Saving a Split for Last-layer Retraining can Improve Group Robustness without Group Annotations

Tyler LaBonte [1] **Vidya Muthukumar** [1] **Abhishek Kumar** [2]

Abstract

Empirical risk minimization (ERM) of neural networks is prone to over-reliance on spurious correlations and poor generalization on minority groups. The recent *deep feature reweighting* technique (Kirichenko et al., 2023) achieves state-of-the-art group robustness via simple last-layer retraining, but it requires held-out group annotations to construct a group-balanced reweighting dataset. We examine this impractical requirement and find that last-layer retraining can be surprisingly effective without group annotations; in some cases, a significant gain is solely due to class balancing. Moreover, we show that instead of using the entire training dataset for ERM, dependence on spurious correlations can be reduced by holding out a small split of the training dataset for class-balanced last-layer retraining. Our experiments on four benchmarks across vision and language tasks indicate that this method improves worst-group accuracy by up to 17% over class-balanced ERM on the original dataset despite using *no additional data or annotations* – a surprising and unexplained result given that the two splits have equally drastic group imbalance.

1. Introduction

Classification tasks in machine learning often suffer from *spurious correlations*: patterns which are predictive of the target class in the training dataset but irrelevant to the true classification function. These spurious correlations, often in conjunction with the target class, create *minority groups* which are underrepresented in the training dataset. For example, in the task of classifying cows and camels, the training dataset may be biased so that a desert background is spuriously correlated with the camel class, creating a minority group of camels on grass backgrounds (Beery et al., 2018). Beyond this simple scenario, spurious correlations have been observed in high-consequence applications including medicine (Zech et al., 2018), justice (Chouldechova, 2016), and facial recognition (Liu et al., 2015).

Neural networks trained via the standard procedure of empirical risk minimization (ERM) (Vapnik, 1998), which minimizes the average training loss, tend to overfit to spurious correlations and generalize poorly on minority groups (Geirhos et al., 2020). Even worse, it is possible for ERM models to rely exclusively on the spurious feature and incur minority group performance that is no better than random guessing (Shah et al., 2020). Therefore, maximizing the model's *group robustness*, quantified by its worst accuracy on any group, is a desirable objective in the presence of spurious correlations (Sagawa et al., 2020).

In contrast to more generic distribution shift settings (*e.g.,* domain generalization (Koh et al., 2021)), the presence of spurious correlations enables the improvement of group robustness merely by addressing model bias (without collecting additional minority group data). The recently proposed *deep feature reweighting* (DFR) (Kirichenko et al., 2023; Izmailov et al., 2022) technique efficiently corrects model bias by retraining the last layer of the neural network, a simple procedure which achieves state-of-the-art group robustness. The key hypothesis underlying DFR is that ERM models which overfit to spurious correlations still learn *core features* that correlate with the ground-truth label on all groups, but they perform poorly because they overweight the spurious features in the last layer. Ostensibly, retraining the last layer on a group-balanced *reweighting dataset* would upweight the core features and improve worst-group accuracy.

DFR compares favorably to existing methods such as *group distributionally robust optimization* (DRO) (Sagawa et al., 2020), which requires group annotations for the entire training dataset. However, DFR still necessitates additional held-out data and group annotations to achieve maximal performance (Kirichenko et al., 2023). This requirement limits its practical application, as the groups are often unknown ahead of time or difficult to annotate (*e.g.,* due to financial, privacy, or fairness concerns).

[1]Georgia Institute of Technology, Atlanta, GA USA [2]Google DeepMind, Mountain View, CA USA. Correspondence to: Tyler LaBonte <tlabonte@gatech.edu>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

Our contributions. We examine the performance of last-layer retraining in the absence of group annotations on four well-established benchmarks for group robustness across vision and language tasks.[1]

First, we show that while group balance is the most important factor in DFR performance, in some cases a significant gain is solely due to class balancing. Therefore, we propose *class-balanced last-layer retraining* as a simple but strong baseline for group robustness without group annotations. Our experiments show that, on average over the four datasets, this method achieves 94% of DFR worst-group accuracy compared to 76% without class balancing. Moreover, while class-balanced ERM was recently proposed as a competitive baseline for worst-group accuracy (Idrissi et al., 2022), we observe that class-balanced last-layer retraining renders balancing in the ERM stage optional, as retraining results in similar performance regardless of ERM balance.

The performance of class-balanced last-layer retraining reveals a "free lunch" with practical ramifications. We show that instead of using the entire training dataset for ERM, dependence on spurious correlations can be reduced by holding out a small split of the training dataset for last-layer retraining. Our results indicate this method improves worst-group accuracy by up to 17% over class-balanced ERM on the original dataset despite using *no additional data or annotations* – a surprising and unexplained result given that the two splits have equally drastic group imbalance.

2. Preliminaries

Setting. We consider classification tasks with input domain $\mathcal{X}$ and target classes $\mathcal{Y}$. We assume $\mathcal{S}$ is a set of *spurious features* such that each sample $x \in \mathcal{X}$ is associated with one feature $s \in \mathcal{S}$. In conjunction with the target class, the spurious features partition the dataset into groups $\mathcal{G} = \mathcal{Y} \times \mathcal{S}$. While the groups may be heavily imbalanced in the training dataset, we desire a model which is invariant to the spurious feature and has roughly uniform performance over $\mathcal{G}$. Therefore, we evaluate *worst-group accuracy*, *i.e.,* the minimum accuracy among all groups (Sagawa et al., 2020).

