Mitigating Spurious Correlation in Images by Intervention

Fahimeh HosseiniNoohdani
fhosseini@ce.sharif.edu

Mohammad-Mahdi Samiei
mohmahsamiei@gmail.com

Parsa Hosseini
parsa.hosseini@sharif.edu

Mahdieh Soleymani Baghshah
soleymani@sharif.edu

Sharif University of Technology
Tehran, Iran

Abstract

Though proven to be strongly effective on in-distribution data for image classification, standard ERM training fails when faced with out-of-distribution samples. One case of such failure is when there is a high spurious correlation between some features of the data and the property of interest. Deep Feature Reweighting (DFR) was proposed to face this challenge by retraining the last layer of a model on a group-balanced subset of the data to reduce the dependence of models on spurious features and further enhance their attention to the already learnt core ones. Although feature reweighting can alleviate relying on spurious features, the last layer features may not include pure core features. We propose a method based on interventions on input images to lessen the fake correlation between the spurious sections and the labels while retraining the last layer of a model. Based on our observation that models trained with ERM still highly attend to the core, causal part of images, we first generate masks for images using class activation maps. Afterwards, we make two types of interventions on images by masking or combining them, and retrain the last layer of the model on the augmented data. Along with its high interpretability, this method only needs data group labels for the model selection phase and has an overall better worst group accuracy compared to previous methods with the same amount of supervision on the group labels.

1. Introduction

While deep neural networks are capable of superhuman performance, they still fail when faced with out-of-distribution test data [3, 17, 1]. Studies have shown that these models tend to make their predictions according to simple features with high correlation with the label, even if these correlations are unstable across data distributions [20, 17]. Relying on these spurious correlations instead of the stable ones causes the model to overfit on the training data and fail on out-of-distribution samples, for which those previous correlations don't hold. As an example, [3] shows that a model trained for classifying cows and camels is prone to use the background as a shortcut for prediction, instead of the foreground object. For example, when a camel is unexpectedly placed on a a green field, the model may make a mistake.

Most 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, 6, 16]. Another line of work approaches the problem of spurious correlation by balancing minority/majority groups of data in different classes, to remove spurious correlations. Among these works, [5] proposes DFR, which retrains the last layer of the models with group-balanced data to make models robust to spurious correlations. Nonetheless, during retraining, DFR needs to know the minority and majority groups of data from each class, and thus, requires group labels. On the other hand, methods like [10] try to learn a robust model without access to group labels during training.

This study aims to mitigate spurious correlations by intervening on the input images during the retraining of the last layer, without the knowledge of data groups in the training phase. [5], observes that ResNet-50 [4] trained with standard empirical risk minimization (ERM) on biased datasets is still able to extract the core features of data, and states that retraining only the last layer of the model on a balanced subset of the data is sufficient to make it robust to spurious correlations. However, due to the nature of CNN architecture, features extracted from segments of an image are highly influenced by the area around them. Hence, even by last layer retraining, the model may still fall under the influence of spurious correlations between parts of the image

Table 1: Average xGradCAM scores for the foreground (FGS) and the background (BGS) in the Waterbirds training data. The columns Mask/FG and Mask/BG respectively indicate the percentage of the foreground and the background pixels covered by the mask made by selecting the 15000 pixels with the highest score according to xGradCAM.

Group	FGS	BGS	Mask/FG	Mask/BG
land-land	0.67	0.26	80.7%	20.3%
land-water	0.64	0.14	85.6%	19.7%
water-land	0.66	0.26	75.8%	19.3%
water-water	0.62	0.37	64.6%	23%

and the label. To reduce the impact of the spurious parts of the image on the core ones, we make interventions [14] on the spurious segments in individual images, in order to reduce the spurious correlations. We propose two intervention types, one by removing the spurious regions, and another by switching spurious parts of two different images. After making these interventions, we can retrain the last layer of the model on the augmented data with standard ERM and still achieve results comparable with DFR without requiring group label information during training.

When a model trained with ERM is faced with data exhibiting spurious correlations, it will make predictions based on the spurious features and thus fail on out-of-distribution samples. In an initial attempt to explain this failure, one might assume that the model has been focusing on spurious features rather than the core ones. Nevertheless, as observed in our experiments, even a model trained with ERM has learned these core features in many data points. Inspired by this observation, we make masks for images by selecting pixels with the highest scores based on xGradCAM [18]. Whereas this method has similarities to MaskTune [2], there is a key difference between the two methods. MaskTune, with the idea that models trained with ERM pay attention to spurious pixels, assumes that the pixels with the highest xGradCAM scores are spurious. On the other hand, we claim that the high-scoring pixels are the core ones, and take a step further than masktune to make combinations of various images in addition to simply masking the data.

