ExMap: Leveraging Explainability Heatmaps for Unsupervised Group Robustness to Spurious Correlations

Rwiddhi Chakraborty, Adrian Sletten, Michael C. Kampffmeyer
Department of Physics and Technology, UiT The Arctic University of Norway
firstname[.middle initial].lastname@uit.no

Abstract

Group robustness strategies aim to mitigate learned biases in deep learning models that arise from spurious correlations present in their training datasets. However, most existing methods rely on the access to the label distribution of the groups, which is time-consuming and expensive to obtain. As a result, unsupervised group robustness strategies are sought. Based on the insight that a trained model's classification strategies can be inferred accurately based on explainability heatmaps, we introduce ExMap, an unsupervised two stage mechanism designed to enhance group robustness in traditional classifiers. ExMap utilizes a clustering module to infer pseudo-labels based on a model's explainability heatmaps, which are then used during training in lieu of actual labels. Our empirical studies validate the efficacy of ExMap - We demonstrate that it bridges the performance gap with its supervised counterparts and outperforms existing partially supervised and unsupervised methods. Additionally, ExMap can be seamlessly integrated with existing group robustness learning strategies. Finally, we demonstrate its potential in tackling the emerging issue of multiple shortcut mitigation[1].

1. Introduction

Deep neural network classifiers trained for classification tasks, have invited increased scrutiny from the research community due to their overreliance on spurious correlations present in the training data [4, 5, 9, 31, 38]. This is related to the broader aspect of Shortcut Learning [10], or the Clever Hans effect [15], where a model picks the path of least resistance to predict data, thus relying on shortcut features that are not causally linked to the label. The consequence of this phenomenon is that, although such models may demonstrate impressive mean accuracy on the test data, they may still fail on challenging subsets of the data, i.e. the groups [7, 8, 27]. As a result, group robustness is a natural objective to be met to mitigate reliance on spurious correlations. Thus, instead of evaluating models based on mean test accuracy, evaluating them on *worst group accuracy* has been the recent paradigm [12, 21, 25, 40], resulting in the emergence of group robustness techniques. By dividing a dataset into pre-determined groups of spurious correlations, classifiers are then trained to maximize the *worst group accuracy* - As a result, the spurious attribute that the model is most susceptible to is considered the shortcut of interest.

In Figure 1, we illustrate the group robustness paradigm. Given a dataset, a robustness strategy takes as input the group labels and retrains a base classifier (such as Expected Risk Minimization, i.e. ERM) to improve the worst group accuracy (G3 in this case). GroupDRO [28] was one of the early influential works that introduced the group robustness paradigm. Further, it demonstrated a strategy that could indeed improve worst group accuracy. One limitation of this approach was the reliance on group labels in the training data, which was replaced with the reliance on group labels in the validation data in successive works [13, 19]. However, while these efforts have made strides in enhancing the accuracy of trained classifiers for underperforming groups, many hinge on the assumption that the underlying groups are known apriori and that the group labels are available, which is often impractical in real-world contexts. An unsupervised approach, as illustrated in Figure 1, would ideally estimate pseudo-labels that could be inputs to any robustness strategy, leading to improved worst group robustness. An example of such a fully unsupervised worst group robustness approach is (GEORGE) [32]. GEORGE clusters the penultimate layer features in a UMAP reduced space, demonstrating impressive results on multiple datasets. In this work, we instead show that clustering *explainability heatmaps* instead, is more beneficial in improving worst group robustness. Intuitively, this stems from the fact that a pixel-attribution based explainability method in input space focuses only on the relevant image features (pixel space) in the task, discarding other entangled features irrelevant for the final prediction.

In our work, we circumvent the need for a group labeled

[1] Code available at https://github.com/rwchakra/exmap

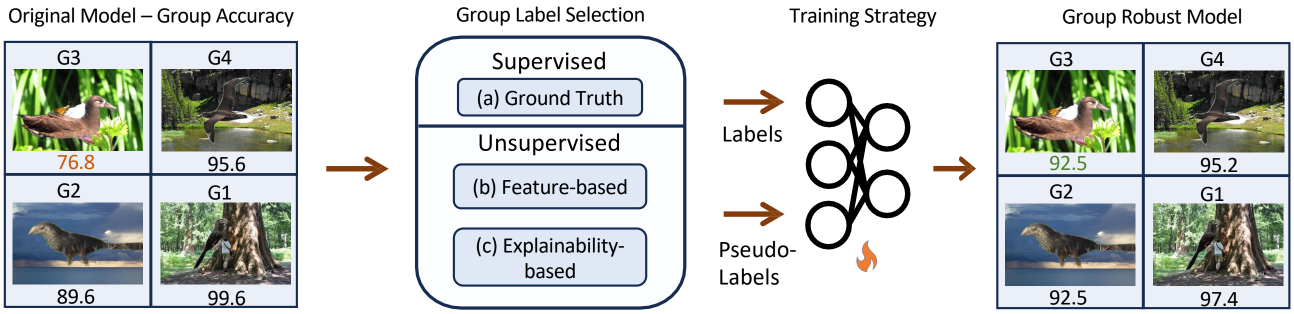

Figure 1. To improve the original models worst group accuracy, most current approaches rely on supervised group labels (a), which requires extensive annotation processes. Unsupervised approaches have relied on extracting pseudo labels based on the models feature representations (b), where information can be highly entangled. ExMap instead infers group pseudo labels based on explainability heatmaps (c), leading to improved worst group performance.

