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

Supplementary Material

In this supplementary material, we present additional details about the following:

- The datasets used - C-MNIST, Waterbirds, CelebA, Urbancars, Urbancars single shortcut variants, Waterbirds (FG-Only).
- Experimental Setup - The details on the heatmap extraction and clustering phase in ExMap.
- Additional results providing further intuition on how ExMap captures underlying group information.
- More results on the robustness of our method with standard errors.
- The connection between group robustness and fair clustering.
- Limitations and Societal Impact.

1. Datasets

We present the number of examples from each group for all the datasets, and the process of generating them. For C-MNIST, we used the same setup as in [2]. For Waterbirds and CelebA, we use the same setup as in [6, 8]. For Urbancars we use the same setup as in [7].

1.1. C-MNIST

We create a dataset where we have control of the number of elements in each group and what the spurious attribute is.

The Colored-MNIST dataset is a synthetic dataset based on the well-known MNIST. The MNIST dataset is a collection of several thousands of examples of handwritten digits (0-9). The images are single-channelled (black and white) and have a size of 28x28 pixels, and are accompanied by a label giving the ground truth.

We use the original data split, 60000 train and 10000 test. Since the original dataset does not have a validation set, we use the last 10000 images of the training set as the validation set.

We convert the dataset into a 2 class problem by modifying the task. This is done by simply going over to classify the numbers as smaller or equal to 4 ($y = 0$: value $<= 4$), and larger than 4 ($y = 1$: value > 4). To create the spurious attributes we make use of colors. Red is used as the first spurious attribute ($s = 0$: RGB $= (255, 0, 0)$), and green is used as the second spurious attribute ($s = 1$: RGB $= (0, 255, 0)$). Naturally, the images will need to be made 3-channeled to account for this change.

As we are interested in combating spurious correlations we create the dataset in a way such that there are correlations between the classes and spurious attributes. We use 99% correlation. That means that 99% of images from one class will have the same colour, while the remaining 1% will have the other colour. The amount of correlation was deliberately chosen so that ERM worst group accuracy is low. Table 1 shows the number of images in each group for each split.

Split	Total Data	Groups			
		Group 0 (y=0, s=0)	Group 1 (y=0, s=1)	Group 2 (y=1, s=0)	Group 3 (y=1, s=1)
Train	50,000	254	25,284	24,231	231
Val	10,000	45	5,013	4,893	49
Test	10,000	48	5,091	4,815	46

Table 1. Data splits in the Colored-MNIST dataset.

1.2. Waterbirds

Waterbirds [10] is a synthetic dataset created with the purpose of testing a model's reliance on background. The dataset consists of RGB images depicting different types of birds on different types of backgrounds. The different types of birds are divided into 2 classes, landbirds ($y = 0$) and waterbirds ($y = 1$). The different backgrounds are also divided into 2 and represent the spurious attributes of this dataset: land background ($s = 0$) and water background ($s = 1$). The group distributions across the different splits are presented in Table 2.

The Waterbirds dataset is created by using 2 other datasets, the Caltech-UCSD Birds-200-2011 (CUB) dataset [13] and the Places dataset [14]. The CUB dataset contains images of birds labelled by species and their segmentation masks. To construct the Waterbirds dataset the labels in the CUB dataset are split into 2 groups, where waterbirds are made up of seabirds (albatross, auklet, cormorant, frigatebird, fulmar, gull, jaeger, kittiwake, pelican, puffin, or tern) and waterfowls (gadwall, grebe, mallard, merganser, guillemot, or Pacific loon), while the remaining classes are labelled as landbirds. The birds are cropped using the pixel-level segmentation masks and pasted onto a water background (categories: ocean or natural lake) or land background (categories: bamboo forest or broadleaf forest) from the Places dataset.