We will often refer to datasets and models as *group-balanced* or *class-balanced*, meaning that in expectation, the dataset is composed of an equal number of samples from groups in $\mathcal{G}$ or classes in $\mathcal{Y}$, respectively. This balance can be achieved by training on a subset with equal data from each group/class, or sampling from the data so that each minibatch is balanced in expectation (Idrissi et al., 2022). To make the latter more concrete, for group balancing we first sample $s \sim \text{Unif}(\mathcal{S})$, then sample $x \sim \hat{p}(\cdot|s)$ where $\hat{p}$ is the training distribution; class balancing is the same with $\mathcal{Y}$ instead of $\mathcal{S}$. We use the minibatch sampling approach for both class-balanced ERM and class-balanced last-layer retraining, and we provide a comparison with the subset method in Appendix A.

Deep feature reweighting. The recently proposed *deep feature reweighting* (DFR) (Kirichenko et al., 2023; Izmailov et al., 2022) method achieves state-of-the-art worst-group accuracy by performing ERM on the training dataset, then retraining the last layer of the neural network on a group-balanced held-out dataset, called the *reweighting dataset*. In the original implementation, half the validation set is used to construct the reweighting dataset: all data from the smallest group is included and the other groups are randomly downsampled to that size. Then, the feature embeddings (*i.e.,* the outputs of the penultimate layer) of the reweighting dataset are pre-computed and used to train a logistic regression with explicit ℓ_1-regularization. The results are averaged over 10 randomly subsampled reweighting datasets, and a hyperparameter search is performed over ℓ_1 regularization strength on the other half of the validation set.

To emphasize practicality and efficiency, our implementation of DFR has some differences from the original. *(i)* Instead of logistic regression, we train the last layer on the reweighting dataset via minibatch optimization using SGD and AdamW (Loshchilov & Hutter, 2019) for the vision and language tasks, respectively. This fits well into standard training pipelines and avoids pre-computing the feature embeddings and writing them to disk, which can be slow and memory-intensive. *(ii)* To reduce the number of hyperparameters, we use a fixed-value ℓ_2 regularization instead of searching over ℓ_1 regularization strength. We observed similar performance for ℓ_1 and ℓ_2 regularization, which we believe is because ℓ_1-regularized gradients do not induce sparsity (Langford et al., 2009). *(iii)* We sample uniformly at random from the groups in the held-out dataset (to get group balanced minibatches) instead of averaging over group-balanced subsets of the data (Idrissi et al., 2022). We compare the implementations in detail in Appendix A.

Datasets and models. We study four datasets which are well-established as benchmarks for group robustness across vision and language tasks, detailed in Appendix B.

- *Waterbirds* (Welinder et al., 2010; Wah et al., 2011; Sagawa et al., 2020) is an image classification dataset where the task is to predict whether a bird is a landbird or a waterbird. The spurious feature is the image background: more landbirds are present on land backgrounds than waterbirds, and vice versa.
- *CelebA* (Liu et al., 2015; Sagawa et al., 2020) is an image classification dataset where the task is to predict whether

[1] Following previous work in this setting (Sagawa et al., 2020; Liu et al., 2021; Nam et al., 2022; Kirichenko et al., 2023; Izmailov et al., 2022), we assume access to a small validation set with group annotations for model selection. Concurrent research has explored the removal of this requirement (Lee et al., 2023).

Table 1: **Last-layer retraining on the held-out dataset.** While unbalanced (UB) last-layer retraining decreases performance, class-balanced (CB) last-layer retraining is competitive with state-of-the-art methods for group robustness without group annotations. CB ERM is trained on the combined training and held-out datasets using class-balanced minibatches, while JTT and RWY-ES only use the training dataset. DFR uses group annotations on the held-out dataset, while Group DRO-ES requires them for the training dataset. We list the mean and standard deviation over three independent runs.

Method	Group annotations	Worst-group test accuracy			
		Waterbirds	CelebA	CivilComments	MultiNLI
Class-balanced ERM	✗	$81.9_{\pm 3.4}$	$67.2_{\pm 5.6}$	$61.4_{\pm 0.7}$	$\mathbf{69.2_{\pm 1.6}}$
JTT (Liu et al., 2021; Idrissi et al., 2022)	✗	$85.6_{\pm 0.2}$	$75.6_{\pm 7.7}$	$67.8_{\pm 1.6}$	$67.5_{\pm 1.9}$
RWY-ES (Idrissi et al., 2022; Izmailov et al., 2022)	✗	$74.5_{\pm 0.0}$	$\mathbf{76.8_{\pm 7.7}}$	$78.9_{\pm 1.0}$	$68.0_{\pm 0.4}$
UB last-layer retraining	✗	$88.0_{\pm 0.8}$	$41.9_{\pm 1.4}$	$57.6_{\pm 4.2}$	$64.6_{\pm 1.0}$
CB last-layer retraining	✗	$\mathbf{92.6_{\pm 0.8}}$	$73.7_{\pm 2.8}$	$\mathbf{80.4_{\pm 0.8}}$	$64.7_{\pm 1.1}$
DFR (our impl.)	✓	$92.4_{\pm 0.9}$	$87.0_{\pm 1.1}$	$81.8_{\pm 1.6}$	$70.8_{\pm 0.8}$
DFR (Kirichenko et al., 2023; Izmailov et al., 2022)	✓	$91.1_{\pm 0.8}$	$89.4_{\pm 0.9}$	$78.8_{\pm 0.5}$	$72.6_{\pm 0.3}$
Group DRO-ES (Sagawa et al., 2020; Izmailov et al., 2022)	✓	$90.7_{\pm 0.6}$	$90.6_{\pm 1.6}$	80.4	73.5

