^p), y)$;
end
// Find the optimal mask
$p^* \leftarrow$ FindElbow($l_p$);
$m \leftarrow$ MaskWithProportion($x\_score$, $p^*$);
return $m$;
```

4.2. Decompose-and-Compose (DaC)

Viewing images as combinations of components gives us the upper hand of being able to intervene [19] on them. This could be done by combining components of different images to create novel combinations less seen by the model. To be more specific, in the case of spurious correlation, by combining components of different selected inputs, novel images that are more similar to underrepresented data could be created and used during the training of the model as a means of upweighting the underrepresented groups. Here we propose a method for obtaining new datapoints from minority groups, by combining the ones from the majority groups. Given dataset $\mathcal{D}$, we consider a pretrained classifier $f_\theta(x) = w \circ g_\phi(x)$ with standard ERM on $\mathcal{D}_{tr}$, in which g_ϕ is a feature extractor and w is a linear predictor. Afterwards, inspired by [8], we intend to retrain only w on an augmented variation of $\mathcal{D}_{tr}$ that is prepared by intervening on samples to produce new ones to upweight the underrepresented or minority groups.

Selecting low-loss examples and decomposing them. For a given training batch $\mathcal{B}$, we first select the subset $\mathcal{B}' = \{(x^{(i)}, y^{(i)}) \in \mathcal{B} | (x^{(i)}, y^{(i)})$ is among q portion of training samples with the lowest loss$\}$ and q is the hyperparameter denoting the portion of the selected samples. For each sample $(x^{(i)}, y^{(i)}) \in \mathcal{B}'$, first, by Algorithm 1, the mask $m(x^{(i)})$ will be obtained. Then, the two parts of image x is found as $x^{(i)} \odot m(x^{(i)})$ and $x^{(i)} \odot (1 - m(x^{(i)}))$. But still, we do not know which of these two parts consists more of the causal regions.

As explained in Sec. 3, based on the predictive power of the causal and non-causal parts of the images, the model may attend more to one of them . We inspect both of these assumptions and find which of them yields better results on the validation data. Therefore, we consider $causalflag$ as a hyperparameter in Algorithm 2, and when this flag indicates the non-causal assumption, masks are inverted, as shown in Lines 10-13, to obtain more causal regions.

Composing the causal and non-causal parts of two images. Given a sample $(x^{(i)}, y^{(i)}) \in \mathcal{B}'$, another sample $(x^{(j)}, y^{(j)}) \in \mathcal{B}'$ is selected randomly such that $y^{(i)} \neq y^{(j)}$. Then, we combine the two images as below:

$$\begin{aligned} \hat{x}^{(i)}_{\text{comb}_j} &= m(x^{(i)}) \odot x^{(i)} \\ &+ (1 - m(x^{(i)})) \odot (1 - m(x^{(j)}))x^{(j)} \\ &+ (1 - m(x^{(i)})) \odot m(x^{(j)})b, \end{aligned} \tag{1}$$

where b is a $1 \times 1 \times 3$ vector indicating the mean of $\mathcal{B}$ across the color channels. This formula constructs the combined image by putting the selected parts of $x^{(i)}$ and masked parts of $x^{(j)}$ that are not located on the selected parts of $x^{(i)}$ together, and filling the remaining parts of the image by the default value b. The reason for setting the remaining areas equal to b is to retain the statistics of the batch as much as possible. Finally, we define $\mathcal{M} = \{(\hat{x}^{(i)}_{\text{comb}_j}, y^{(i)}) | (x^{(i)}, y^{(i)}) \in \mathcal{B}'\}$ as the combined samples.

