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

Supplementary Material

7. Related Work

7.1. Mitigating Spurious Correlation

It has long been known that deep models trained under standard ERM settings are vulnerable to spurious correlations [2, 24, 29]. This problem has been addressed in the literature under terms such as shortcut learning [16, 38] and simplicity bias [28, 30].
Most well-known works in the literature approach mitigating spurious correlation by either group balancing or sample reweighting. Group DRO [24], which is one of the best-performing methods proposed so far, uses group annotations to minimize the worst group error. SUBG [6] trains a model by ERM on a random group-balanced subset of data and has proven to be effective on several benchmarks. Following [6], DFR [8] states that models trained with ERM are capable of extracting both core and non-core features of an image and proposes to retrain only the last layer of the predictor on a group-balanced subset of training or validation set to make models robust to spurious correlation. Although these methods have acceptable performance, they require group labels of the training or validation set for the training. This assumption is not feasible in many scenarios and has been addressed by several methods that aim to train robust models without access to the group labels. Among these methods, some introduce methods for reweighting or pseudo-labelling samples for last layer retraining [10, 20]. AFR[20] upweights samples for which a model trained with ERM assigns a low probability to the correct class. DD-DFR [10] assigns pseudo group labels to samples based on the change of the model's prediction on them when adding dropout to the model. In addition to this line of work, [13, 17] introduce methods for fine-tuning whole models without group knowledge by upweighting or upsampling the data misclassified by a model trained by ERM. JTT [13] upsamples datapoints which are misclassified by a model trained by ERM, with the assumption that these samples are mostly from under-represented groups.

The methods mentioned above have a common assumption that the misclassified samples or samples with high loss are mostly from minority groups. While this holds true for some samples, in many cases, the main reason behind the model's high loss on a sample is the complexity of its causal regions. On the contrary, we manually make combined images that are theoretically from the minority groups, as a means for upweighting under-represented samples.

7.2. Data Augmentation for Bias Mitigation

A line of work uses data augmentation for enhancing models' generalizationability [35, 37, 39]. Inspired by mixup [41], LISA [39] selectively interpolates datapoints across different groups or different labels to train an invariant predictor. DISC [35] utilizes a concept bank to detect spurious concepts and intervenes on samples using these concepts to balance the spurious attributes. In addition to these works, few works use synthetic data augmentation for balancing the training data [7, 21, 22, 34]. GAN debiasing [22] uses a GAN to generate images and intervene on them in the latent space. FFR [21] combines synthetic data augmentation and loss-based debiasing methods (such as Group DRO [24]) for mitigating spurious correlation.

Almost all the methods based on data augmentation require the knowledge of the spurious attributes or group labels, or use additional concept banks or generative models for detecting and intervening on spurious attributes. DaC on the other hand, augments the training data with none of the mentioned requirements.

7.3. Attention-based Masking for Out-of-Distribution Generalization

Some other works were proposed for removing the irrelevant parts of images by masking [11, 43]. CaaM [33] proposes a causal attention module that generates data partitions and removes confounders progressively to enhance models' generalizability. [36] masks patches of images based on the class activation map and refills them from patches of other images and utilizes these samples for representation distillation with a pretrained model. Decoupled-Mixup [14] distinguishes discriminative and noise-prone parts of images and fuses these parts by mixup separately. MaskTune [3] which is the most similar work to ours, based on the assumption that models trained with ERM mostly focus on parts of the image with high spurious correlation to the label, masks parts of the image with the highest scores according to xGradCAM. Then a new model is fine-tuned on the masked data.

None of the methods mentioned above, except MaskTune, strive to extract the causa parts of images in order

to determine the true label of the newly obtained images. However, a key point in DaC is that it distinguishes the causal parts from the non-causal regions to be able to make combined images and determine their label. Additionally, as discussed in Sec. 3, it cannot be simply assumed that the focus of models trained with ERM is on non-causal parts of images, which is the most noticeable downfall of MaskTune, that we aimed to solve to an extent.

8. Details on Experiments

8.1. Datasets

In this study, We compared methods on four datasets with distribution shifts. The first three datasets are related to correlation shift and the last one includes diversity shift between the train and test sets according to the categorization introduced in [40].