dataset by introducing ExMap, a novel two stage mechanism: First, we extract explainability heatmaps from a trained (base) model on the dataset of interest (we use the validation set *without* group labels). Next, we use a clustering module to produce pseudo-labels for the validation data. The resulting pseudo-labels can then be used for any off-the-shelf group robustness learning strategy in use today. ExMap is also flexible in the choice of clustering algorithm. We show that attaching the ExMap mechanism to baseline methods leads to improved performance over the unsupervised counterparts, and further closes the gap to supervised and partially supervised counterparts. Additionally, we demonstrate that ExMap is also useful in the recent *multiple shortcut* paradigm [18], where current popular supervised approaches have been shown to struggle. We conclude with an extended analysis on why clustering explainability heatmaps is more beneficial than raw features. In summary, our contributions include:

1. ExMap: A simple but efficient unsupervised, strategy agnostic mechanism for group robustness that leverages explainability heatmaps and clustering to generate pseudo-labels for underlying groups.
2. An extended analysis that provides intuition and insight into why clustering explainability heatmaps leads to superior results over other group-robustness baseline methods.
3. Demonstrating the usefulness of ExMap in improving worst group robustness in both the single shortcut and multiple shortcut settings.

2. Related Work

Single shortcut mitigation with group labels The paradigm of taking a frozen base model and proposing a shortcut mitigation strategy to maximise *worst group accuracy* was introduced in Group-DRO (gDRO) [28]. However, the requirement of group labels in both training and validation data motivated the proposal of mitigation strategies without training labels. This has resulted partially supervised approaches [33] that only require a small set of group labels as well as in several methods that only require the validation group labels [13, 19, 26]. One such example is DFR[13], which re-trains the final layer of a base ERM model on a balanced, reweighting dataset. Most relevant to our work, GEORGE [32] proposes an unsupervised mechanism to generate pseudo-labels for retraining by clustering raw features, and can therefore be considered the closest method to our proposed ExMap. We show that clustering heatmaps is a more beneficial and intuitive technique for generating pseudo-labels, as attributing the model performance on the input data pixels leads to a more intuitive interpretation of which features are relevant for the task, and which are not. Our method, ExMap, leverages this insight and clusters the heatmaps instead, leading to improved performance over GEORGE and its two variants - GEORGE(gDRO) trained with the Group-DRO strategy, and GEORGE(DFR), trained with the DFR strategy.

Other Strategies for Shortcut Mitigation There are other extant works that mitigate spurious correlations without adopting the group-label based paradigm directly. MaskTune [2], for instance, learns a mask over discriminatory features to reduce reliance on spurious correlations. CVar DRO [17] proposes an efficient robustness strategy using conditional value at risk (CVar). DivDis [16], on the other hand, proposes to train multiple functions on source and target data, identifying the most informative subset of labels in the target data. Discover-and-Cure (DISC) [37] discovers spurious concepts using a predefined concept bank, then intervenes on the training data to mitigate the spurious concepts, while ULA [35] uses a pretrained self-supervised model to train a classifier to detect and mitigate spurious correlations. While these approaches do not directly adopt the group-label, we show that the proposed explainability heatmap-based approach is more efficient in improving the worst-group accuracy.

Multi-Shortcut Mitigation The single shortcut setting is a simpler benchmark as the label is spuriously correlated with only a single attribute. However, real world datasets are challenging, and may contain multiple spurious attributes correlated with an object of interest. As a result, when one spurious attribute is known, mitigating the reliance on this attribute may exacerbate the reliance on another. The recently introduced Whac-A-Mole [18] dilemma for multiple shortcuts demonstrates this phenomenon with datasets containing multiple shortcuts (e.g. background and co-occurring object). Single shortcut methods fail to mitigate *both* shortcuts at once, leading to a spurious conservation principle, where if one shortcut is mitigated, the other is exacerbated. The authors introduce Last Layer Ensemble (LLE) to mitigate multiple shortcuts in their datasets, by training a separate classifier for each shortcut. However, LLE's reliance on apriori knowledge of dataset shortcuts is impractical in the real world. We evaluate ExMap in this context and show that it is effective as an unsupervised group robustness approach to the multi-shortcut setting.

Heatmap-based Explainability The challenge of attributing learned features to the decision making of a model in the image space has a rich history. The techniques explored can be differentiated on a variety of axes. LIME, SHAP, LRP [3, 11, 22] are early model-agnostic methods, while Grad-CAM and Integrated Gradients[30, 34] are gradient based attribution methods. We use LRP in this work owing to its popularity, but in principle, the heatmap extraction module can incorporate any other method widely in use today. LRP is a backward propagation based technique relying on the relevance conservation principle across each neuron in each layer. The output is a set of relevance scores that can be attributed to a pixel wise decomposition of the input image. Heatmap-based explainability techniques have also been used in conjunction with clustering, in the context of discovering model strategies for classification, and disparate areas such as differential privacy [6, 15, 29].