As will be shown in Appendix 9, most low-loss datatpoints are the ones from the majority groups. By combining the causal and non-causal parts of two majority datapoints from different labels, we make new datapoints from minority groups, thus group-balance the training data without access to group annotation. More precisely, suppose that the model attends more to the causal parts across the dataset. Consider $(x^{(i)}, y^{(i)})$ and $(x^{(j)}, y^{(j)})$ as two samples from majority groups where $x^{(i)} = \psi(\tilde{c}^{(i)}, \tilde{s}^{(i)})$ and $x^{(j)} = \psi(\tilde{c}^{(j)}, \tilde{s}^{(j)})$ and $y^{(i)} \neq y^{(j)}$. Therefore, $\tilde{s}^{(i)}$ and $\tilde{s}^{(j)}$ are in accordance with $y^{(i)}$ and $y^{(j)}$ respectively. Now, in the combined datapoint $\hat{x}^{(i)}_{\text{comb}_j} = \psi(\tilde{c}^{(i)}, \tilde{s}^{(j)})$, the non-

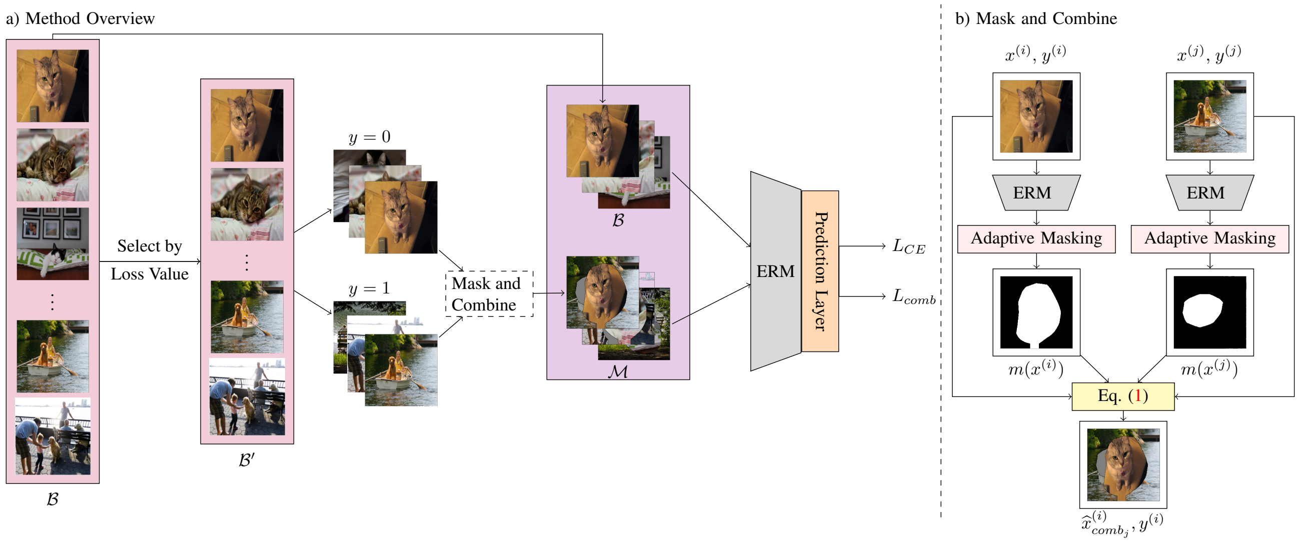

Figure 4. (a) An overview of our DaC method. For each batch, a q portion of samples with the lowest loss is selected. Then images of different labels are combined by the *Mask and Combine* module. The overall loss to update the model's last layer parameters is a weighted sum of the loss on the original batch (L_{CE}) and the combined data (L_{comb}). The algorithm for this method is shown in Algorithm 2. (b) The Mask and Combine module. The two input images $x^{(i)}$ and $x^{(j)}$ are masked by Algorithm 1. Afterwards, The selected part of $x^{(i)}$ and the masked parts of $x^{(j)}$ are combined, and the remaining gaps are filled with the mean value of the batch. The new combined image has the same label as $x^{(i)}$ and is used for training the last layer of the model.

causal part $\tilde{s}^{(j)}$ does not have the spurious value corresponding to $y^{(i)}$, which makes this datapoint from the minority groups. It is worth mentioning that this combination tends to be a minority sample independent of whether $m(x^{(i)}) \odot x^{(i)}$ or $(1 - m(x^{(i)})) \odot x^{(i)}$ reflects the cause more.

