Avoiding Spurious Correlations: Bridging Theory and Practice

Thao Nguyen*
University of Washington
thaottn@cs.washington.edu

Vaishnavh Nagarajan
Google Research
vaishnavh@google.com

Hanie Sedghi
Google Research, Brain team
hsedghi@google.com

Behnam Neyshabur
Google Research, Blueshift team
neyshabur@google.com

Abstract

Distribution shifts in the wild jeopardize the performance of machine learning models as they tend to pick up spurious correlations during training. Recent work [11] has characterized two specific failure modes of out-of-distribution (OOD) generalization, and we extend this theoretical framework by interpreting existing algorithms as solutions to these failure modes. We then evaluate them on different image classification datasets, and in the process surface two issues that are central to existing robustness techniques. For the algorithms that require access to group information, we demonstrate how the existing annotations included in standard OOD benchmarks are unable to fully capture the spurious correlations present. For methods that don't rely on group annotations during training, the validation set they utilize for model selection carries assumptions that are not realistic in real-world settings. This leads us to explore how the choice of distribution shifts represented by validation data would affect the effectiveness of different OOD robustness algorithms.

1 Introduction

Neural networks are commonly trained and evaluated with the assumption that the training distribution matches the test distribution. However, when this assumption does not hold because of distribution shifts in real-world deployments, machine learning models can perform in unexpected and undesirable ways [5, 17]. Previous work has shown how models can utilize different shortcuts to perform well on the task [4], a notable example of which is the presence of spurious correlations — features that correlate with the labels but are actually not causal, such as image backgrounds [16, 18, 23].

The explanation for why empirical risk minimization (ERM) approaches tend to rely on spurious correlations can be summarized by two lines of work. The first [1, 19, 21] assumes that both invariant and spurious features are only partially predictive of the label. As a result, both will be used to maximize accuracy. The other line of work [6, 20, 13] demonstrates how spurious features are simpler to learn than invariant ones and thus are preferred by gradient descent (i.e. simplicity bias), even when both kinds of features are fully predictive of the labels. To provide a concrete framework for understanding OOD generalization failures, [11] uses a series of tasks that are easy to succeed at to pinpoint fundamental mechanisms by which spurious correlations affect gradient-descent-trained linear classifiers: they induce *geometric* and *statistical* skews in the data.

Applying this theoretical framework to existing techniques for improving OOD generalization, we seek to tackle both types of skews in practice. We start by characterizing the algorithms in

*Work done as a member of the Google AI Residency.

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

consideration based on the failure mode that they address, and empirically evaluate their effectiveness when they are used alone or in conjunction with another, on several standard OOD benchmarks. Through this comparison, we find that there is no single solution that works best for all the datasets we examined. As the algorithms chosen for our investigation rely on the availability of group labels, we also investigate how well the official spurious feature annotations from these widely adopted OOD benchmarks encapsulate all the spurious correlations present in the data distribution. We find that even when different subgroups are balanced (in terms of the number of datapoints) by undersampling, there still exist substantial performance gaps among the groups. This suggests the presence of more complicated subgroup structures that cannot be captured by a single feature annotation alone.

In addition to the algorithms we examined, other existing robustness methods [10, 12] that don't require group labels at training time are often tuned with validation data that is representative of the distribution shift found at test time, in terms of group information. This prompts us to design a new validation set that closely resembles the degree of spurious correlation found in the training set instead, and explore how this affects the relative effectiveness of different algorithms.

In summary, our contributions are:

- We revisit existing algorithms for OOD generalization, and show how they could be used alone or in combination with each other to address the geometric and statistical skews defined in [11].
- We empirically evaluate the effectiveness of these methods on 3 image classification datasets with varying properties (dataset size, degree of spurious correlations, difficulty of learning the spurious feature). We find that no single approach consistently yields the best performance.
- We also repeat these evaluations in a more realistic setting that assumes no knowledge of distribution shifts at test time, as this information is currently implied in the official validation sets that most existing robustness algorithms use for hyperparameter tuning.
- We demonstrate that undersampling the training sets to balance the representation of different subgroups is not sufficient in capturing all the spurious correlations present — in each of the downsampled datasets, certain groups still appear to be intrinsically harder to learn than the rest.