3. Worst Group Robustness

In this section, we provide notation and brief background of the group robustness problem. We are given a dataset D with image-label pairs being defined as $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents an image, y_i is its corresponding label, and N is the number of pairs in the dataset. The model's prediction for an image x is $y_{\text{pred}} = \hat{f}(x)$. The cross-entropy loss for true label y and predicted label y_{pred} is given by $L(y, y_{\text{pred}}) = -\sum_{c=1}^{C} y_c \log(y_{\text{pred},c})$, where C is the number of classes. Then, an ERM classifier simply minimizes the average loss over the training data:

$$\hat{f} = \arg\min_{f} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1}$$

where $\hat{f}$ is the model obtained after training. Next, given the validation data D, we assume that for the class label set $L = \{c_1, c_2, ..., c_k\}$ there exists a corresponding spurious attribute set $A = \{a_1, a_2, ..., a_m\}$, such that the group label set $G : L \times A$. For example, in CelebA, typically a : Gender (Male/Female), and c: Blonde Hair (Blonde/Not Blonde). In this case $L = \{0, 1\}$, and $A = \{0, 1\}$. Then, the optimization can be described as the worst-expected loss over the validation set, conditioned on the group labels and the spurious attributes:

$$\hat{f}^* = \arg\min_{f} \max_{(c_i, a_j) \in G} \mathbb{E}_{(x,y)\in D}[L(y, f(x))|c_i, a_j] \tag{2}$$

As discussed before, recent works aim to design strategies over the (base) trained model to minimize this objective. For example, JTT collects an error set from the training data, and then upweights misclassified examples during the second training phase. DFR reweights the features responsible for misclassifications during the first phase in its finetuning stage. Note, however, that both these methods rely on the validation set *group labels* G_{val} to finetune the network. We consider the case where $G_{train} = \phi$ and $G_{val} = \phi$. We do not have access to group labels, and must therefore infer pseudo-labels in an unsupervised manner so that existing group robustness methods can be used.

4. Leveraging Explainability Heatmaps for Group Robustness – ExMap

In this section, we describe ExMap, an intuitive and efficient approach for unsupervised group robustness to spurious correlations. ExMap is a two-stage method, illustrated in Figure 2. In the first stage, we extract explainability heatmaps for the model predictions. In the second stage, we cluster the heatmaps to generate pseudo-labels. These pseudo-labels can then be used on any off-the-shelf group robustness strategy in use today. In our work, we demonstrate the strategy agnostic nature of ExMap by running it on two popular strategies - JTT and DFR.

4.1. Explainability Heatmaps

We use LRP [3] in this stage to generate pixel attributions in the input space. This allows us to focus only on the relevant features for the task. Specifically, given the validation data $D_{val} = \{(x_i, y_i)\}_{i=1}^{M}$, we use pixel wise relevance score $r_x = (\mathcal{LRP}(x)) \forall x \in D_{val}$. Specifically, for each data point x, the relevance score is defined on a per-neuron-per layer basis. For an input neuron n_k at layer k, and an output neuron n_l at the following layer l, the relevance score R_k is intuitively a measure of how much this particular input n_k contributed to the output value n_l:

$$R_k = \sum_{l:k \longrightarrow l} \frac{z_{kl}}{z_l + \varepsilon \cdot \text{sign}(z_l)} \tag{3}$$

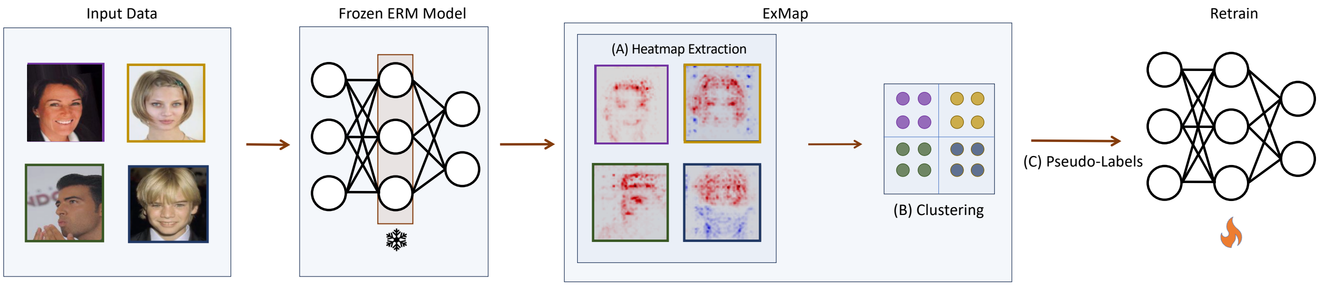

Figure 2. Our Proposed Method: ExMap facilitates group-robustness by extracting explainability heatmaps from the frozen base ERM model for the validation data (A). These heatmaps are then clustered (B) to obtain pseudo-labels for the underlying groups, which are subsequently chosen for the retraining strategy (C).