By the above intervention that combines pairs of data, we generate data in order to break the spurious correlation between the non-causal parts of images and labels. Finally, the loss function during retraining the last layer of the model is defined as below:

$$L_{CE} = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} l(f_\theta(x), y), \tag{2}$$

$$L_{\text{comb}} = \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} l(f_\theta(x), y) \tag{3}$$

$$L_{\text{total}} = L_{CE} + \alpha L_{\text{comb}}, \tag{4}$$

in which l is the cross-entropy loss and α is the hyperparameter determining the importance of the combinations. An overview of our Decompose and Compose (DaC) method in addition to its algorithm is shown in Fig. 4 and Algorithm 2 respectively.

5. Experiments

5.1. Compared Methods

In this paper, we consider 6 baselines besides ERM. *DFR* [8] argues that even in the presence of spurious correlations, neural network classifiers still learn the core features. Following this, they show that simple retraining of the last layer with group-balanced data can be sufficient to make the model robust to spurious correlation. We evaluated their method on our ResNet50 [5] backbone trained with ERM. *Group DRO* [24] aims to minimize the worst-case loss across groups with strong regularization. *LISA* [39] is a data-augmenting technique that aims to learn invariant predictors By intervening on samples with either the same labels but different domains or the same domains but different labels. *MaskTune* [3] argues that in the presence of spurious correlations, ERM models attend more to the spurious parts of images. Therefore, they fine-tune an ERM model for one epoch using a masked version of the training data to force the model to focus on the core parts. *CNC* [42] is a contrastive learning method designed to align representations of samples within the same class that have different spurious attributes, while also distinguishing between samples of dissimilar classes that share similar spurious features. *JTT* [13] first uses a model trained with ERM to detect misclassified samples. These samples are subsequently upweighted when training a new model on the dataset.

Algorithm 2: Decompose-and-Compose (DaC)

```
Input: Model $f_\theta(.) = w \circ g_\phi(.)$; Dataset $\mathcal{D}_{tr}$; Loss
       function $l(.,.)$; Hyperparameters $\alpha$, $q$, causalflag
for epoch=$1, 2, \dots K$ do
    for batch $\mathcal{B}$ in $\mathcal{D}_{tr}$ do
        $b \leftarrow mean(\mathcal{B})$
        $\mathcal{B}' \leftarrow q$ portion of samples in $\mathcal{B}$
          with the lowest loss
        $\mathcal{M} \leftarrow \{\}$
        for each image $(x, y) \in \mathcal{B}'$ do
            Pick $(x', y') \in \mathcal{B}'$ s.t. $y \neq y'$
            $m \leftarrow$ AdaptiveMasking$(f, x, y, l)$
            $m' \leftarrow$ AdaptiveMasking$(f, x', y', l)$
            if casualflag=False then
                $m \leftarrow 1 - m$
                $m' \leftarrow 1 - m'$
            end
            $\hat{x}_{comb} = m \odot x$
            $+(1 - m) \odot (1 - m')x' + (1 - m) \odot m'b$
            $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\hat{x}_{comb}, y)\}$
        end
        $L_{CE} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} l(f_\theta(x), y)$
        $L_{comb} \leftarrow \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} l(f_\theta(x), y)$
        $L_{total} \leftarrow L_{CE} + \alpha L_{comb}$
        $w \leftarrow$ UpdateWeights$(L_{total})$
    end
end
```

5.2. Setup

The experiments are done on four datasets: Waterbirds [24], CelebA [15], Metashift [12], and Dominoes [18]. The details for these datasets are in Appendix 8.