a person is blond or not. The spurious feature is gender; there are more blond women than blond men.

- *CivilComments* (Borkan et al., 2019; Koh et al., 2021) is a text classification dataset where the task is to predict whether a comment is toxic or not. The spurious feature is the presence of one of the following categories: male, female, LGBT, black, white, Christian, Muslim, or other religion. More toxic comments contain one of these categories than non-toxic comments, and vice versa.
- *MultiNLI* (Williams et al., 2018; Sagawa et al., 2020) is a text classification dataset where the task is to predict whether a pair of sentences is a contradiction, entailment, or neither. The spurious feature is a negation in the second sentence – more contradictions have this property than entailments or neutral pairs.

We utilize a ResNet-50 (He et al., 2016) pretrained on ImageNet-1K (Russakovsky et al., 2015) for Waterbirds and CelebA, and a BERT (Devlin et al., 2019) model pretrained on Book Corpus (Zhu et al., 2015) and English Wikipedia for CivilComments and MultiNLI. We use half the validation set for feature reweighting (Kirichenko et al., 2023; Izmailov et al., 2022) and half for model selection with group annotations (Sagawa et al., 2020; Liu et al., 2021; Kirichenko et al., 2023; Nam et al., 2022; Izmailov et al., 2022). See Appendix B for further details.

3. Class-balanced last-layer retraining

In this section, we investigate the necessity of a group-balanced reweighting dataset for DFR and show that *class-balanced last-layer retraining* is a simple but strong baseline for group robustness without group annotations. To enable a fair comparison with our implementation of DFR (see Section 2), class-balanced last-layer retraining follows the same training procedure, except the reweighting dataset is constructed by sampling uniformly over the classes $\mathcal{Y}$ instead of the groups $\mathcal{G}$.

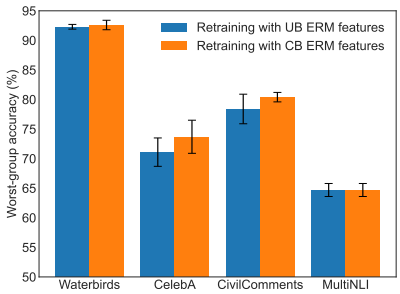
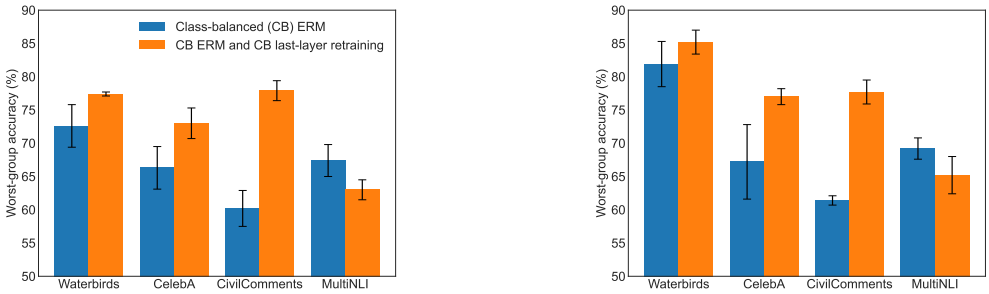


Figure 1: **Class-balanced last-layer retraining renders class-balanced ERM optional.** We compare class-balanced (CB) last-layer retraining on the held-out dataset initialized with unbalanced (UB) or CB ERM features. Contrasting with Idrissi et al. (2022), our results suggest that CB ERM is not necessary for improved group robustness as long as the last layer alone is retrained with class balancing. We list the mean and standard deviation over three independent runs. See Appendix C for detailed results.

In Table 1, we compare unbalanced and class-balanced last-layer retraining on the held-out dataset against DFR and other methods not using group annotations. Class balancing is essential for good performance: on average over the four datasets, class-balanced last layer retraining achieves 94% of DFR worst-group accuracy – competitive with other state-of-the-art techniques – compared to 76% without class balancing. While group balance is still the most important factor in DFR performance, our results suggest that in some cases (namely, Waterbirds and CivilComments) a significant amount of the worst-group accuracy benefit of DFR is solely due to class balancing.