2 Experimental Setup and Background

2.1 Failure Modes and Solutions

For the sake of our discussion, consider a simplified binary classification setting where any datapoint can be written as $(\mathbf{x}_{inv}, x_{sp})$ denoting the invariant feature (potentially high-dimensional) and a spurious feature. The label y of any point is either $+1$ or -1. While the label can be perfectly predicted by looking at x_{inv}, we also have that the spurious feature is partially correlated with the label, i.e., $\Pr_{\mathcal{D}}[x_{sp} \cdot y > 0] > 0.5$, where $\mathcal{D}$ is the data distribution. In other words, in a *majority* of points, the spurious feature is positively aligned with the label and in a *minority* of points, it is negatively aligned. By analyzing gradient descent on linear classifiers and its infinite-time-equivalent max-margin, [11] show that there are two main kinds of failure modes caused by the presence of spurious features. We briefly summarize these below.

2.1.1 Statistical skews

One mode of failure arises from the slow convergence rate of gradient descent (on the cross-entropy loss), relative to the max-margin classifier. That is, even in datasets where the max-margin classifier succeeds in using only the invariant features, finite-time (stochastic) gradient descent (SGD) still relies on spurious features unless trained for an exponentially long time. This is a purely statistical effect in that, in principle, it can be straightforwardly countered if we upweight or oversample the minority datapoints (through duplication) while running GD. This is in contrast to the second type of failure mode below which cannot be addressed by such statistical tricks.

2.1.2 Geometric skews

Another kind of failure mode is one that can occur to both GD and max-margin classifiers due to certain geometric effects in the dataset. Refer to Figure 1, which is taken from [11], for an illustration of the 2D setting. In particular, it is often the case that the "invariant margin" (i.e., the margin of separation in the purely-invariant feature subspace) of the minority group is much larger than that of the majority group. This arises from a property of the invariant features: when there are more uniquely

sampled points, we are likely to see more difficult points that make the data harder to separate. As a result of this imbalance of margins across groups, any classifier with a margin-maximizing bias can be shown to make use of the spurious features. Note that simply upsampling the minority group cannot address this issue since it does not affect the number of unique datapoints, and thereby the margins of minority and majority groups.

2.1.3 How do different algorithms address these skews?

As already noted above, the idea of upsampling the minority can only address statistical skews. If we were to however *downsample* the majority (to equal the size of the minority), we would not only nullify statistical skews, but also roughly equalize the margins in the majority and minority groups, thereby nullifying any geometric skews as well.

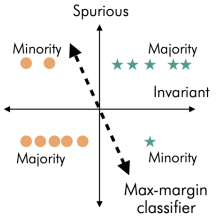

Figure 1: **Geometric failure mode (from [11]).**

An alternative way to address geometric skews is to directly scale down the margins of the minority datapoints, in effect making it equal to those of the majority, so models are encouraged to separate majority and minority groups using only the invariant feature. Formally, let S_{maj} and S_{min} be the majority and minority groups respectively. Then, we seek a classifier that maximizes the following notion of margins:

$$\arg \max_{\mathbf{w}, \|\mathbf{w}\|=1} \min \left(\{c \cdot y(\mathbf{w} \cdot \mathbf{x}) | \mathbf{x} \in S_{min}\}, \{y(\mathbf{w} \cdot \mathbf{x}) | \mathbf{x} \in S_{maj}\} \right), \tag{1}$$

where c is a small multiplicative constant intended to scale down the minority margins. A formal setup of this algorithm can be found in Appendix D of [11] or in related work [8, 14, 2]. As [8] propose, one can extend this approach for max-margin classifiers to deep neural networks trained with cross-entropy loss, by multiplying the *logits* of minority datapoints with a small constant (to be tuned) before computing the loss. Theoretical expansion of this case is beyond the scope of our work, given our focus on connecting robustness methods to real-world evaluations.