Algorithm 1 Generating Pseudo-labels using G-ExMap

```
Input: Dataset $D_{val}$, ERM Model $\mathcal{M}$, DataLoader $\mathcal{L}$
Output: Pseudo-labels $\hat{G}$
procedure GENERATEPSEUDOLABELS($D, \mathcal{M}, \mathcal{L}$)
    $R \leftarrow \varnothing$                                ▷ Initialize heatmap set
    for each batch $x$ in $\mathcal{L}$ do
        pred $\leftarrow \arg\max_i \mathcal{M}(x)_i$
        for each layer $k, l$ do
            Compute $z_l \leftarrow n_k w_{kl}$
        end for
        Compute LRP relevance $r_x$ for $x$ using Eq. 3
        Add $r_x$ to $R$
    end for
    Cluster $R$ using G-ExMap method:
    $\hat{A} \leftarrow$ Cluster$(R)$                 ▷ Estimated spurious labels
    Combine class labels $L$ with $\hat{A}$
                    $\hat{G} \leftarrow L \times \hat{A}$
    return Pseudo-labels $\hat{G}$
end procedure
```

where $z_l = n_k w_{kl}$ for the weight connection w_{kl}. R_k is computed for all the neurons at layer k, and backpropagated from the output layer to the input layer to generate pixel level relevance scores r_x for each data input x. We can thus build the heatmap set $R = \{r_x | x \in D_{val}\}$. This process is summarized in Algorithm 1.

4.2. Clustering

In the second stage, we cluster the LRP representations from the first stage. The intuition here is that over the data, the heatmaps capture the different strategies undertaken by the model for the classification task [15]. The clustering module helps identify dominant model strategies used for the classification task. By identifying such strategies and resampling in a balanced manner, ExMap guides the model to be less reliant on the dominant features across the data, i.e. the spurious features. The heatmaps serve as an effective proxy to describe model focus areas. We have two options in choosing how to cluster: Local-ExMap (L-ExMap), where we cluster heatmaps on a per-class basis, and Global-ExMap (G-ExMap), where we cluster all the heatmaps at once, and segment by class labels. We present the G-ExMap results in this paper, owing to better empirical results.

Specifically, given the Heatmap set R as described in Algorithm 1, the estimated spurious labels are generated by the global clustering method, $\hat{A} = \text{Cluster}(R)$ where Cluster (.) represents a clustering function. Now, given class label set L and estimated spurious label set $\hat{A}$, we can generate our pseudo-*group* label set $\hat{G} = L \times A$ by selecting each $a_i \in \hat{A}$, and each $c_k \in L$, to create $\{c_k, a_i\}$ $\forall k, i$.

In principle, it doesn't matter what clustering method we use, but that the clustering process itself outputs useful pseudo-labels. For our work, we leverage spectral clustering with an eigengap heuristic, inspired by SPRAY [15]. Later, we show that the choice of the clustering method does not have a significant effect on the results. The outputs, which are the pseudo group labels for the validation data D_{val}, can now be used as labels in lieu of ground truth labels to train any group robustness strategy in use. Note how in principle, *any* method that uses group labels (training or validation) would benefit from this approach. To apply this to the training set D, one would simply repeat Algorithm 1 on D. In this work, we apply ExMap to two common group robustness strategies - JTT [19] and DFR [13]. Thus, we demonstrate the strategy-agnostic nature of our approach that can be applied to any off-the-shelf method using group labels today.

5. Experiments

In this section we first present the datasets, baselines, and experimental setup. Next, we present the results and discussion[2].

[2] A discussion of the limitations and societal impact can be found in the supplementary material.

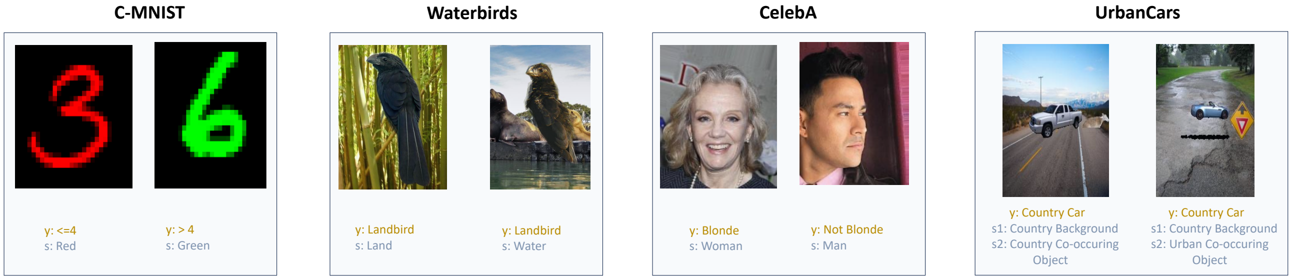

Figure 3. The datasets used in our work, visualized with respect to the class labels, and the shortcuts s. For the complete list of datasets and more details, please refer to the supplementary material.