Waterbirds This dataset is created by combining bird photos from the Caltech-UCSD Birds-200-2011 [31] dataset with image backgrounds from the Places dataset [44]. The birds are labelled 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 [24].

CelebA CelebA celebrity face dataset in the presence of spurious correlations was proposed by [24]. In this dataset the binary label is assigned to the hair colour and the gender is the attribute with spurious correlation with the label [15].

Dominoes: This dataset, synthesized in a manner similar to [18], consists of paired images: one from CIFAR10 and one from MNIST. The CIFAR10 image, either an automobile or a truck, serves as the target label. Meanwhile, the MNIST image, a zero or a one, acts as the spurious part. The spurious correlation between MNIST digits and the label is 90%.

Metashift: Our setup for Metashift dataset follows [35]. The target is to classify cats and dogs, and spurious features are objects and backgrounds, namely sofa, bed, bench, and bike. The test images are from backgrounds that are not present in the training set.

8.2. Details on the CelebA Dataset

As mentioned in Sec. 5.3, in addition to the spurious correlation between gender (which can be inferred from the facial features) and hair colour, some hair attributes contribute to hair volume such as hair wave and baldness, which are correlated with the hair colour. The number of people with each hair colour and specific attributes is extracted from the CelebA metadata and shown in Tabs. 2 and 3. According to the statistics, while about 0.05% of blond people wear hats or are bald, more than 8 per cent of people who are not blond wear hats or are bald. Similarly, the percentage of blond people with wavy hair is more than 1.5 times greater than the ones that are not blond. Additionally, our eye observations from the dataset indicate that there is a correlation between the length of hair and its colour, as short hair is more co-occurred with non-blond hair. It is worth mentioning that since the attribute of hair length was not available in the CelebA metadata, we assessed this claim by eye observation. A few examples of randomly selected samples from each hair colour are shown in Fig. 14.

Table 2. Number of people with wavy hair with each hair colour.

	Blond = -1	Blond = 1
Wavy = -1	121761	16094
Wavy = 1	50855	13889

Table 3. Number of people with each hair colour that are bald or wear a hat.

	Blond = -1	Blond = 1
Bald = -1 $\wedge$ Hat = -1	158440	29817
Bald = 1 $\vee$ Hat = 1	14176	166

8.3. ERM Training Details

Similar to [8], we used SGD optimizer with learning rate 10^{-3} and momentum 0.9 for all datasets. We used weight decay 10^{-3} for Waterbirds, Metashift and Dominoes dataset and 10^{-4} for CelebA. The batch size for CelebA, Waterbirds, Metashift, and Dominoes were 128, 32, 16, and 16 respectively. The model was trained for 100 epochs on the Waterbirds and Metashift datasets, and for 30 and 15 epochs on the CelebA and Dominoes.

Table 4. Hyperparameters for DaC

Dataset	Hyperparameters		
	epochs	α	q
Waterbirds	20	10	0.6
CelebA	15	5	0.2
MetaShift	30	6	0.5
Dominoes	20	6	0.8

8.4. DaC Training Details

For all datasets, Adam optimizer with a learning rate of 0.5×10^{-2}, and step learning rate scheduler with step size 5 and gamma 0.5 were used. The batch size was 64 for all datasets. To encourage the diversity of training data during retraining the last layer, in cases when the selected samples with low loss in each batch were only from one class, we randomly combined the selected images with others from

Table 5. Mean and worst group accuracy on the validation sets of four datasets when applying DaC using the original or inverted masks.

Invert	Waterbirds		CelebA		Metashift		Dominoes	
Mask	Worst	Average	Worst	Average	Worst	Average	Worst	Average
✗	88.4	93.1	84.6	91.2	79	79	19.6	63.1
✓	22.6	63.2	83.9	90.1	45	60.1	89.2	93.0

the same class. No regularization terms were used for retraining the last layer of the model. The proportions for creating the curve of the loss with respect to the amount of masking in adaptive masking did not contain 1, since masking the whole image would trivially increase the loss on the masked image significantly. More details regarding the number of epochs, and optimal values for α and q are in Tab. 4. Batch size, α and q were selected from $\{32, 64\}$, $\{1, ..., 10\}$, and $\{0.2, 0.4, 0.5, 0.6, 0.8, 1\}$ respectively, and the criteria for hyperparameter selection was the worst group accuracy on the validation set.