2. Problem Definition

Given access to dataset $D = \{(X^{(i)}, y^{(i)})\}_{i=1}^{N}$ exhibiting spurious correlations, the goal is to train a classifier M that can perform equally well on data from both training and new distributions. It is assumed that the difference between distributions is mostly reliant on a shift of spurious correlations. To address the problem more specifically, each image X in D consists of two parts: C which is the core (or causal) part of X, and S which is part of X that has a spurious correlation with Y. The problem arises when a model trained with standard ERM relies on S to predict Y and fails to generalize to datasets in which the correlation between S and Y does not necessarily hold. If the correlation between S and Y in the training set can be reduced, models even trained by ERM are more likely to rely on the causal features instead of the spurious ones.

3. Method

3.1. Motivation

DFR [5] states that models trained with ERM are capable of extracting both the core and spurious features of an image, and reweighting the coefficients of features in the last layer of the predictor is sufficient for robustness to spurious correlations. DFR has proven effective across various benchmarks; however, a downfall of this approach is its assumption that the group label of each validation data point is available.

Methods have been proposed to robustify models even in the absence of group annotations [10, 13, 8, 27, 7] among which some are extensions of DFR [7]. For a more detailed overview of the related work refer to the Appendix.

A more interpretable approach to retraining the last layer is to detect parts of the image that have a spurious correlation with the label, perturb them to generate new augmentations of the original data, and retrain only the last layer of the base model which is trained with ERM. Moreover, these augmentations are solely based on masking the training data and combining them and do not include any predefined augmentations based on any knowledge of the spurious features. The simplicity of augmentation along with only retraining the last layer, makes our method efficient in addition to high intrinsic interpretability.

MaskTune [2] hypothesizes that models trained with ERM mostly focus on parts of the image with a high spurious correlation to the label. However, as shown in our experiments, this hypothesis does not hold unless the spurious correlation is so high that the model is encouraged to mainly focus on the non-core parts of the image. To be more precise, we assume that models trained with ERM mostly attend to the core parts of the image. Table 1 illustrates that a model trained with ERM mostly focuses on the core parts of the images (i.e. birds). This is also proven in our experiments on CelebA, in which our proposed methods have shown an acceptable performance.

3.2. Mask and Combine

Given dataset D, we first train a classifier $M = f \circ g$, in which f is a feature extractor and g is a linear predictor, with standard ERM on D. Afterwards, inspired by [5], we retrain only g on an augmented variation of D. For $(X^{(i)}, y^{(i)}) \in B$, in which B is the training batch containing $(X^{(i)}, y^{(i)})$, first, the attention scores of pixels of $X^{(i)}$ are computed

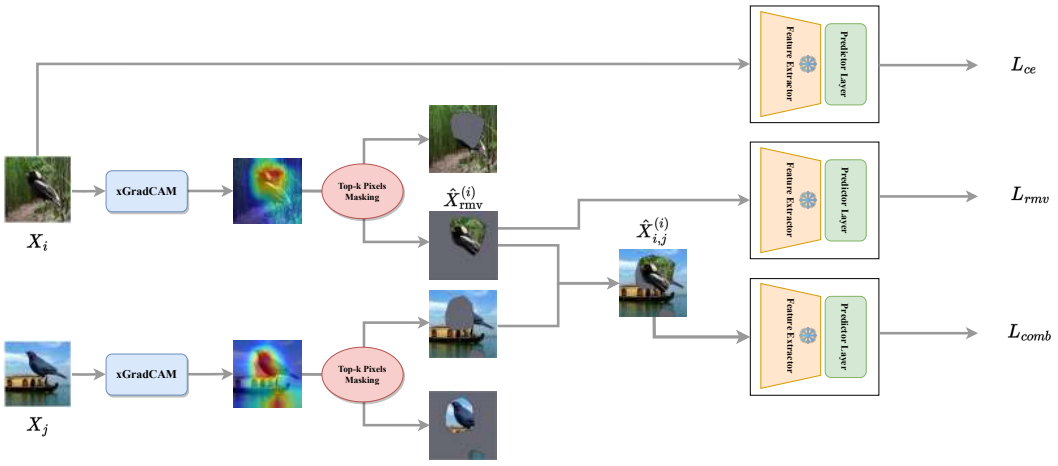

Figure 1: Method overview. Given a model trained with ERM on the dataset, we first make masks for the images based on the xGradCAM score. Afterwards, masked and combined augmentation for $X^{(i)}$ are constructed using the masks. Finally, the model is updated according to the classification loss for the original image and its masked and combined versions.