The official train-test split of the CUB dataset is used, and 20% of the training set is used to create the validation set. The group distribution for the training set is such that

Split	Total Data	Groups			
		Group 0 (y=0, s=0)	Group 1 (y=0, s=1)	Group 2 (y=1, s=0)	Group 3 (y=1, s=1)
Train	4,795	3,498	184	56	1,057
Val	1,199	467	466	133	133
Test	5,794	2,255	2,255	642	642

Table 2. Data splits in the Waterbirds dataset.

most images (95%) depict bird types with corresponding backgrounds, to represent a distribution that may arise from real-world data. This distribution turns the background into a spurious feature. Take note that there is a distribution shift from the training split to the validation and test splits which are both more balanced, and include many more elements for the minority group. The creators of the dataset argue that they do this to more accurately gauge the performance of the minority groups, something that might be difficult if there are too few examples. They also do this to allow for easier hyperparameter tuning.

1.3. Celeb-A

CelebA here is a reference to a part of the CelebA celebrity face dataset [9] that was introduced by [10] as a group robustness dataset. From the original dataset, the feature *Blond_Hair* is used as the class, meaning that the images are divided into people who are not blonde ($y = 0$) and blonde ($y = 1$). Meanwhile, as a spurious attribute, we use the feature *Male* from the original dataset, which divides into female ($s = 0$) and male ($s = 1$). The official train-val-test split of the CelebA dataset is used. Note in Table 3, that the splits are likely randomly created, which results in equally group-distributed splits. Across all splits the group (blonde, male) is the smallest.

This dataset tests for model reliance on strongly correlated features in a real-world dataset. Observe in Table 3 that $g_3 = (y = 1, s = 1)$ which represents blonde males is severely underrepresented compared to the other groups, hence we expect the model to learn gender as a spurious feature for the class blonde.

Split	Total Data	Group 0 (y=0, s=0)	Group 1 (y=0, s=1)	Group 2 (y=1, s=0)	Group 3 (y=1, s=1)
Train	162,770	71,629	66,874	22,880	1,387
Val	19,867	8,535	8,276	2,874	182
Test	19,962	9,767	7,535	2,480	180

Table 3. Data splits in the CelebA dataset.

1.4. Urbancars

We use Urbancars, as proposed by [7]. There are 4000 images per target class, i.e. 8000 images in total. The target class is the car type (country/urban), while the two shortcuts are the background type (country/urban), and co-occurring object (country/urban). For the exact list of the cars, objects, and background, please see [7].

1.5. Urbancars single shortcut variants

The original Urbancars data has eight group combinations due to two classes, and two shortcuts (Background and Co-Occurring object). For the single shortcut variants, we merge the 4 extra groups for one particular shortcut, to leave 4 groups for the other. For example: To create Urbancars (BG), we merge the 4 groups from the other shortcut (CoObj), to create four groups containing the single shortcut of background for each of the two classes. A similar procedure is adopted to create Urbancars (CoObj).

1.6. Waterbirds (FG-Only)

This dataset is created to evaluate how well the trained models circumvent background reliance on the Waterbirds dataset, since background is the shortcut in the data. We remove the backgrounds in all the images only on the test set. In Figure 1, we present some examples.

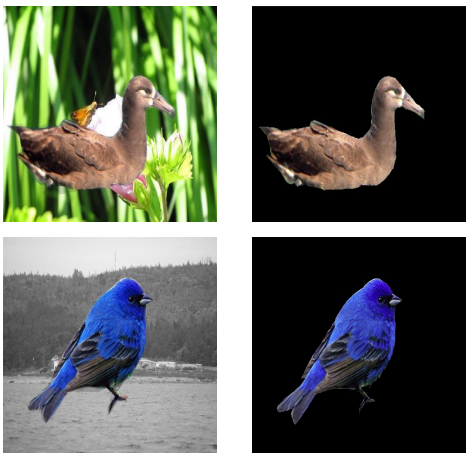

Figure 1. Waterbirds (FG Only)

2. Experimental Setup

In this section, we present more details on the heatmap extraction phase, clustering choice, and the hyperparameters used.

2.1. Heatmap Extraction

Following SPRAY [3], which reports good results across different downsizing of heatmaps, we sweep predominantly

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
GEORGE (DFR)	✗/✗	71.7±0.1	95.2±0.3	69.1±0.9	83.6±1.0	76.9±0.9	91.4±1.0
DFR+ExMap (ours)	✗/✗	**72.5**±0.2	94.9±0.3	**71.4**±0.8	93.2±0.2	**79.2**±0.7	93.2±0.3

Table 4. Group/mean test accuracy with std. Results over 5 runs.