8.5. Original Masks or Inverted Ones?

As mentioned in Sec. 4.2, we train the model in two settings corresponding to the ERM casual attention and ERM non-causal attention assumptions. For the former setting, i.e. ERM casual attention, we keep the parts obtained by adaptive masking as the causal parts while for the later one, i.e. ERM non-causal attention, the parts remained by adaptive masking are considered as non-causal and thus we invert the masks in order to obtain the causal regions for DaC. Based on the worst group accuracy of the model trained by DaC on the validation set in these two settings, it can be determined whether the parts to which the model generally pays more attention are causal or non-causal. The results for both cases are in Tab. 5. According to the results, unlike the Dominoes, on Waterbirds and Metashift the model attends more to the causal components. Regarding the CelebA dataset, it seems that the attention of the model does not grasp the entire hair parts in the image, hence, the inverted mask still contains a proportion of the hair. This was also reflected in the results in Tab. 1, in which, unlike other datasets, our model has a lower performance on CelebA. For more details on the CelebA dataset, refer to Sec. 8.2.

8.6. Details on the Kneedle Algorithm

As mentioned in Sec. 5.2, we use the *Kneedle* algorithm for finding the optimal amount of masking in Algorithm 1. This optimal amount is indicated by the *elbow* (i.e. the point with the highest curvature) of the curve of the loss with respect to the amount of masking. Since we only have access to a finite number of points from this curve, we use the Kneedle algorithm, which identifies elbows in a finite set of points from a curve.

The Kneedle method is based on the concept that knee points approximate the local maxima when the set of points is rotated about a specific line. This line is determined by the first and last points and is chosen to preserve the overall behaviour of the set. By rotating the curve about this line, knee/elbow points are identified as the points where the curve deviates most from the straight line segment connecting the set's endpoints. This approximation effectively captures the points of maximum curvature for the discrete set of points. The algorithm works as follows:

1. Smoothing: it applies a smoothing spline or other smoothing methods to data.
2. Normalization: It normalizes smoothed data by min-max normalization to function well regardless of the magnitude of data values.
3. Difference Computation: It defines D_d as the set of differences between x- and y- values. The knee is where the difference curve changes from horizontal to sharply decreasing.
4. Local Maxima Calculation: It identifies the local maxima of the difference curve as candidate knee points.
5. Threshold Calculation: For each local maximum (x_{lmx_i}, y_{lmx_i}) in the difference curve it defines the T_{lmx_i} which is based on the average difference between consecutive x values in the difference curve and a sensitivity parameter, S. This parameter determines how aggressive the method is. Smaller values for S, detect knees quicker, and large values are more conservative.

$$T_{lmx_i} = y_{lmx_i} - S.\frac{\sum_{i=1}^{n-1} x_{i+1} - x_i}{n-1}$$

6. Knee Declaration: If any difference value (x_i, y_i), where $j > i$, drops below the threshold $y = T_{lmx_i}$ before the next local maximum in the difference curve is reached, the method declares that local maximum as a knee point. Kneedle's run time for any given n pairs of $x-$ and $y-$ values is bounded by $\mathcal{O}(n^2)$.

8.7. Training Time

Since the ERM model used for computing the attribution scores of the pixels is fixed, extracting the attention heatmap and adaptive masking is done as a preprocess. Hence, during training, the previously prepared and saved masks are used, similar to MaskTune [3]. Additionally, since the optimal percentage of the masked pixels in *Adaptive Masking*

is selected among a small number of candidates, the time complexity of FindElbow is constant. The training time of several methods (excluding the ERM phase of the methods) on Waterbirds is shown in Tab. 6.