using a visual explanation method [23, 18]. Due to its efficiency, we use xGradCAM [18] for this means. Afterwards, a $H \times W \times 1$ binary mask $m^{(i)}$ for $X^{(i)}$ is made by selecting the top-K pixels with the highest scores. As stated before, the pixels with the highest score usually indicate the core parts of $X^{(i)}$. Now, according to $m^{(i)}$ two types of perturbations are applied on $X^{(i)}$:

Removing the non-core parts. The masked areas of $X^{(i)}$ are set to an uninformative value in the new image. The new obtained image can be formulated as below:

$$\hat{X}^{(i)}_{\text{rmv}} = m^{(i)} X^{(i)} + (1 - m^{(i)})b, \tag{1}$$

where b is a $1 \times 1 \times 3$ vector indicating the mean of B across the colour channels. The reason for setting the masked areas equal to b is to change the statistics of the batch as insignificantly as possible.

Combining the core and non-core parts of two images. Given two images $X^{(i)}, X^{(j)} \in B$ such that $y^{(i)} \neq y^{(j)}$, a combination of the two images containing the core of $X^{(i)}$ and non-core parts of $X^{(j)}$ is computed as below:

$$\begin{aligned}\hat{X}^{(i)}_{i,j} = m^{(i)} X^{(i)} + (1 - m^{(i)})(1 - m^{(j)}) X^{(j)} \\ + (1 - m^{(i)})(m^{(j)})b.\end{aligned} \tag{2}$$

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 background b.

Both the augmentations above generate data in order to break the spurious correlation between the non-core parts of the images and the labels. Finally, the loss function during retraining the last layer of the model is defined as below:

$$\mathcal{L}_{CE} = \frac{1}{N} \sum_{i=1}^{N} l(\theta, X^{(i)}, y^{(i)}), \tag{3}$$

$$\mathcal{L}_{\text{rmv}} = \frac{1}{N} \sum_{i=1}^{N} l(\theta, \hat{X}^{(i)}_{\text{rmv}}, y^{(i)}), \tag{4}$$

$$\mathcal{L}_{\text{comb}} = \frac{1}{N} \sum_{i=1}^{N} l(\theta, \hat{X}_{i,j_i}, y^{(i)}) \quad \text{where } y^{(j_i)} \neq y^{(i)}, \tag{5}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{\text{rmv}} + \beta \mathcal{L}_{\text{comb}}. \tag{6}$$

in which $l(\theta, ., .)$ is the is cross-entropy loss when the parameters of g equal θ. α and β are hyperparameters determining the effectiveness of each augmentation type.

4. Experiments

4.1. Datasets

Waterbirds Dataset. This dataset is created by combining bird photos from the Caltech-UCSD Birds-200-2011 [22] dataset with image backgrounds from the Places dataset [29]. The birds are labeled as either waterbirds or landbirds and are placed on either water or land backgrounds. Waterbirds are more frequently shown on water backgrounds, while landbirds are more often shown on land [17].

CelebA Dataset. CelebA celebrity face dataset in the presence of spurious correlations was proposed by [17]. In this dataset the binary label is assigned to the hair color and

Table 2: Worst group and mean test accuracy on Waterbirds and CelebA using ResNet-50. Group Labels column shows the amount of group label supervision by each method.

Method	Group Labels	Waterbirds		CelebA	
		Mean (%)	Worst (%)	Mean (%)	Worst (%)
ERM	✗	90.2	74.8	95.9	41.7
DFR	✓✓	94.2 ± 0.4	92.9± 0.2	91.3±0.3	**88.3±1.1**
MaskTune	✗	93.0±0.7	86.4±1.9	**91.3±0.1**	78.0±1.2
JTT	✓	93.3	86.7	88.0	81.1
MaC (ours)	✓	**94.3±0.9**	**93.0±0.7**	89.6±1.3	81.8±1.8

the gender is the attribute with spurious correlation with the label [11].