(a) Training dataset only

(b) Combined training and held-out datasets

Figure 2: **A "free lunch" in group robustness.** We compare class-balanced (CB) ERM on the entire dataset to splitting the dataset and performing CB ERM on the first split and CB last-layer retraining on the second. This technique improves worst-group accuracy on Waterbirds, CelebA, and CivilComments by up to 17% despite using *no additional data or annotations*. We believe it underperforms on MultiNLI because there is not enough data in the first split, *i.e.,* ERM performance can still be improved by collecting more data. We searched over four splits, but holding out 5% of data consistently performed well overall. We plot the mean and standard deviation over three independent runs. See Appendix C for detailed results.

Moreover, our experiments in Figure 1 indicate that class-balanced last layer retraining has similar performance regardless of whether it is initialized with class-unbalanced or class-balanced ERM features. Contrasting with Idrissi et al. (2022), this shows that class balancing in the ERM stage is optional. This result has practical relevance for expensive models pre-trained without class balancing (*e.g.,* large language models), as the benefits of class balancing can be reaped by simply retraining the last layer instead of training a new class-balanced ERM model.

4. A "free lunch" in group robustness

Motivated by the promising results of Section 3, where we performed class-balanced last-layer retraining on a fixed held-out set (namely, half the validation set), we now ask *how can we best utilize a realistic training dataset?* In particular, practical applications often work with a predetermined data, annotation, and compute budget. Within this budget, and with no explicit held-out dataset, would one achieve better group robustness by using the entire dataset for ERM or by holding out a subset for last-layer retraining?

We investigate this question on our four benchmark datasets by randomly splitting the initial dataset into two, then performing class-balanced ERM training on the first (larger) split and class-balanced last-layer retraining on the second (smaller) split. Figure 2a illustrates the results of our experiments on the training dataset and Figure 2b on the combined training and held-out datasets. In each case, we perform a hyperparameter search over dataset split proportions: namely $80/20$, $85/15$, $90/10$, and $95/5$ for ERM and last-layer retraining respectively, where the ERM baseline corresponds to a $100/0$ split. With that said, the $95/5$ split consistently performed well overall, and therefore we recommend holding out 5% of data in practice. Due to a larger quantity of data, we expect all numbers to be higher in Figure 2b compared to Figure 2a (especially on Waterbirds, which has a more group-balanced validation set).

Figure 2 indicates that last-layer retraining on a held-out dataset split substantially improves worst group accuracy on Waterbirds, CelebA, and CivilComments. It decreases performance on MultiNLI, which is the only dataset where adding held-out data from the same distribution significantly increases ERM worst-group accuracy compared to DFR. Specifically, class-balanced ERM achieves $67.4 \pm 2.4\%$ worst-group accuracy on the training dataset and $69.2 \pm 1.6\%$ on the combined training and held-out datasets, while our DFR implementation achieves $70.8 \pm 0.8\%$. Therefore, we hypothesize that last-layer retraining on the second split can improve group robustness *only if there is enough data for ERM to perform near-optimally* on the first split, *i.e.,* if the performance of ERM on the first split is limited by dataset bias rather than sample variance.

Based on this hypothesis, our answer to the posed question is: *if ERM performance is stable when holding out 5% of data, perform last-layer retraining on the held-out dataset instead of ERM on the initial dataset*. We call this technique a "free lunch" because it improves worst-group accuracy with *no additional data or annotations* beyond ERM. Therefore, we believe this method is especially relevant to practitioners, and it can be easily implemented with little change to data processing or model training workflows.

A remaining question is *why* last-layer retraining improves group robustness; since the training and held-out datasets have equally drastic group imbalance, it is counterintuitive that reducing the quantity of data used for ERM and performing last-layer retraining can increase worst-group accuracy.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint 1907.02893*, 2019. Cited on page 11.

Beery, S., van Horn, G., and Perona, P. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 2018. Cited on page 1.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of African-American English. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016. Cited on page 11.

Borkan, D., Dixon, L., Sorenson, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *World Wide Web (WWW)*, 2019. Cited on page 3.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2018. Cited on page 11.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2016. Cited on pages 1 and 11.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning (ICML)*, 2021. Cited on page 11.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*, 2019. Cited on pages 3 and 9.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. Cited on page 11.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. Cited on pages 1 and 11.

Goel, K., Gu, A., Li, Y., and Ré, C. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations (ICLR)*, 2021. Cited on page 11.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. Annotation artifacts in natural language inference data. In *North American Association for Computational Linguistics (NAACL)*, 2018. Cited on page 11.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(1):357–362, 2020. Cited on page 10.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018. Cited on page 11.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on pages 3 and 9.

Hovy, D. and Søgaard, A. Tagging performance correlates with author age. In *Association for Computational Linguistics (ACL)*, 2015. Cited on page 11.

Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. Cited on page 10.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2022. Cited on pages 2, 3, 4, 10, and 11.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. Cited on pages 1, 2, 3, 8, 10, and 11.

Kim, E., Lee, J., and Choo, J. BiaSwap: removing dataset bias with bias-tailored swapping augmentation. In *International Conference on Computer Vision (ICCV)*, 2021. Cited on page 11.

Kim, N., Hwang, S., Ahn, S., Park, J., and Kwak, S. Learning debiased classifier with biased committee. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. Cited on page 11.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023. Cited on pages 1, 2, 3, 8, 10, and 11.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021. Cited on pages 1 and 3.