Another popular algorithm to address spurious correlations is group DRO [7, 15]. Here the objective is to minimize the worst-case cross-entropy loss of different groups of points. This should eventually converge to the standard max-margin and therefore not help with geometric skews. On the other hand, an alternative called group-adjusted DRO proposed in [18] makes the minority group's objective harder by adding a constant to its loss (we use this version in the tables of results). By doing so, we believe this mechanism should help address geometric skews as it makes all the groups equally hard to learn. However, it remains uncertain how groupDRO counters the effect of statistical skews.

2.2 Datasets

We experiment with 3 datasets that have served as popular benchmarks for OOD generalization algorithms [10, 3, 19]. They are chosen for their varying dataset sizes, degrees of spurious correlation (both natural and artificially induced), as well as levels of difficulty for learning the spurious feature:

- Colored MNIST[1]: the task is to classify the digit images into 2 classes (digits < 5 and digits $\geq$ than 5). The label is spuriously correlated with the color of the digit, which could be either "red" or "green". We modify the procedure from [1] to remove the noise introduced to the labels.
- Waterbirds [18]: the task is to classify bird images as either "landbird" or "waterbird". The label is spuriously correlated with the image background which could be either "land" or "water" places.
- CelebA [18]: the task is to classify celebrity pictures as having either "blond" or "black" hair. The label is spuriously correlated with the gender which is either "male" or "female". We modify the original setup in [18] by using only a fraction of the data from the (female, dark hair) group to keep the size of this "minority" group small.

For each dataset, we use the official group annotations that come with the dataset to split the training data into 4 groups, 2 minority and 2 majority. More details on the construction and composition of each dataset can be found in Appendix A.

3 Results

The algorithms of interest have been described and linked to the respective skews in Section 2.1.3. Refer to Appendix B for further details on hyperparameters and model choices. For upsampling

(downsampling), we follow the standard practice and oversample (undersample) datapoints from each group to match the size of the biggest (smallest) group in the original dataset. Hyperparameter tuning is done based on the worst-group accuracy of the validation set, given previous work [10] which shows that worst-group test accuracies are significantly lower when average validation accuracy is used for model selection instead of worst-group validation accuracy.

Table 1: **Comparison of models trained with different methods, tuned on the official validation set, which mirrors the distribution shift found at test time.** The table is sectioned based on the skew(s) that each approach addresses. The numbers in brackets capture the performance changes compared to the baseline (ERM). We observe that the best-performing approach is not consistent across all datasets, and methods that address complementary skews when combined often yield complementary performance benefits.

Algorithm	ColoredMNIST	Waterbirds	CelebA
ERM	93.1	71.7	53.3
Upsampling	96.1 (+3.0)	86.0 (+14.3)	85.0 (+31.7)
Margin Scaling	95.2 (+2.1)	81.9 (+10.2)	57.7 (+4.4)
GroupDRO	**97.4 (+4.3)**	**91.1 (+19.4)**	88.4 (+35.1)
Downsampling	96.1 (+3.0)	87.6 (+15.9)	**88.9 (+35.6)**
Margin Scaling + Upsampling	96.2 (+3.1)	85.0 (+13.3)	87.8 (+34.5)
GroupDRO + Upsampling	96.5 (+3.4)	87.6 (+15.9)	86.7 (+33.4)

We report worst-group test accuracy from the best performing model trained with each method in Table 1. Across all 3 datasets, we tend to obtain larger performance boosts from tackling geometric skews (e.g. through GroupDRO) than statistical skews (e.g. through upsampling).