Similar to all the works mentioned in Sec. 5.1, the model we use in our experiments is ResNet-50 pre-trained on ImageNet. For ERM training, on all datasets except Dominoes, we used random crop and random horizontal flip as data augmentation, similar to [3, 8]. Retraining the last layer of the model did not require data augmentation. Also, to reduce the strong disturbance of class imbalance, we used class-balanced data to retrain the last layer on CelebA, which is the same approach we took to reproduce the results of [3]. Model selection and hyper-parameter fine-tuning are done according to the worst group accuracy on the validation set. For all the datasets, the value for α and the proportion q of the selected data (according to their loss) for combining have been chosen from $\{1, 2, \dots 10\}$, and $\{0.2, 0.4, 0.5, 0.6, 0.8, 1\}$ respectively. For adaptive masking, we used [26] python implementation to determine the optimal amount of masking.

In addition to the main method (DaC), we test another version of our method, named DaC-C, which uses all the correctly classified samples for making combined data, and removes the hyperparameter q. Thus, it uses correct classification as a way for selecting low-loss samples.

For all datasets, we have trained the model in two settings: one by assuming that the model generally attends to the causal parts, and the other by the assumption that the model trained by ERM attends more to the non-causal parts. For all datasets except the Dominoes, the former has better worst group accuracy on the validation set.

The details for training the base ERM model and training the last layer of the model with DaC are in Appendix 8.

5.3. Results

The results of our experiments along with reported results for DFR [8], Masktune [3], LISA [39], Group DRO [24], and JTT [13] on four benchmarks are illustrated in Table Tab. 1. Both the worst group accuracy and the average group accuracy as the most commonly used metrics to evaluate robustness again spurious correlation have been reported. Similar to [8], the Group Info column shows whether the label of the group (majority/minority) to which datapoints belong is available for training or validation data. Among the methods in Tab. 1, only DFR requires group info of validation data during the training phase (and not just for model selection), which is shown by ✓✓.
The results of the methods annotated with $*$ are reproduced by our own experiments. As for the other methods, the results on the Waterbirds and CelebA datasets are from their original paper. The results for Metashift are reported by [35]. Three methods among the baselines, i.e. DFR, Group DRO, and LISA need the group label during the training phase as mentioned in Tab. 1. According to these results, our method outperforms other methods that don't require group labels during training with a large margin in both mean and worst group accuracy metrics on Waterbirds, Dominoes, and Metahift datasets. Moreover, although the proposed method does not need the group label of the training data, it outperforms Group DRO and LISA on Waterbirds and Metashift datasets and is on par with DFR on these datasets. It is worth mentioning that the CelebA dataset does not match the type of spurious correlation for which our method has been designed as mentioned in Sec. 2. More precisely, in addition to the face (including gender features) that can be considered as a non-causal part for classifying the hair colour, there is also a spurious attribute for the causal part (i.e. hair) in this problem. There is a spurious correlation between the volume of the hair and the label, as will be discussed in Appendix 8.

Unlike most previous methods, which usually do not generalize well to samples with diversity shift, our method can perform well also for diversity shift if it is due to novel compositon in the scene. For more analysis of DaC-C,

Table 1. A comparison of mean and worst group accuracy of several methods, including ours, on four datasets. The Group Info column shows whether each method uses group labels of train/validation data, with ✓✓indicating that group info is used in both the training and validation phases. The mean and std are reported over 3 runs on different seeds. The bold and underlined numbers indicate the best results among all methods, and methods not requiring group annotation, respectively.