Figure 2. ExMap based misclassifications on challenging examples (Waterbirds): In each of these images, the object of interest (bird) is co-habited by dominant peripheral objects such as humans and other birds. These situations are challenging for the classifier to discern the relevant object from the irrelevant ones.

over the following downsizings: [224, 112, 100, 56, 28, 14, 7, 5, 3]. The downsizing of heatmaps additionally helps in speeding up the clustering process and mitigating potential out-of-memory issues.

2.2. Clustering choice

Since ExMap if flexible to the choice of clustering algorithm, we experiment with spectral clustering, UMap reduced KMeans [6], and KMeans. We use the eigengap heuristic with spectral to automatically choose the number of clusters, and sweep over different cluster sizes for the KMeans based methods. We look for the largest gap in among the first 10 eigenvalues. Otherwise we test for 2-15 clusters for kmeans (overclustering as practiced in [12]).

2.3. Hyperparameters

We use the same hyperparameters for DFR and JTT as in the original papers [6, 8].

For DFR, we perform the following steps:

- Given (pseudo)group labels we create a retraining set by subsampling each group to the size of the smallest group. These are then used to retrain the last layer. After being passed through the feature extractor, each sample is normalised based on the data used to retrain the last layer.
- Similar to [6], we use logistic regression with L1-loss.
- The strength of L1 is swept over [1.0, 0.7, 0.3, 0.1, 0.07, 0.03, 0.01]. The sweep is performed by randomly splitting the retraining dataset in 2, and performing retraining with one half and evaluating the performance with the other. This is performed 5 times with different splits and the best strength is chosen based on highest worst (pseudo)group accuracy.

Method	Accuracy (%)	
	Waterbirds WGA / Mean	**CelebA** WGA / Mean
Base (ERM)	76.8 / 98.1	41.1 / 95.9
GEORGE (DFR)	91.7 ± 0.2 / 96.5 ± 0.1	83.3 ± 0.2 / 89.2 ± 0.2
DFR+ExMap	**92.5** ± **0.1** / 96.0 ± 0.3	**84.4** ± **0.5** / 91.8 ± 0.2

Table 5. Group / mean test accuracy with std. Results over 5 runs.

- When L1 strength has been selected, we retrain using the whole retrain set. This is performed 20 times with different subsamplings. The weights from each subsampling are averaged (this is viable according to DFR authors) to yield the final last layer weights. The normalisation of data is also averaged across the 20 runs.

For the ERM model, we perform the following steps:

- We use Resnet-18 for CMNIST, Resnet-50 for the others. We start with imagenet-pretrained Resnet-50 similar to previous work as it was observed to perform better. For all settings we replace the final fully connected layer to reflect the nature of our problems, i.e. 2 classes.
- Learning rate: 3e-3, weight decay: 1e-4, cosine learning rate scheduler.
- Batch size:We use batch size of 32 for Waterbirds and Urbancars, 100 for CelebA, and 128 for C-Mnist.
- Epochs: We train for 100 epochs on Waterbirds and Urbancars, 20 for CelebA, and 10 for C-Mnist.
- We use early stopping using the best mean (weighted) validation accuracy

For GEORGE, we perform the following steps:

- Acquire feature extractor (base ERM) outputs.
- Max normalise features.
- Cluster features as exmap or using UMAP+kmeans. We use 2 dimensions for the UMAP reduction, and high number of clusters (overclustering regime following [12]).

3. Capturing of Group Information

In addition to why ExMap representations are better for downstream group robustness over raw classifier features, we are also interested in what kind of group information the ExMap representations capture. The advantage of heatmaps are that they capture only the relevant features, while previous approaches that cluster in the feature space are prone to be effected by features that are irrelevant for the final prediction. To further substantiate our findings, we generate additional results to demonstrate that ExMap in-

Methods	Group Info	Waterbirds		CelebA	
	Train/Val	WGA(%)↑	Mean(%)	WGA(%)↑	Mean(%)
Base (ERM)	✗/✗	76.8	98.1	41.1	95.9
BPA	✗/✗	71.3	87.1	83.3	90.1
DFR+ExMap (ours)	✗/✗	**92.5**	96.0	**84.4**	91.8

Table 6. Comparison with Fair Clustering: Worst group and mean accuracy on Waterbirds and CelebA.