Datasets We use CelebA [20], Waterbirds [28, 36, 39], C-MNIST [1], and Urbancars [18]. In CelebA, the class label to be predicted is hair colour (Blonde/Not Blonde), and the spurious attribute is gender (Male/Female). For Waterbirds, the class label is the bird type (waterbird/landbird), and the spurious attribute is the background (land/water). In C-MNIST, the class label is if the number is smaller than or equal to four. Any number lesser than or equal to four is assigned blue, while all numbers greater than four are assigned the color red, with a correlation of 99%. Thus, the spurious attribute is the color. For Urbancars, the class label is the car type (country/urban), and the spurious attributes are the background and co-occuring object (both country/urban). We create two variants of Urbancars: The first variant is Urbancars (BG), where only the background object is the spurious attribute. The second variant is Urbancars (CoObj), where the co-occuring object is the spurious attribute. We present single shortcut results on CelebA, Waterbirds, C-MNIST, Urbancars (BG) and Urbancars (CoObj). For the multiple shortcut setting, we use the original UrbanCars dataset with both shortcuts [18]. An overview over the considered datasets can be found in Figure 3. We present more dataset details in the supplementary.

Baselines We use the unsupervised approaches DivDis, MaskTune, and two variants of GEORGE (with gDRO and DFR) as the baselines in our work. We also adapt LfF, JTT, and CVar DRO to the unsupervised setting as additional baselines. We train the ERM model using an Imagenet-pretrained Resnet-50, and use the open source implementations of the baselines to generate our results. Specifically, we implement GEORGE(DFR), ExMap, and JTT. Remaining results are reported from [2], [19], and [16].

Setup We make sure to use the same hyperparameters from the baseline papers to reproduce the results. We utilise a composite of LRP rules to get the explainability heatmaps as recommended by [14, 23]. Following their recommendations we use LRP-ϵ for the dense layers near the output of the model with small epsilon ($\epsilon \ll 1$), followed by LRP-γ for the convolutional layers.

For the spectral clustering, we use the affinity matrix, and cluster-QR [29] to perform the clustering. The eigengap heuristic is applied to the 10 smallest eigenvalues of the Laplacian matrix to select the number of significant clusters to use. We demonstrate later that using a simpler clustering approach such as k-means can also emit reasonable results. For more details on the affinity matrix, clustering and pseudo-label generation, please see the supplementary.

5.1. Results: Single Shortcut

In Table 1 we present the single shortcut results for the datasets. First, we note that with no supervision, ExMap based DFR improves significantly upon ERM. Second, we note the improved performance of ExMap based DFR over the unsupervised baselines, including GEORGE, our closest baseline. Further, since the DFR-based GEORGE and ExMap significantly outperforms the other baselines, we present results comparing these two methods on C-MNIST, Urbancars (BG) and Urbancars (CoObj) in Table 2. In both tables, we demonstrate the superiority of clustering heatmaps to generate pseudo-labels instead of the raw features as in GEORGE. These results also show that the groups inferred by ExMap are indeed useful for worst group robustness to spurious correlations. Third, we note the gap in performance between DFR and ExMap based DFR. Since the former uses validation labels, we expect an increased accuracy, but we can report better performance on Waterbirds, and within 3% 2% 8% and 6% of the DFR results on the remaining datasets. On CelebA, our results are within 5% of Group-DRO, which demonstrates the best overall results. However, note that Group-DRO is a fully supervised approach, using labels from both the training and validation sets. For all datasets, we are able to outperform GEORGE, our closest baseline. As discussed before, while mean accuracy is not the appropriate metric to track in the group robustness setting (ERM has the best overall mean accuracy but the worst overall worst group accuracy), we can still confirm that ExMap based DFR does not witness significant drops in performance.

Methods	Group Info	Waterbirds		CelebA	
	Train/Val	WGA(%)↑	Mean(%)	WGA(%)↑	Mean(%)
Base (ERM)	✗/✗	76.8	98.1	41.1	95.9
Group DRO	✓/✓	91.4	93.5	88.9	92.9
EIIL	✓/✓	87.3	93.1	81.3	89.5
BARACK	✓/✓	89.6	94.3	83.8	92.8
CVar DRO	✗/✓	75.9	96.0	64.4	82.5
LfF	✗/✓	78.0	91.2	77.2	85.1
JTT	✗/✓	86.7	93.3	81.1	88.0
DFR	✗/✓	92.1	96.7	86.9	91.1
GEORGE (gDRO)	✗/✗	76.2	95.7	53.7	94.6
CVar DRO	✗/✗	62.0	96.0	36.1	82.5
LfF	✗/✗	44.1	91.2	24.4	85.1
JTT	✗/✗	62.5	93.3	40.6	88.0
DivDis*	✗/✗	81.0	-	55.0	-
MaskTune	✗/✗	86.4	93.0	78.0	91.3
GEORGE (DFR)	✗/✗	91.7	96.5	83.3	89.2
DFR+ExMap (ours)	✗/✗	**92.5**	96.0	**84.4**	91.8

Table 1. Worst group and mean accuracy on the test sets of the different datasets. The Group Info column showcases for each method whether group labels are used for that split of the data (✗= does not use group labels, ✓= uses group labels). We report the average results over 5 runs after hyperparameter tuning. Gray rows represent supervised approaches. *DivDis does not report mean test accuracy results.