LaBonte, T. Milkshake: Quick and extendable experimentation with classification models. http://github.com/tmlabonte/milkshake, 2023. Cited on page 10.

Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *Journal of Machine Learning Research (JMLR)*, 10(1):777–801, 2009. Cited on page 2.

Lee, Y., Yao, H., and Finn, C. Diversify and disambiguate: Learning from underspecified data. In *International Conference on Learning Representations (ICLR)*, 2023. Cited on page 2.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021. Cited on pages 2, 3, and 11.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. Cited on pages 1, 2, and 11.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. Cited on pages 2 and 10.

maintainers, T. and contributors. TorchVision: PyTorch's computer vision library. *GitHub*, 2016. Cited on page 10.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Association for Computational Linguistics (ACL)*, 2019. Cited on page 11.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: Training debiased classifier from biased classifier. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 11.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations (ICLR)*, 2022. Cited on pages 2, 3, and 11.

Niven, T. and Kao, H.-Y. Probing neural network comprehension of natural language arguments. In *Association for Computational Linguistics (ACL)*, 2019. Cited on page 11.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning for Health*, 2019. Cited on page 11.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. In *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Automatic Differentiation*, 2017. Cited on page 10.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 11.

Qiu, S., Potapczynski, A., Izmailov, P., and Wilson, A. G. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning (ICML)*, 2023. Cited on page 11.

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. The elephant in the room. *arXiv preprint 1808.03305*, 2018. Cited on page 11.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(1):211–252, 2015. Cited on pages 3 and 9.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. Cited on pages 1, 2, 3, 10, and 11.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 1.

Shetty, R., Schiele, B., and Fritz, M. Not using the car to see the sidewalk: Quantifying and controlling the effects of context in classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on page 11.

Singla, S. and Feizi, S. Salient ImageNet: how to discover spurious features in deep learning? In *International Conference on Learning Representations (ICLR)*, 2022. Cited on page 11.

Sohoni, N. S., Dunnmon, J. A., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. Cited on page 11.

Sohoni, N. S., Sanjabi, M., Ballas, N., Grover, A., Nie, S., Firooz, H., and Ré, C. BARACK: partially supervised group robustness with guarantees. In *International Conference on Machine Learning (ICML) Workshop on Spurious Correlations, Invariance, and Stability*, 2022. Cited on page 11.

Tatman, R. Gender and dialect bias in YouTube's automatic captions. In *Association for Computational Linguistics (ACL) Workshop on Ethics in Natural Language Processing*, 2017. Cited on page 11.

Vapnik, V. *Statistical Learning Theory*. Wiley, 1998. Cited on page 1.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. Cited on page 2.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD birds 200. Technical report, California Institute of Technology, 2010. Cited on page 2.

William Falcon and the PyTorch Lightning maintainers and contributors. PyTorch Lightning. *GitHub*, 2019. Cited on page 10.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Association for Computational Linguistics (NAACL)*, 2018. Cited on page 3.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) System Demonstrations*, 2020. Cited on page 10.

Xiao, K. Y., Engstrom, L., Ilyas, A., and Mądry, A. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations (ICLR)*, 2021. Cited on page 11.

Xu, Y., He, H., Shen, T., and Jaakkola, T. Controlling directions orthogonal to a classifier. In *International Conference on Learning Representations (ICLR)*, 2022. Cited on page 11.

Yaghoobzadeh, Y., Mehri, S., Tachet, R., Hazen, T., and Sordoni, A. Increasing robustness to spurious correlations using forgettable examples. In *European Association for Computational Linguistics (EACL)*, 2021. Cited on page 11.

Yang, Y.-Y., Chou, C.-N., and Chaudhuri, K. Understanding rare spurious correlations in neural networks. *arXiv preprint 2202.05189*, 2022. Cited on page 11.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15:e1002683, 2018. Cited on pages 1 and 11.

Zhang, J., Lopez-Paz, D., and Bottou, L. Rich feature construction for the optimization-generalization dilemma. In *International Conference on Machine Learning (ICML)*, 2022a. Cited on page 11.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning (ICML)*, 2022b. Cited on page 11.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *International Conference on Computer Vision (ICCV)*, 2015. Cited on pages 3 and 9.

A. Additional experiments

In this section we include the detailed results of additional experiments and ablations.

Table 2: **Empirical risk minimization.** First, we compare class-unbalanced (CU) and class-balanced (CB) ERM on the training dataset vs. the combined training and held-out datasets (*i.e.,* the training dataset plus half the validation dataset). We list the mean and standard deviation over three independent runs. Note that the Waterbirds validation dataset has a different distribution than the training dataset (and, in particular, is group-balanced), and that MultiNLI is class-balanced *a priori*.