4.2. Setup

The model used in experiments is ResNet-50 pretrained on ImageNet. For ERM training, we used random crop and random horizontal flip as data augmentation similar to [2, 5]. We used SGD optimizer with the constant learning rate 10^{-3} and momentum 0.9 in both datasets. For the Waterbirds dataset, the model was trained for 100 epochs with batch size 32 and weight decay 10^{-3}, while for CelebA the model was trained for 30 epochs with batch size 128 and weight decay 10^{-4}.

For retraining the last layer, Adam optimizer with learning rate 0.5×10^{-2}, and step learning rate scheduler with step 5 and $\gamma = 0.5$ were used for both datasets. The training on both datasets was with batch size 32. Regarding the loss terms coefficients, $\alpha = 2, \beta = 3$ and $\alpha = 3, \beta = 0$ were used for Waterbirds and CelebA respectively. Also, to reduce the strong disturbance of class imbalance, we used class-balanced data to retrain the last layer on CelebA. Model selection and hyper-parameter fine-tuning are done according to the worst group accuracy on the validation set. For more implementation details refer to the Appendix.

4.3. Results

The results of our experiments along with reported results for DFR, Masktune, and JTT are illustrated in Table 2. Our method outperforms other methods in both mean and worst group accuracy on Waterbirds, which can partly be due to the higher proportion of spurious parts in comparison to CelebA. This property allows our method to make more sensible combined data, which as shown in Table 2 has a significant effect on the model's worst group accuracy. Also, MaC achieves higher worst group accuracy (the main objective of mitigating spurious correlations), on CelebA compared to MaskTune.

Among the methods in Table 2, JTT has the same amount of supervision on group labels as our method, i.e. it requires validation data group labels only for model selection. As shown in the results, our method has higher mean and worst group accuracies compared to this method.

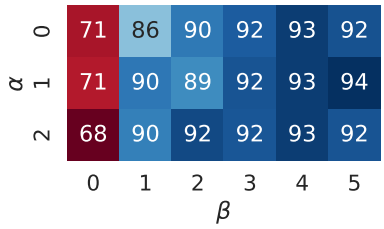

Figure 2: Effect of loss terms coefficients on the worst group accuracy on Waterbirds. The existence of the term $\mathcal{L}_{\text{comb}}$ in the loss function has a considerable effect on the worst group accuracy.

5. Conclusion

In this work, we considered the problem of spurious correlation between segments of images and labels. According to our observations, models trained with standard ERM on biased datasets are still able to attend to core parts of images, which is in agreement with the reports of [5]. To reduce spurious correlations even further, we proposed reducing the effect of the spurious parts of each image to force models to attend to the core segments. To this end, we masked images using xGradCAM to extract the core parts of the images, and by using these masks, we made two types of interventions on the spurious parts of the original data to decrease spurious correlations. The method has proven to be effective on two benchmarks and has comparable performance to DFR, which unlike our approach, requires the knowledge of group labels during training. Although the focus of this work was limited to spurious correlations between the segments of images and the labels, the idea can potentially be applicable in more complicated settings, where there is a spurious correlation between attributes of the object of interest and the label. Further study on more complex settings and more advanced interventions on images is left to future work.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[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. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[3] 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.

[4] 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.

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

[6] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021.

[7] 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.

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

[9] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu. Guided attention inference network. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, 42(12):2996–3010, dec 2020.

[10] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021.

[11] 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.

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

[13] 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*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[14] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

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

[16] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18347–18377. PMLR, 17–23 Jul 2022.

[17] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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, jun 2020. IEEE Computer Society.

[24] 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, October 2021.

[25] Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust fine-tuning, 2023.

[26] Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. Chroma-vae: Mitigating shortcut learning with generative classifiers. In S. Koyejo, S. Mohamed, A.

Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20351–20365. Curran Associates, Inc., 2022.

[27] 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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26484–26516. PMLR, 17–23 Jul 2022.

[28] 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.

[29] 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.

A. Related Work

A.1. Mitigating Spurious Correlation

It has long been known that deep models trained under standard ERM settings are vulnerable to spurious correlations[20, 17, 1]. This problem has been addressed in the literature under terms such as shortcut learning[26, 12] and simplicity bias[21, 19].