We also observe that no method reliably yields the best performance for all the settings that we studied. Those that target different skews (e.g. margin scaling and upsampling, or groupDRO and upsampling) offer *complementary* performance gains when used in combination on some datasets (e.g. CelebA and ColoredMNIST), but this is not always the case (e.g. Waterbirds). This suggests an opportunity for future work to study the implications of combining different OOD robustness methods on the optimization process and generalization.

Given the dependence of the algorithms on spurious feature annotations, we analyze how well the provided group information behaves in accordance with the theoretical framework [11], by looking at the individual group validation accuracies when models are trained on downsampled datasets, an approach that is expected to get rid of most geometric and statistical skews in the original data. In ColoredMNIST, where the spurious correlation is synthetically introduced and controlled, all groups now indeed exhibit similar performance. However, in the case of Waterbirds and CelebA, given the same number of datapoints, certain groups — which are not always the minority — still appear to be harder to learn than the rest (see Figure 2). This suggests the presence of more complex subgroup structures beyond the official annotations that come with these 2 widely used OOD benchmarks.

In addition, we note that the existing validation sets from these benchmarks reflect the distribution shifts at test time, which should not be expected in real-world settings. This issue is also applicable to algorithms that don't require group information during training (e.g. [10, 12]), as they still rely on annotated validation data for hyperparameter tuning. Thus, we investigate what Table 1 would look like in a more "in-the-wild" setting where we have no knowledge of future distribution shifts, and models are selected based on information related to the training distribution alone. We construct a new validation set with similar degrees of spurious correlation as the training set by subsampling from the original validation set (see Appendix A for the group breakdown). Consequently, we find that the relative effectiveness of the robustness methods in consideration could be sensitive to the choice of distribution shifts represented by the validation data (Table 2). For instance, on the Waterbirds dataset, GroupDRO + Upsampling is now much less effective compared to the other methods, and on the ColoredMNIST dataset, Margin Scaling + Upsampling no longer offers complementary performance gains compared to using each method by itself. This observation illustrates the importance of designing OOD generalization techniques that could do without group annotations completely.

4 Conclusion

Based on the theoretical framework introduced in [11], our work connects different techniques for boosting OOD generalization to the types of skews (i.e. geometric or statistical) induced by spurious

Table 2: **Comparison of models trained with different methods, tuned on a new validation set we constructed that follows the training group distribution instead.** We consider a more practical setting that carries no assumption about future distribution shifts: the validation set should resemble the training set in terms of group distribution. When hyperparameters are tuned on this new validation set, the relative effectiveness of certain methods varies substantially from that in Table 1.

Algorithm	ColoredMNIST	Waterbirds	CelebA
ERM	92.1	50.0	53.3
Upsampling	95.1 (+3.0)	83.0 (+33.0)	81.1 (+27.8)
Margin Scaling	94.6 (+2.5)	66.5 (+16.5)	56.7 (+3.4)
GroupDRO	**97.0 (+4.9)**	83.5 (+33.5)	88.3 (+35.0)
Downsampling	94.0 (+1.9)	**85.1 (+35.1)**	**88.4 (+35.1)**
Margin Scaling + Upsampling	94.2 (+2.1)	77.3 (+27.3)	87.8 (+34.5)
GroupDRO + Upsampling	95.8 (+3.7)	62.5 (+12.5)	88.3 (+35.0)