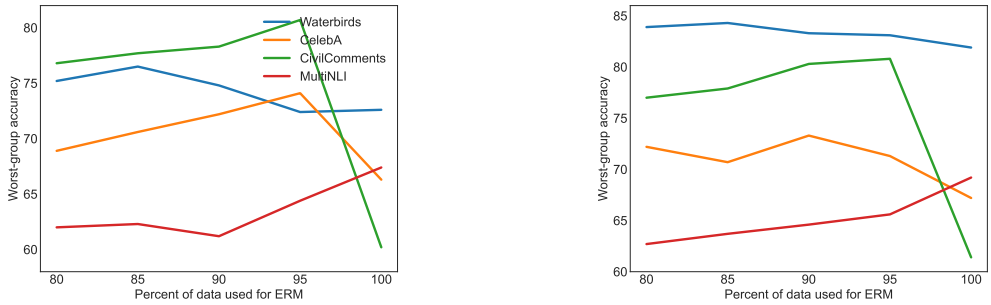
Method	Held-out dataset included	Worst-group test accuracy			
		Waterbirds	CelebA	CivilComments	MultiNLI
CU ERM	✗	$72.4_{\pm 1.0}$	$44.1_{\pm 0.9}$	$63.8_{\pm 6.2}$	$67.4_{\pm 2.4}$
CB ERM	✗	$72.6_{\pm 3.2}$	$66.3_{\pm 3.2}$	$60.2_{\pm 2.7}$	$67.4_{\pm 2.4}$
CU ERM	✓	$81.6_{\pm 1.5}$	$44.5_{\pm 3.4}$	$59.1_{\pm 2.2}$	$69.1_{\pm 1.3}$
CB ERM	✓	$81.9_{\pm 3.4}$	$67.2_{\pm 5.6}$	$61.4_{\pm 0.7}$	$69.2_{\pm 1.6}$

Table 3: **Balancing methodology comparison.** Next, we compare last-layer retraining on the held-out set with different balancing methodologies, each initialized with class-balanced ERM features. In "sampling", we sample from the held-out dataset at a non-uniform rate so that each minibatch is class- or group-balanced in expectation, while in "subset", we train on a random class- or group-balanced subset of the held-out dataset. Specifically, for group-balanced sampling we first sample $s \sim \text{Unif}(\mathcal{S})$, then sample $x \sim \hat{p}(\cdot|s)$ where $\hat{p}$ is the training distribution; class-balanced sampling is the same with $\mathcal{Y}$ instead of $\mathcal{S}$. For subset balancing, we keep all data from the smallest group/class and downsample the others uniformly at random to that size. We list the mean and standard deviation over three independent runs.

Last-layer retraining method	Worst-group test accuracy			
	Waterbirds	CelebA	CivilComments	MultiNLI
Class-unbalanced	$88.0_{\pm 0.8}$	$41.9_{\pm 1.4}$	$57.6_{\pm 4.2}$	$64.6_{\pm 1.0}$
Class-balanced sampling	$92.6_{\pm 0.8}$	$73.7_{\pm 2.8}$	$80.4_{\pm 0.8}$	$64.7_{\pm 1.1}$
Class-balanced subset	$92.1_{\pm 0.9}$	$74.6_{\pm 2.0}$	$80.2_{\pm 1.4}$	$64.5_{\pm 1.3}$
Group-balanced sampling	$92.4_{\pm 0.9}$	$87.0_{\pm 1.1}$	$81.8_{\pm 1.6}$	$70.8_{\pm 0.8}$
Group-balanced subset	$91.6_{\pm 1.9}$	$88.1_{\pm 1.0}$	$79.2_{\pm 1.8}$	$68.3_{\pm 1.8}$
DFR (Kirichenko et al., 2023; Izmailov et al., 2022)	$91.1_{\pm 0.8}$	$89.4_{\pm 0.9}$	$78.8_{\pm 0.5}$	$72.6_{\pm 0.3}$

Table 4: **Necessity of the held-out dataset.** Next, we investigate whether holding out a subset of the training dataset for class-balanced (CB) last-layer retraining is essential, or if retraining on the entire (previously seen) dataset is also effective. For retraining on the training dataset, we use the same class-balanced last-layer retraining procedure as the held-out dataset, but we train for 20 epochs for the vision tasks and 2 epochs for the language tasks. For retraining on the held-out dataset, we report the best over four splits (*i.e.,* the same numbers as Figure 2). Our results suggest that holding out data is necessary to achieve maximal worst-group accuracy, though last-layer retraining on the training dataset interestingly prevents the performance decrease on MultiNLI. We list the mean and standard deviation over three independent runs.

Method	Worst-group test accuracy			
	Waterbirds	CelebA	CivilComments	MultiNLI
CB ERM	$72.6_{\pm 3.2}$	$66.3_{\pm 3.2}$	$60.2_{\pm 2.7}$	$67.4_{\pm 2.4}$
CB last-layer retraining on training dataset	$71.0_{\pm 2.5}$	$66.9_{\pm 1.4}$	$61.9_{\pm 0.8}$	$67.0_{\pm 1.5}$
CB last-layer retraining on held-out dataset	$77.4_{\pm 0.3}$	$73.0_{\pm 2.3}$	$77.9_{\pm 1.5}$	$63.0_{\pm 1.5}$



(a) Training dataset only

(b) Combined training and held-out datasets

Figure 3: **Last-layer retraining dataset split ablation.** Finally, we compare class-balanced (CB) ERM on the entire dataset to splitting the dataset and performing CB ERM on the first split and CB last-layer retraining on the second. We plot worst-group accuracy against the percent of data used for ERM. For example, 90 corresponds to CB ERM on 90% of the dataset and CB last-layer retraining on the remaining 10%, while 100 corresponds to CB ERM on the entire dataset. This technique helps on all datasets except MultiNLI – we hypothesize this is because there is not enough data for ERM to perform statistically optimally. We recommend a $95/5$ split in practice, as it does consistently well in our experiments, performing the best or second-best on each dataset besides Waterbirds. We plot the mean over three independent runs and leave out error bars for readability.