	Group Info	Waterbirds		CelebA		Metashift		Dominoes	
Method	train/val	Worst	Average	Worst	Average	Worst	Average	Worst	Average
DFR*	✗/✓✓	$92.3_{\pm 0.2}$	$93.3_{\pm 0.5}$	$88.3_{\pm 1.1}$	$91.3_{\pm 0.3}$	$72.8_{\pm 0.6}$	$77.5_{\pm 0.6}$	$\mathbf{90}_{\pm \mathbf{0.4}}$	$\mathbf{92.3}_{\pm \mathbf{0.2}}$
Group DRO	✓/✓	$91.4_{\pm 1.1}$	$93.5_{\pm 0.3}$	$88.9_{\pm 2.3}$	$92.9_{\pm 0.2}$	$66.0_{\pm 3.8}$	$73.6_{\pm 2.1}$	-	-
LISA	✓/✓	$89.2_{\pm 0.6}$	$91.8_{\pm 0.3}$	$\mathbf{89.3}_{\pm \mathbf{1.1}}$	$92.4_{\pm 0.4}$	$59.8_{\pm 2.3}$	$70.0_{\pm 0.7}$	-	-
MaskTune*	✗/✗	$86.4_{\pm 1.9}$	$93.0_{\pm 0.7}$	79.4	89.5	$66.3_{\pm 6.3}$	$73.1_{\pm 2.2}$	$65.8_{\pm 4.7}$	$85.6_{\pm 0.7}$
CnC	✗/✓	$88.5_{\pm 0.3}$	$90.9_{\pm 0.1}$	$\underline{88.8_{\pm 0.9}}$	$89.9_{\pm 0.5}$	-	-	-	-
JTT	✗/✓	86.7	93.3	81.1	88.0	$64.6_{\pm 2.3}$	$74.4_{\pm 0.6}$	-	-
Base (ERM)	✗/✗	$70.8_{\pm 0.5}$	$91.6_{\pm 0.1}$	41.7	$\underline{\mathbf{96.0}}$	$61.3_{\pm 3.4}$	$73.9_{\pm 1.5}$	$72.8_{\pm 1.6}$	$88.5_{\pm 0.3}$
DaC-C	✗/✓	$\underline{\mathbf{92.6}_{\pm \mathbf{0.2}}}$	$94.9_{\pm 0.2}$	$76.11_{\pm 0}$	$91.35_{\pm 0.2}$	$76.0_{\pm 0.8}$	$\underline{\mathbf{80.0}_{\pm \mathbf{1.4}}}$	$89.0_{\pm 0.7}$	$92.2_{\pm 0.2}$
DaC	✗/✓	$92.3_{\pm 0.4}$	$\underline{\mathbf{95.3}_{\pm \mathbf{0.4}}}$	$81.9_{\pm 0.7}$	$91.4_{\pm 1.1}$	$\underline{\mathbf{78.3}_{\pm \mathbf{1.6}}}$	$79.3_{\pm 0.1}$	$\underline{89.2_{\pm 0.1}}$	$\underline{92.2_{\pm 0.3}}$

please refer to Sec. 5.4.2.

5.4. Ablation Study

5.4.1 Effect of Combining Images

Combining images proves to be extremely effective in enhancing the worst test group accuracy, which is evident in Fig. 5. It can be seen that the value of $\alpha \geq 1$ highly reduces the reliance on spurious patterns and also the range of the proper value of α is similar in different datasets.

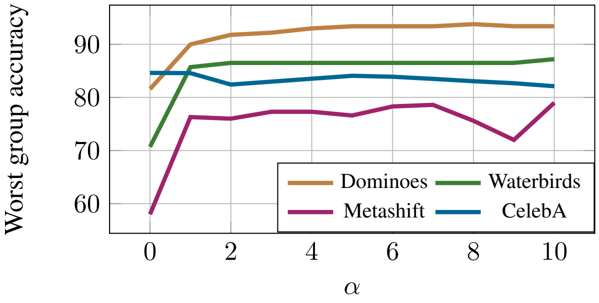

Figure 5. Worst group accuracy on different datasets with respect to α. $\alpha \geq 1$ is enough to increase worst group accuracy rapidly.

5.4.2 Effect of the Proportion of the Selected Data

As the number of the selected samples for combining increases, more datapoint on which the base model has a higher loss will be used for combining. The model is more prone to attending to irrelevant parts in these samples, which results in the combined images being wrongly labelled. As shown in Tab. 1, the model has the best worst group accuracy when we choose the proportion of the selected samples by hyper-parameter tuning. Selecting all the correctly-classified samples omit the need for tuning hyper-parameter q, but it may degrade the accuracy of the method when the ERM overfits on the training samples, since in this case, all samples may be used for combination.