Table 6. The training time of different methods (excluding the ERM training phase) on the Waterbirds dataset on Nvidia A100 GPU

Method	DFR	CnC	JTT	MaskTune	Ours
Time (min)	4	85	58	6.5	18.9

9. More Empirical Observations

In Sec. 3, we claimed that the images on which the model trained with ERM has a low loss show specific properties. This assumption is valid since on the images from the majority groups both the causal and non-causal parts of images are in accordance with the label. Hence, even if the model attends more to the non-causal parts or its attention is divided between the causal and non-causal parts, it will still perform well on the datapoint and obtains a low loss. Fig. 6a illustrates that the images from minority groups are more among the images with high losses. On the other hand, images from majority groups are almost uniformly distributed between loss quantiles, with a slightly higher probability in lower loss quantiles, as shown in Fig. 6b. Since the probability of majority samples is higher than the minority ones across the dataset and $p(\text{low loss}|\text{majority})$ is high, it can be concluded that the probability of a low loss sample being from the majority groups is relatively high.

10. Comparison of Attribution Maps

The class activation maps of models trained with ERM and our method on some samples are illustrated in Figs. 7 to 10.

11. Combined Images

Some examples of combined images and their corresponding label are shown in Figs. 11 to 13.

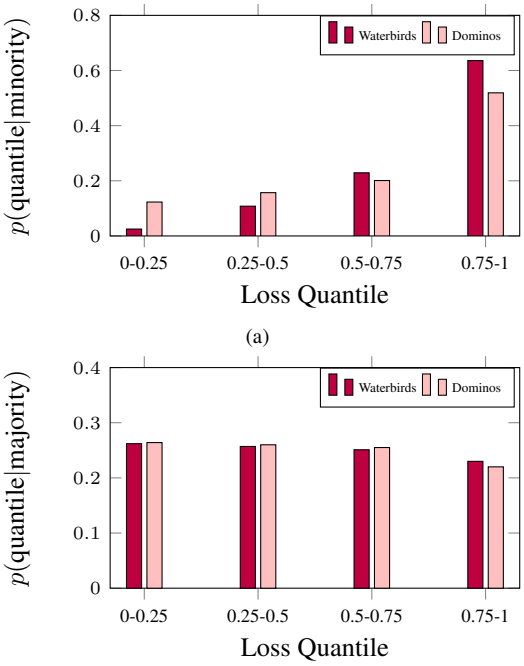

Figure 6. (a) Distribution of training images from minority groups between loss quantiles for the Waterbirds and Dominoes datasets. (b) Distribution of training images from majority groups between loss quantiles for the Waterbirds and Dominoes datasets.

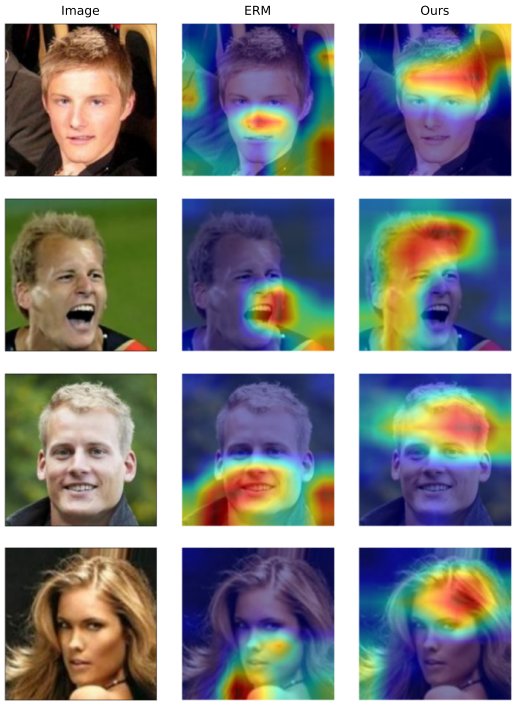

Figure 7. Saliency maps of models trained with ERM and our proposed method on CelebA samples which are misclassified by the base model trained with ERM.