B. Dataset composition and training details

The detailed composition of the four benchmark datasets we study is listed in Table 5.

Table 5: **Dataset composition.** We study four well-established benchmarks for group robustness across vision and language tasks. The class probabilities change dramatically when conditioned on the spurious feature. Note that Waterbirds is the only dataset that has a distribution shift (in particular, the validation dataset is group-balanced conditioned on the classes) and MultiNLI is the only dataset which is class-balanced *a priori*. Probabilities may not sum to 1 due to rounding.

Dataset	Group g		Training distribution $\hat{p}$			Data quantity		
	Class y	Spurious s	$\hat{p}(y)$	$\hat{p}(g)$	$\hat{p}(y\|s)$	Train	Val	Test
Waterbirds	landbird	land	.768	.730	.984	3498	467	2225
	landbird	water		.038	.148	184	466	2225
	waterbird	land	.232	.012	.016	56	133	642
	waterbird	water		.220	.852	1057	133	642
CelebA	non-blond	female	.851	.440	.758	71629	8535	9767
	non-blond	male		.411	.980	66874	8276	7535
	blond	female	.149	.141	.242	22880	2874	2480
	blond	male		.009	.020	1387	182	180
CivilComments	neutral	no identity	.887	.551	.921	148186	25159	74780
	neutral	identity		.336	.836	90337	14966	43778
	toxic	no identity	.113	.047	.079	12731	2111	6455
	toxic	identity		.066	.164	17784	2944	8769
MultiNLI	contradiction	no negation	.333	.279	.300	57498	22814	34597
	contradiction	negation		.054	.761	11158	4634	6655
	entailment	no negation	.334	.327	.352	67376	26949	40496
	entailment	negation		.007	.104	1521	613	886
	neither	no negation	.333	.323	.348	66630	26655	39930
	neither	negation		.010	.136	1992	797	1148

We utilize a ResNet-50 (He et al., 2016) pretrained on ImageNet-1K (Russakovsky et al., 2015) for the vision tasks and a BERT (Devlin et al., 2019) model pretrained on Book Corpus (Zhu et al., 2015) and English Wikipedia for the language tasks. These pretrained models are used as the initialization for ERM on the four datasets we study. We use standard ImageNet normalization with standard flip and crop data augmentation for the vision tasks, and BERT tokenization for the language

tasks. Our implementation uses the following packages: `NumPy` (Harris et al., 2020), `PyTorch` (Paszke et al., 2017), `Lightning` (William Falcon and the PyTorch Lightning maintainers and contributors, 2019), `TorchVision` (maintainers & contributors, 2016), `Matplotlib` (Hunter, 2007), `Transformers` (Wolf et al., 2020), and `Milkshake` (LaBonte, 2023).

For ERM and last-layer retraining, we do not vary any hyperparameters; their fixed values are listed in Table 6. For the dataset split experiments (Section 4), we vary the splits among $80/20$, $85/15$, $90/10$, and $95/5$ for ERM and last-layer retraining respectively, where the ERM baseline corresponds to a $100/0$ split. With that said, the $95/5$ split did consistently well, and therefore we recommend holding out 5% of data in practice; see Appendix A for our ablation study.

Table 6: **ERM and last-layer retraining hyperparameters.** We use standard hyperparameters following previous work (Sagawa et al., 2020; Idrissi et al., 2022; Kirichenko et al., 2023; Izmailov et al., 2022). For last-layer retraining, we keep all hyperparameters the same except the number of epochs on CelebA, which we increase to 100.

Dataset	Optimizer	Initial LR	LR schedule	Batch size	Weight decay	Epochs
Waterbirds	SGD	3×10^{-3}	Cosine	32	1×10^{-4}	100
CelebA	SGD	3×10^{-3}	Cosine	100	1×10^{-4}	20
CivilComments	AdamW (Loshchilov & Hutter, 2019)	1×10^{-5}	Linear	16	1×10^{-4}	10
MultiNLI	AdamW (Loshchilov & Hutter, 2019)	1×10^{-5}	Linear	16	1×10^{-4}	10

C. Additional tables

In this section, we provide detailed results for Figures 1 and 2.

Table 7: **Table for Figure 1.** We compare different combinations of ERM and last-layer retraining with or without class balancing. Notably, class-balanced last-layer retraining enables nearly the same worst-group accuracy whether the ERM incorporates class-balancing or not. In Figure 1, we only plot the results which use class-balanced last layer retraining (*i.e.,* the last two rows of the table). We use the entire held-out set for last-layer retraining, and we list the mean and standard deviation over three independent runs.