Methods	Group Info	C-MNIST		Urbancars (BG)		Urbancars (CoObj)	
	Train/Val	WGA(%)↑	Mean(%)	WGA(%)↑	Mean(%)	WGA(%)↑	Mean(%)
Base (ERM)	✗/✗	39.6	99.3	55.6	90.2	50.8	92.7
DFR	✗/✓	74.2	93.7	77.5	81.0	84.7	88.2
GEORGE (DFR)	✗/✗	71.7	95.2	69.1	83.6	76.9	91.4
DFR+ExMap (ours)	✗/✗	**72.5**	94.9	**71.4**	93.2	**79.2**	93.2

Table 2. Worst Group accuracy and mean accuracy on C-MNIST, Urbancars (BG), and Urbancars (CoObj). We use GEORGE as the baseline, since both GEORGE and ExMap significantly outperform other unsupervised methods on Waterbirds and CelebA. Gray rows represent supervised approaches.

5.2. Results: Multiple Shortcuts

Here, we present the results on the UrbanCars data, which contains multiple shortcuts in the images - the background and the co-occurring object in the image. This dataset was introduced in the recent work on multiple shortcut mitigation [18], where the authors show that mitigating one shortcut may lead to a reliance on another shortcut in the data, rendering the single shortcut setting incomplete (the Whac-a-Mole problem). The authors introduce a new set of metrics for the task - The **BG Gap**, which is the drop in accuracy between mean and cases when only the background is uncommon, the **CoObj Gap** which is the drop in accuracy between mean and cases when only the co-occurring object is uncommon, and the **BG+CoObj Gap**, the drop when both the background and the co-occurring object are uncommon. A mitigation strategy should witness a smaller drop from the original accuracy when compared to others. In Table 3, we present the ExMap based DFR results with respect to DFR, ERM, and GEORGE(DFR). We also present results of three variants of DFR: DFR (Both), which is retraining on the original UrbanCars data with both shortcuts. DFR(BG) retrains on UrbanCars with only the background shortcut, and DFR(CoObj) retrains with only the co-occuring object shortcut. Red values indicate an increase in gap when compared to ERM, which is undesirable (the Whac-A-Mole dilemma). Note that the first three DFR methods have access to the group labels, while GEORGE and ExMap do not. Table 3 demonstrates some important results: First, that DFR + ExMap consistently posts lower drops than the base ERM model. Second, that ExMap does not witness an increase in gap on any of the metrics compared to ERM, unlike GEORGE(DFR), which

Method	BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	-8.2	-14.2	-69.0
DFR (Both)	-4.6	-5.4	-14.2
DFR (BG)	-0.3	-29.2 (× 2.06)	-33.2
DFR (CoObj)	-16.3 (× 1.99)	-0.5	-19.1
GEORGE (DFR)	-7.0	-15.4 (×1.08)	-63.4
DFR+ExMap (ours)	**-5.9**	**-9.9**	**-30.7**

Table 3. Multiple Shortcuts on UrbanCars. Red values indicate the Whac-A-Mole dilemma: Mitigating one shortcut exacerbates reliance on the other (compared to ERM). ExMap proves to be the most robust in this setting, and outperforms GEORGE, its direct unsupervised counterpart.

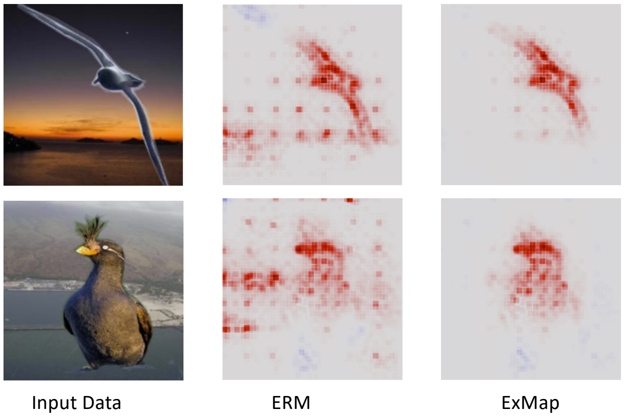

Figure 4. ERM and ExMap Heatmaps - Left: The Input images. Middle: ERM model explanations. Right: Improved group robustness using ExMap. Our method helps improve the focus on relevant attributes, in turn improving the pseudo-label estimation for retraining.

exhibits the Whac-A-Mole dilemma for the CoObj Gap. Finally, the DFR variants exhibit the Whac-A-Mole dilemma: For a DFR variant retrained on a particular shortcut, the reliance on that shortcut is mitigated (e.g. DFR (BG) mitigates the BG Gap), but the other shortcut reliance is exacerbated (DFR (BG) exhibits a higher CoObj Gap than ERM). Note that DFR uses the validation group labels, and hence will be more useful in mitigating shortcuts than our unsupervised setting. In fact, as demonstrated in [18], training separate classifiers for each shortcut is the best approach to mitigating multiple shortcuts, which explains DFR's best overall results. However, this setting assumes availability to the shortcut labels, which ExMap does not assume. Yet, it demonstrates a robust performance for the multi-shortcut setting even in the unsupervised setting, outperforming GEORGE, its closest unsupervised competitor.