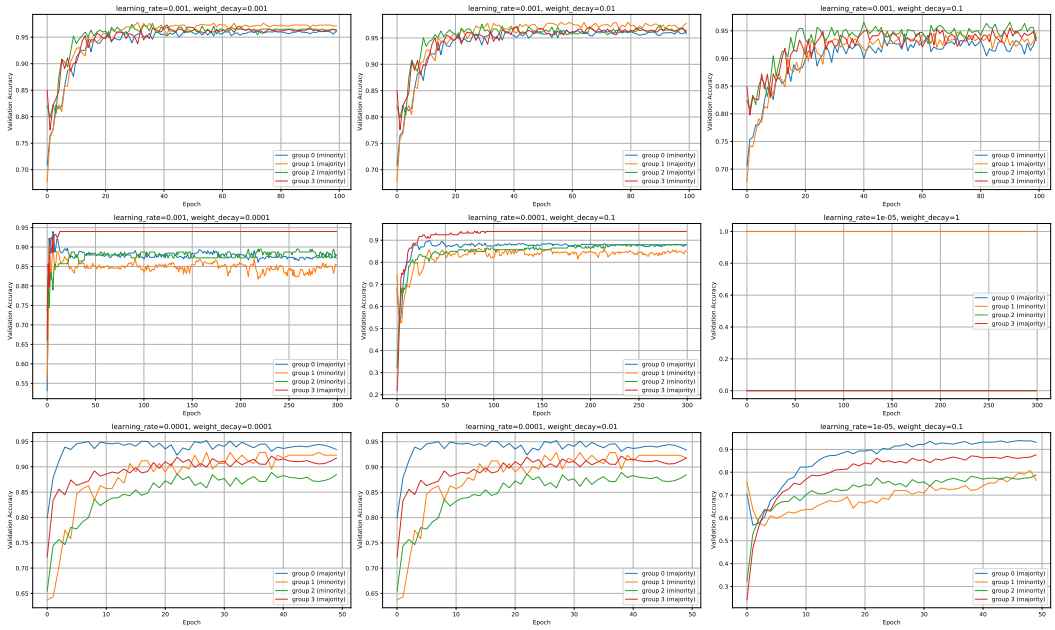

Figure 2: **Validation accuracies of subgroups in the downsampled datasets (rows), across different hyperparameter settings (columns) used.** While training with ColoredMNIST downsampled data (top row) results in the ideal scenario of having the same performance across different subgroups — suggesting that the skews induced by spurious correlations are completely eliminated — performance gaps between subgroups are still discernible when training on Waterbirds and CelebA downsampled datasets (bottom 2 rows). We hypothesize that this is because the spurious feature (i.e. "color") is synthetically introduced and thus carefully controlled in ColoredMNIST, whereas the group information that comes with the other 2 datasets (and widely adopted in existing OOD benchmarks) doesn't fully capture all sources of spurious correlations present in the data.

features that they address. We look at how these methods perform in practice on 3 standard image classification datasets with different extents of spurious correlations, as well as how effectively they work in combination to address both types of skews. We then increase the resemblance to real-world setups further, by tuning models on a new validation set that doesn't assume any knowledge of distribution shifts at test time. In the process, we find that the best approach varies by dataset, and the structure of the validation set plays an important role in the effectiveness of different approaches (relative to the ERM baseline). This necessitates further study into a more thorough way to evaluate existing OOD robustness methods, even those that don't require group labels at training time. Last but not least, we also demonstrate how performances across subgroups still vary substantially even when all the groups are represented equally in the training set. This presents an opportunity for future work to study automatic group structure detection, or explore other possible sources of spurious correlations that may be present in standard OOD benchmarks beyond the given annotations.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019.

[3] E. Creager, J.-H. Jacobsen, and R. Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[4] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[5] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[6] K. L. Hermann and A. K. Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *arXiv preprint arXiv:2006.12433*, 2020.

[7] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[8] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *arXiv preprint arXiv:2103.01550*, 2021.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[11] V. Nagarajan, A. Andreassen, and B. Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

[12] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin. Learning from failure: Training debiased classifier from biased classifier. *arXiv preprint arXiv:2007.02561*, 2020.

[13] K. Nar, O. Ocal, S. S. Sastry, and K. Ramchandran. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *arXiv preprint arXiv:1901.08360*, 2019.

[14] H. Narasimhan and A. K. Menon. Training over-parameterized models with non-decomposable objectives. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[15] Y. Oren, S. Sagawa, T. B. Hashimoto, and P. Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[17] A. Rosenfeld, R. Zemel, and J. K. Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[18] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[19] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[20] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*, 2020.