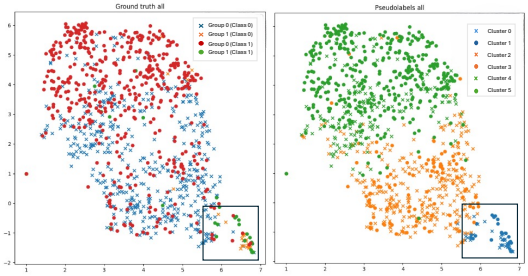

Figure 3. Groups in UrbanCars: (Left) Ground truth group labels per class. We observe minority groups (the spurious correlations) in the highlighted bottom right corner. (Right) Pseudo-labels learned by ExMap based clustering reveals a similar overall structure, conserving the dominant groups (green and yellow), while capturing the minority groups (blue cross and circle) as well.

deed captures the underlying group information. In Figure 3, we plot the pseudo-labels for UrbanCars (CoObj) after ExMap based clustering. ExMap captures both the dominant groups and the minority groups in the dataset, as indicated by the pseudo-labels learned. We also note that ExMap does not necessarily learn the same number of groups as in the ground truth data, since this information is assumed unavailable. The key observation from this figure is that ExMap is successful in identifying the dominant and minority group structure in the data. The group robust learner (such as DFR) can then sample across these groups in a balanced manner while retraining, leading to mitigation against spurious correlations.

4. Robustness Analysis

Our results in Table 1 and Table 2 in the main text are presented as the average of five runs. To illustrate the robustness of the compared approaches, we further provide the standard deviation for the ExMap and the main competitor in Table 4 and Table 5. We observe that results are robust across runs.

5. Connections to Fair Clustering

Given the close relationship between group robustness and the domain of fair clustering [1, 4, 5], we briefly comment on their connection and the potential of the insights of ExMap in the fair clustering setting. The domains of fair clustering and group robustness differ slightly, with the former aiming to improve mean accuracy independent of sensitive attributes, while the latter aim to maximize worst group accuracy. Therefore, there is a natural connection between these two research areas. Sensitive attributes in fair clustering can be regarded as a special type of spurious correlation, causally unrelated to the task. Recent work in fair clustering has therefore adopted some of the insights from the field of group robustness [11]. However, these approaches adopt a GEORGE inspired approach (cluster in raw features space), which we demonstrate to be sub-optimal in the context of group robustness. While an in-depth exploration of this is out-of-scope for this work, it could present an interesting avenue of future work. In Table 6, we present the ExMap results on Waterbirds and CelebA with respect to the method introduced in [11].

6. Limitations and Societal Impact

There are certain intuitive failure cases where the ExMap approach is not as efficient. This occurs when the images themselves are quite challenging to discern the objects of interest (the class), from other peripheral objects in the scene. In Figure 2, we present some examples of misclassifications by ExMap based DFR. In these images, we can see that the object of interest (bird), is co-habited by other dominant objects in the scene, such as humans and other birds. This creates an exceptionally challenging task for the classifier to discern the relevant features for the task. We recognise the need for robustness across challenging examples in datasets as motivation for future work. With regard to social impact, we recognise that model robustness to spurious correlations is an important first step in ensuring fair, transparent, and reliable AI that can be deployed in safety critical domains in the real world. Elucidating why models classify as they do, and specific failure cases uncovers shortcomings in exclusively choosing mean test accuracy as a metric. As a result, probing models for their weaknesses is as important as exemplifying their strengths.

References

[1] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Clustering without over-representation. *International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019. 4

[2] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and

David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019. 1

[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. 2

[4] Ioana O Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*, 2019. 4

[5] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021. 4

[6] 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, 3

[7] 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. 1, 2

[8] 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, 3

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision*, 2015. 2

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

[11] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16742–16751, 2022. 4

[12] 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. 3

[13] 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. 1

[14] 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. 1