6. Conclusion

In this paper, we first showed that whether the models trained with standard ERM pay more attention to the causal or spurious parts of images in a dataset depends on the predictiveness of these parts across the entire dataset. Furthermore, in most realistic datasets, due to the lower correlation of non-causal parts with the label compared to the causal ones, ERM usually shows causal attention. We then utilized attribution maps of an ERM model on images to decompose them and find the significantly more attended parts, by monitoring the classification loss of the ERM model on masked images. According to this decomposition of images, we also suggested a method for combining images with low loss which helps to mitigate the spurious correlation and diversity shift. This method has proven to be highly effective on four benchmarks and has a comparable performance with methods that require minority/majority group annotation of training data, unlike ours. Although this research was primarily focused on spurious correlations between parts of images and labels, the idea could potentially be extended to more complex scenarios where there is a spurious correlation between attributes of the objects in a scene and the label. Further research on more accurate methods for distinguishing causal and non-causal parts, and more advanced interventions on images is left for future work.

Acknowledgments

We would like to thank Mr. Mohammad-Mahdi Samiei for his insightful and constructive comments.

References

[1] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021. 1

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2020. 1, 3

[3] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 6, 7, 1, 3

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, page 472–489, Berlin, Heidelberg, 2018. Springer-Verlag. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[6] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022. 1, 3

[7] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14992–15001, 2021. 1

[8] P. Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *ArXiv*, abs/2204.02937, 2022. 1, 3, 5, 6, 7, 2

[9] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 1

[10] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Dropout disagreement: A recipe for group robustness with fewer annotations. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 1

[11] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 1

[12] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022. 7

[13] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 1, 6, 7

[14] Haozhe Liu, Wentian Zhang, Jinheng Xie, Haoqian Wu, Bing Li, Ziqi Zhang, Yuexiang Li, Yawen Huang, Bernard Ghanem, and Yefeng Zheng. Decoupled mixup for generalized visual recognition, 2022. 1

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 7, 2

[16] Nihal Murali, Aahlad Manas Puli, Ke Yu, Rajesh Ranganath, and kayhan Batmanghelich. Shortcut learning through the lens of early training dynamics, 2023. 1

[17] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1

[18] Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. Agree to disagree: Diversity through disagreement for better transferability. In *The Eleventh International Conference on Learning Representations*, 2023. 4, 7, 2

[19] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. 3, 5

[20] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. *ICML 2023*. 1

[21] Maan Qraitem, Kate Saenko, and Bryan A. Plummer. From fake to real: Pretraining on balanced synthetic images to prevent bias. *ArXiv*, abs/2308.04553, 2023. 2, 1

[22] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9301–9310. Computer Vision Foundation / IEEE, 2021. 1

[23] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 1

[24] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 1, 3, 6, 7, 2

[25] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 1

[26] Ville A. Satopaa, Jeannie R. Albrecht, David E. Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171, 2011. 7

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 4

[28] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[29] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. 1

[30] Puja Trivedi, Danai Koutra, and Jayaraman J. Thiagarajan. A closer look at model adaptation using feature distortion and simplicity bias. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011. 2

[32] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 4

[33] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3091–3100, 2021. 1

[34] Xinyue Wang, Yilin Lyu, and Liping Jing. Deep generative model for robust imbalance classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[35] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. *arXiv preprint arXiv:2305.00650*, 2023. 2, 7, 1

[36] Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust finetuning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20301–20310, 2023. 2, 1

[37] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:6502–6509, 2020. 1

[38] Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. Chroma-vae: Mitigating shortcut learning with generative classifiers. In *Advances in Neural Information Processing Systems*, pages 20351–20365. Curran Associates, Inc., 2022. 1

[39] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 25407–25437. PMLR, 2022. 6, 7, 1

[40] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7947–7958, 2022. 2

[41] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1

[42] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *Proceedings of the 39th International Conference on Machine Learning*, pages 26484–26516. PMLR, 2022. 6

[43] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5219–5227, 2017. 1

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2