[5] states that models trained with ERM are capable of extracting both core and non-core features of an image and proposes to reweight the coefficients of features in the last layer of the predictor to make models robust to spurious correlations. Although this method has acceptable performance, it assumes that the group label of each validation data point is available, which is not feasible in many scenarios. Many methods have been proposed to retrain the last layer of models in the absence of group annotations [15, 7], most of which, are based on reweighting or selecting a subset of data for last layer retraining. As an example [15] upweights samples for which a model trained with ERM assigns a low probability to the correct class. In addition to this line of work, [10, 13, 8, 2, 27] introduce methods for fine-tuning whole models without group knowledge. [10] upweights the data misclassified by a model trained by ERM to fine-tune a whole new model. [8] trains two diverse classifiers and selects the one that performs better on the data from the minority group. [2], which is the most similar work to ours, hypothesizes that models trained with ERM mostly focus on parts of the image with high spurious correlation to the label, and proposes to mask parts of the image with the highest scores according to xGradCAM. Then a new pretrained model is trained on the masked data.

A.2. Attention-based Masking for Out-of-Distribution Generalization

Some other works were proposed for removing the irrelevant parts of images by masking [28, 9]. [24] proposes a causal attention module that generates data partitions and removes confounders progressively to enhance models' generalizability. [25] masks patches of images based on the class activation map and refills them from patches of other images and utilizes these samples for distillation with a pretrained model.

B. Implementation Details

α and β were selected from $\{0, 1, 2\}$ and $\{0, 1, 2, 3, 4, 5\}$ respectively according to worst group accuracy on the validation set. The models were trained for 6 and 10 epochs on CelebA and Waterbirds respectively and the results of the model with the highest worst group accuracy on the validation set was reported.

C. Visualizations

In the following, the class activation map of models trained with ERM and our method on some samples are illustrated.

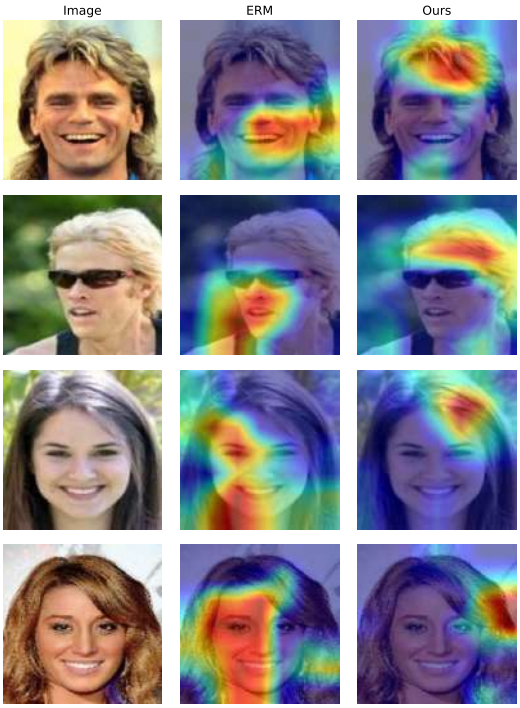

Figure 3: Saliency maps of models trained with ERM and our proposed method on CelebA samples which are misclassified by the base model trained with ERM.

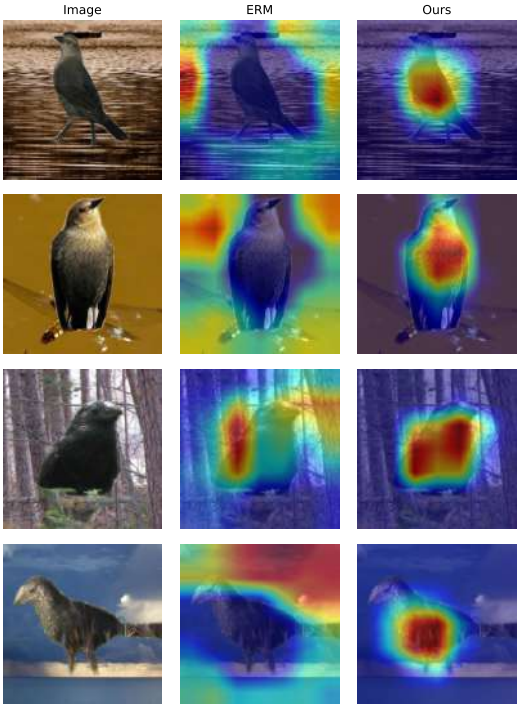

Figure 4: Saliency maps of models trained with ERM and our proposed method on Waterbirds samples which are misclassified by the base model trained with ERM.