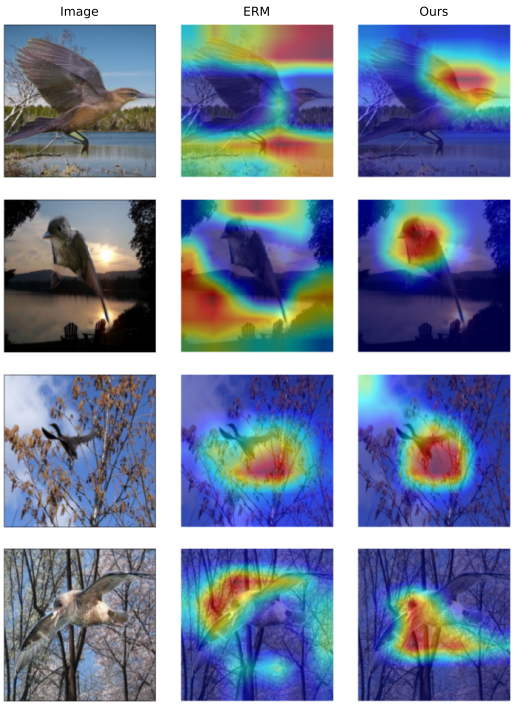

Figure 8. Saliency maps of models trained with ERM and our proposed method on Waterbirds samples which are misclassified by the base model trained with ERM.

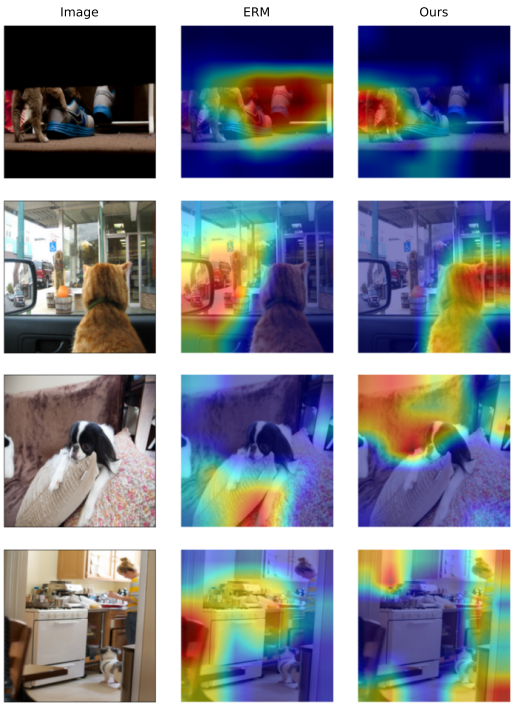

Figure 9. Saliency maps of models trained with ERM and our proposed method on Metashift samples which are misclassified by the base model trained with ERM.

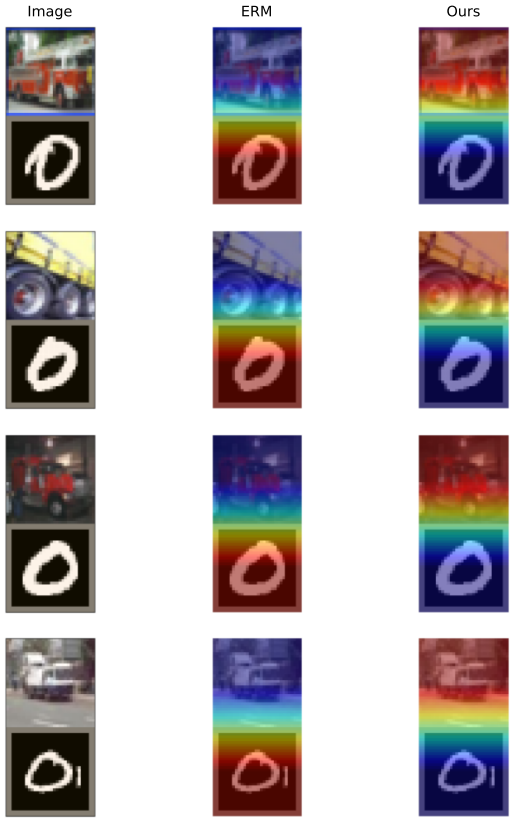

Figure 10. Saliency maps of models trained with ERM and our proposed method on Dominoes samples which are misclassified by the base model trained with ERM.