[21] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[23] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[24] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Appendix

A Dataset Details

A.1 ColoredMNIST

We extract 12000 random images from the official MNIST training set [9] to create a validation set, on which hyperparameters are tuned. Following previous work [1], to construct ColoredMNIST dataset, we divide the MNIST images into 2 classes: $y = 0$ for digits 0-4 and $y = 1$ for digits 5-9. We then color all images in the first class red and the second class green. With probability 0.1, we flip the color of each training image. We perform a similar procedure for validation and test data, but with probability 0.9.

Table 3: **Group breakdown (in terms of number of images) for ColoredMNIST dataset.**

Dataset	Digit < 5, Green	Digit < 5, Red	Digit ≥ 5, Green	Digit ≥ 5, Red
Train	2454	21758	21365	2423
Validation (same as train domain)	70	621	604	69
Validation (same as test domain)	5324	621	604	5451
Test	4545	523	465	4467

A.2 Waterbirds

The dataset is constructed from placing segmented bird images from the Caltech-UCSD Birds-200-2011 dataset [22] on top of backgrounds taken from the Places dataset [24]. Each bird image is broadly classified as either waterbird or landbird, and background images are divided into land and water backgrounds, with waterbirds (landbirds) appearing more frequently on water (land) backgrounds.

Table 4: **Group breakdown (in terms of number of images) for Waterbirds dataset.**

Dataset	Landbird, Land background	Landbird, Water background	Waterbird, Land background	Waterbird, Water background
Train	3498	184	56	1057
Validation (same as train domain)	467	25	7	133
Validation (same as test domain)	467	466	133	133
Test	2255	2255	642	642

A.3 CelebA

In this dataset, the label ("blond" or "dark" hair color) is spuriously correlated with the demographic information ("male" or "female" gender), with blond females and dark-hair males appearing more frequently. We modify the group composition in [18] to make sure the infrequent associations ("dark-hair females" in particular) also have much fewer datapoints in the dataset.

Table 5: **Group breakdown (in terms of number of images) for CelebA dataset.**

Dataset	Blond, Female	Blond, Male	Dark Hair, Female	Dark Hair, Male
Train	22880	1387	4297	66874
Validation (same as train domain)	2874	174	532	8276
Validation (same as test domain)	2874	182	8535	8276
Test	2480	180	9767	7535

B Training Details

All models are trained with stochastic gradient descent (SGD) with momentum 0.9. Specific details about architecture and training hyperparameters are deferred to the individual subsections. For margin scaling, we tune the multiplicative factor c for logits of minority examples over the values in $[0.01, 0.05, 0.1, 0.25, 0.5, 0.75]$. For GroupDRO objective, we tune the group adjustment term C over the values in $[1, 2, 3, 4, 5]$, following [18].

B.1 ColoredMNIST

For the ColoredMNIST task, we use a multi-layer perceptron that consists of 2 hidden layers with 512 neurons in each. All methods are trained with batch size 128, for 50 epochs. We tune learning rate and l_2 regularization strength with the following configurations: $[(1e-2, 1e-3), (1e-2, 1e-2), (1e-2, 1e-1), (1e-3, 1e-3), (1e-3, 1e-2), (1e-3, 1e-1)]$.

B.2 Waterbirds

All methods are trained with ResNet-50 for 300 epochs, using batch size 64. The model weights are pretrained on ImageNet. We tune the hyperparameters over the 3 pairs of learning rate and l_2 regularization strength used by [18]: $[(1e-3, 1e-4), (1e-4, 1e-1), (1e-5, 1)]$.

B.3 CelebA

We use ResNet-50 pretrained on ImageNet for CelebA models, with batch size 128 and 50 epochs of training. Similar to Waterbirds, we also use the 3 pairs of learning rate and l_2 regularization strength from previous work [18] to tune hyperparameters: $[(1e-4, 1e-4), (1e-4, 1e-2), (1e-5, 1e-1)]$.