6. Analysis

In this section, we present analysis and ablations along five axes: First, we demonstrate how the clustering of heatmaps is more useful than the clustering of features. Second, we demonstrate the usefulness of the ExMap representations when compared to ERM with respect to the classification task. Third, we provide more insight into what the learned clusters by ExMap capture in the data. Fourth, we demonstrate that ExMap is robust to the choice of clustering method, by performing an ablation on the clustering method using k-means instead of spectral clustering. Finally, to demonstrate that ExMap is strategy-agnostic, we use JTT as a retraining strategy using ExMap pseudo-labels, and are able to demonstrate robust performance with respect to JTT trained on true validation labels.

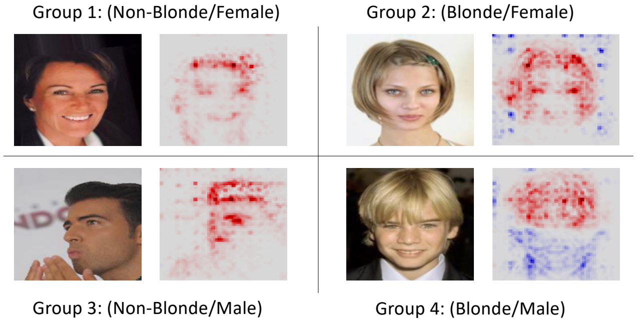

Figure 5. ExMap Heatmaps on CelebA: Each entry represents a group. The positive and negative attributions help interpret which features the model considers spurious (Blue), and which features are helpful (Red).

Methods	Mean (FG-Only %)	Mean (%)	Drop ↓
ERM	44.2	98.1	53.9
DFR	64.7	94.6	29.9
GEORGE (DFR)	73.2	96.5	23.3
DFR+ExMap (ours)	78.5	96.0	**17.5**

Table 4. Waterbirds (FG-Only). All methods exhibit a reliance on the background shortcut in Waterbirds, but ExMap posts the lowest drop, demonstrating its robustness.

6.1. The benefit of heatmaps over features

In this section we add more insight into why leveraging heatmaps for worst group robustness is more useful over features, as for example done in GEORGE. Specifically, we illustrate how heatmap based clustering mitigates reliance on the image background, the spurious attribute in the Waterbirds dataset. The results illustrate a common intuition - Explainability heatmaps highlight only the features relevant for prediction, ignoring those that are not.

Circumventing background reliance Here, we present results on Waterbirds with the spurious attribute, i.e. background, removed. We call this variant Waterbirds (FG-Only), following [13]. Please refer to the supplementary section for examples. An effective group robustness method would not witness a sharp drop in test accuracy if the model does not rely on the background. In Table 4, we present these results.

We can clearly see that the heatmap clustering strategy

Methods	Group Info	WGA(%) ↑	Mean(%) ↑
Base (ERM)	✗/✗	76.8	98.1
DFR	✗/✓	92.1	96.7
DFR+ExMap (SC)	✗/✗	92.6	96.0
DFR+ExMap (KMeans)	✗/✗	92.5	95.9

Table 5. Worst group accuracy and mean accuracy on Waterbirds with two different clustering methods - kmeans and Spectral.

mitigates the background reliance better than the feature based clustering strategy of GEORGE (lowest drop among all methods). This is also intuitive as the heatmap attributions focus on only the relevant features for prediction, discarding the rest (see Figure 4).

6.2. Qualitative Analysis

ExMap improves explanations upon retraining We visualize the heatmaps and predictions of ERM and Exmap based DFR in Figure 4. This is an image of the Waterbirds dataset that ERM misclassifies. This is reflected on the heatmap, as ERM fails to capture the relevant features. On the other hand, ExMap based DFR correctly classifies the image and focuses on the correct object region of interest (bird), instead of the spurious attribute (background).

ExMap improves Model Strategy In Figure 5, each entry represents a particular group. The positive and negative relevance scores correspond to the features that the model considers relevant and spurious respectively. ExMap uncovers the strategy used to make the prediction: In all four groups, we see ExMap helps the model uncover the hair color as a strategy. In fact, in Group 4, the model also learns that the facial features (Gender) are *negatively* associated with the prediction task (Hair Color), which is what we desire from our method. The model has learned the shortcut between man and not-blonde hair, hence ExMap uncovers the negative relevance in the face, effectively uncovering this shortcut. These examples impart a notion of interpretability to our results, as we are able to explain why the model made a particular prediction, and what shortcuts are uncovered.

6.3. Ablation Analysis

Robustness to choice of clustering method Our proposed method does not depend on any particular clustering algorithm. Although we used spectral clustering, one can also use the simpler K-means [24] to capture the clusters for pseudo-labelling. In Table 5, we present the results on Waterbirds. We are able to demonstrate that there is no significant difference in the worst group robustness performance for the clustering method we choose[3]. Both improve upon the base ERM and DFR models, and hence, both are useful.