ERM class balancing	Last-layer retraining class balancing	Worst-group test accuracy			
		Waterbirds	CelebA	CivilComments	MultiNLI
✗	✗	$87.5_{\pm 0.7}$	$45.4_{\pm 2.3}$	$54.7_{\pm 5.8}$	$64.5_{\pm 0.8}$
✓	✗	$88.0_{\pm 0.8}$	$41.9_{\pm 1.4}$	$57.6_{\pm 4.2}$	$64.6_{\pm 1.0}$
✗	✓	$92.3_{\pm 0.4}$	$71.1_{\pm 2.4}$	$78.4_{\pm 2.5}$	$64.7_{\pm 1.1}$
✓	✓	$92.6_{\pm 0.8}$	$73.7_{\pm 2.8}$	$80.4_{\pm 0.8}$	$64.7_{\pm 1.1}$

Table 8: **Table for Figure 2.** We compare class-balanced (CB) ERM on the entire dataset to splitting the dataset and performing CB ERM on the first (larger) split and CB last-layer retraining on the second (smaller) split. The results with and without the held-out dataset correspond to Figure 2a and 2b respectively. We search over four splits and report the best with respect to worst-group validation accuracy; see Figure 3 for results on each split. We list the mean and standard deviation over three independent runs.

Method	Held-out dataset included	Worst-group test accuracy			
		Waterbirds	CelebA	CivilComments	MultiNLI
CB ERM	✗	$72.6_{\pm 3.2}$	$66.3_{\pm 3.2}$	$60.2_{\pm 2.7}$	$67.4_{\pm 2.4}$
CB last-layer retraining	✗	$77.4_{\pm 0.3}$	$73.0_{\pm 2.3}$	$77.9_{\pm 1.5}$	$63.0_{\pm 1.5}$
CB ERM	✓	$81.9_{\pm 3.4}$	$67.2_{\pm 5.6}$	$61.4_{\pm 0.7}$	$69.2_{\pm 1.6}$
CB last-layer retraining	✓	$85.2_{\pm 1.8}$	$77.0_{\pm 1.2}$	$77.7_{\pm 1.8}$	$65.2_{\pm 2.8}$

D. Additional related work

Spurious correlations. The performance of empirical risk minimization (ERM) in the presence of spurious correlations has been extensively studied (Geirhos et al., 2020). In vision, ERM models are widely known to rely on spurious attributes like background (Sagawa et al., 2020; Xiao et al., 2021), texture (Geirhos et al., 2019), and secondary objects (Rosenfeld et al., 2018; Shetty et al., 2019; Singla & Feizi, 2022) to perform classification. In language, ERM models often utilize syntactic or statistical heuristics as a substitute for semantic understanding (Gururangan et al., 2018; Niven & Kao, 2019; McCoy et al., 2019). This behavior can lead to bias against demographic minorities (Hovy & Søgaard, 2015; Blodgett et al., 2016; Tatman, 2017; Hashimoto et al., 2018; Buolamwini & Gebru, 2018) or failure in high-consequence applications (Liu et al., 2015; Chouldechova, 2016; Zech et al., 2018; Oakden-Rayner et al., 2019).

Robustness and group annotations. If group annotations are available in the training dataset, *group distributionally robust optimization* (DRO) (Sagawa et al., 2020) can improve robustness by minimizing the worst-group loss, while other techniques learn invariant or diverse features (Arjovsky et al., 2019; Goel et al., 2021; Zhang et al., 2022a; Xu et al., 2022). Methods which use only partial group annotations include *deep feature reweighting* (DFR) (Kirichenko et al., 2023; Izmailov et al., 2022), which retrains the last layer on a group-balanced held-out set, and *spread spurious attribute* (Nam et al., 2022), which performs DRO with group pseudo-labels. However, since the groups are often unknown ahead of time or difficult to annotate in practice, there has been significant interest in methods which do not utilize group annotations except for model selection. The bulk of these techniques train one or more auxiliary models to pseudo-label the minority group (Sohoni et al., 2020; Nam et al., 2020; Yaghoobzadeh et al., 2021; Creager et al., 2021; Kim et al., 2021; 2022; Sohoni et al., 2022; Zhang et al., 2022b); notably, *just train twice* (JTT) (Liu et al., 2021) upweights samples misclassified by an early-stopped model. Other techniques reweight or subsample the classes (Idrissi et al., 2022) or train with robust losses and regularization (Pezeshki et al., 2021; Yang et al., 2022).

Unlike DFR, our methods utilize no held-out group annotations, and unlike JTT, we do not use an auxiliary model to identify the minority group, but show that class-balanced last-layer retraining improves worst-group accuracy without explicit group pseudo-labels. The concurrent work of Qiu *et al.* (Qiu et al., 2023) proposed a similar method to ours; while they use a tunable loss to upweight misclassified samples, our method shows that no tuning or upweighting is required to achieve similar results. Finally, while our results partially corroborate the findings of Idrissi et al. (2022) that class balancing during ERM is effective for group robustness, we observe that class-balanced last-layer retraining renders class balancing in the first ERM stage optional (see Figure 1). We compare our results with previous methods in Table 1.