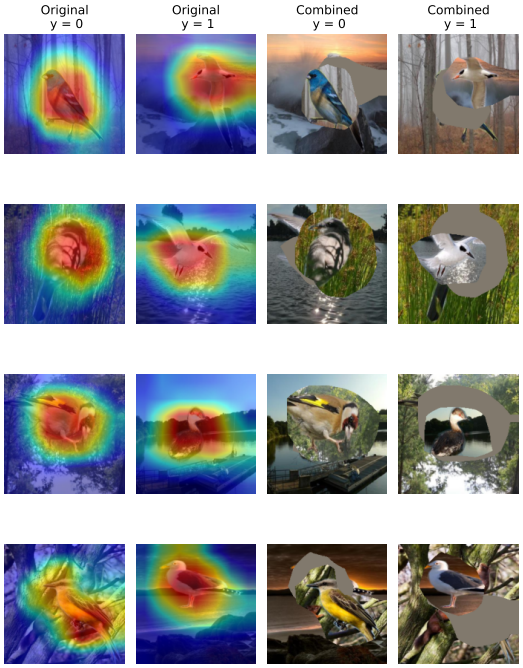

Figure 11. Low loss training samples in the Waterbirds and their combinations.

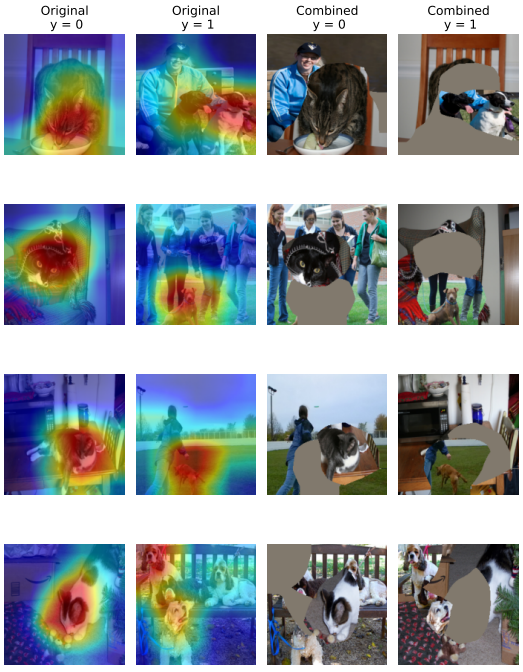

Figure 12. Low loss training samples in the Metashift and their combinations.

Figure 13. Low loss training samples in the Dominoes and their combinations.

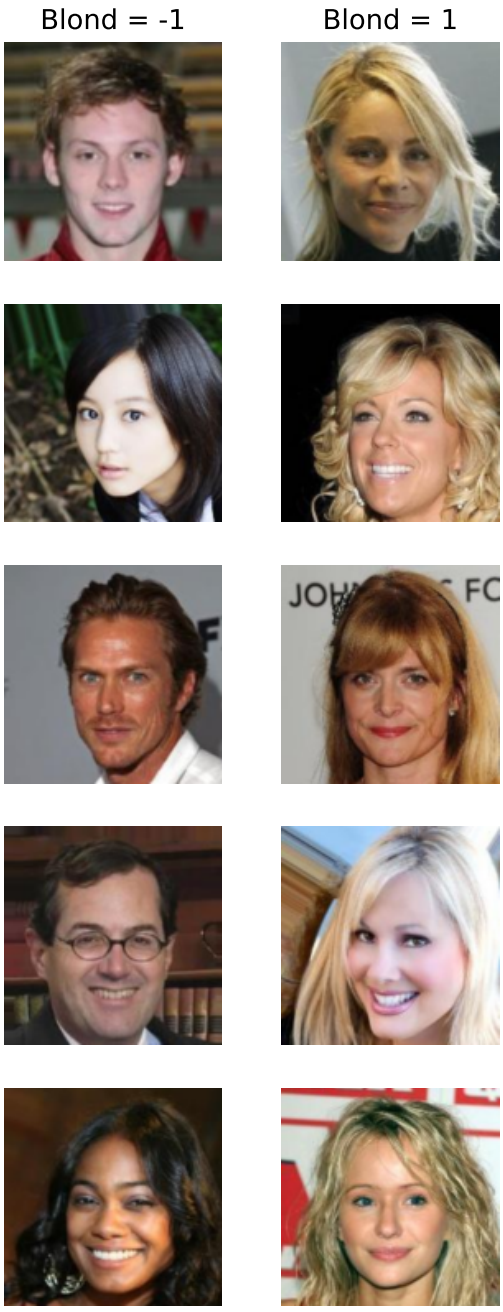

Figure 14. Some samples from the CelebA dataset.

References

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

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

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

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

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

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

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

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

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

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

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

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

[13] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 1, 6, 7

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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