Methods	Group Info	WGA(%) ↑	Mean(%) ↑
Base (ERM)	✗/✗	41.1	95.9
DFR	✗/✓	92.1	96.7
DFR+ExMap (ours)	✗/✗	92.6	96.0
JTT	✗/✓	86.7	93.3
JTT+ExMap (ours)	✗/✗	86.9	90.0

Table 6. Worst group and mean accuracy on Waterbirds for two different retraining strategies - JTT and DFR.

Thus, ExMap is more about demonstrating the usefulness of a heatmap clustering pseudo-labelling module rather than the specifics of the clustering method itself.

Robustness to choice of learning strategy All the results presented until now focus on the DFR backbone for shortcut mitigation. We mention previously that ExMap is strategy-agnostic, meaning that it can be applied to any off-the shelf method in use today. In Table 6, we show the results after applying ExMap to the JTT method on Waterbirds. We demonstrate similar performance to using JTT (originally uses validation labels) simply by using the pseudo labels proposed by ExMap. Additionally, we are able to improve over ERM's poor worst group accuracy as well.

7. Conclusion

The group robustness paradigm for deep learning classifiers raises important questions for when deep learning models succeed, but more importantly, *when they fail*. However, most of current research focuses on the setting where group labels are available. This assumption is impractical for real-world scenarios, where the underlying spurious correlations in the data may not be known apriori. While recent work investigating unsupervised group robustness mechanisms have shown promise, we show that further improvements are possible. In our work, we propose ExMap, where we cluster explainable heatmaps to generate pseudo-labels for the validation data. These pseudo-labels are then used on off-the-shelf group robustness learning mechanisms in use today. In addition to showing why using heatmaps over raw features is useful in this setting, our results demonstrate the efficacy of this approach on a range of benchmark datasets, in both the single and multi-shortcut settings. We are able to further close the gap to supervised counterparts, and outperform partially supervised and unsupervised baselines. Finally, ExMap opens up interesting avenues to further leverage explainability heatmaps in group robust learning.

Acknowledgements

This work was financially supported by the Research Council of Norway (RCN) FRIPRO grant no. 315029. It was further funded through the RCN Centre for Research-based Innovation funding scheme (Visual Intelligence, grant no. 309439), and Consortium Partners.

[3] Note, empirical results illustrated that k-means results were robust to the number of clusters, K, given that K was chosen sufficiently large.

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019. 5

[2] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. *Advances in Neural Information Processing Systems*, 2022. 2, 5

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10, 2015. 3

[4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. *International Conference on Machine Learning*, 2019. 1

[5] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2018. 1

[6] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *Workshop on Artificial Intelligence Safety*, 2019. 3

[7] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664, 2023. 1

[8] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021. 1

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2018. 1

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

[11] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26:982–993, 2017. 3

[12] 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. 1

[13] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *International Conference on Learning Representations*, 2022. 1, 2, 4, 7

[14] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. *International Joint Conference on Neural Networks*, pages 1–7, 2020. 5

[15] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10, 2019. 1, 3, 4

[16] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Out-of-distribution robustness via disagreement. *International Conference on Learning Representations*, 2023. 2, 5

[17] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020. 2

[18] Zhiheng Li, I. Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Cantón Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023. 2, 3, 5, 6, 7

[19] 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. *International Conference on Machine Learning*, pages 6781–6792, 2021. 1, 2, 4, 5

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision*, 2015. 5

[21] Vishnu Suresh Lokhande, Kihyuk Sohn, Jinsung Yoon, Madeleine Udell, Chen-Yu Lee, and Tomas Pfister. Towards group robustness in the presence of partial group labels. *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. 1

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 3

[23] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019. 5

[24] Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. *Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 63–67, 2010. 8

[25] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *International Conference on Learning Representations*, 2021. 1

[26] Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Neural Information Processing Systems*, 2020. 2

[27] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4227–4237, 2019. 1

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

[29] Lukas Schulth, Christian Berghoff, and Matthias Neu. Detecting backdoor poisoning attacks on deep neural networks by heatmap clustering. *ArXiv*, abs/2204.12848, 2022. 3, 5

[30] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016. 3

[31] Sahil Singla and Soheil Feizi. Causal imagenet: How to discover spurious features in deep learning? *CoRR*, abs/2110.04301, 2021. 1

[32] 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, 2

[33] Nimit Sharad Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, and Christopher Re. Barack: Partially supervised group robustness with guarantees. *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. 2

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 2017. 3

[35] Christos Tsirigotis, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron C Courville. Group robust classification without any group information. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd-birds-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[37] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. *International Conference on Machine Learning*, 2023. 2

[38] 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. *International Conference on Machine Learning*, 162:26484–26516, 2022. 1

[39] Bolei Zhou, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Journal of Vision*, 17(10):296–296, 2017. 5

[40] Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. *International Conference on Machine Learning*, pages 12857–12867, 2021. 1