The Group Robustness is in the Details: Revisiting Finetuning under Spurious Correlations

Tyler LaBonte[1]* John C. Hill[2] Xinchen Zhang[3]
Vidya Muthukumar[3,1] Abhishek Kumar†

[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
[2]School of Computer Science, Georgia Institute of Technology
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology

Abstract

Modern machine learning models are prone to over-reliance on spurious correlations, which can often lead to poor performance on minority groups. In this paper, we identify surprising and nuanced behavior of finetuned models on worst-group accuracy via comprehensive experiments on four well-established benchmarks across vision and language tasks. We first show that the commonly used class-balancing techniques of mini-batch upsampling and loss upweighting can induce a decrease in worst-group accuracy (WGA) with training epochs, leading to performance no better than without class-balancing. While in some scenarios, removing data to create a class-balanced subset is more effective, we show this depends on group structure and propose a mixture method which can outperform both techniques. Next, we show that scaling pretrained models is generally beneficial for worst-group accuracy, but only in conjuction with appropriate class-balancing. Finally, we identify spectral imbalance in finetuning features as a potential source of group disparities — minority group covariance matrices incur a larger spectral norm than majority groups once conditioned on the classes. Our results show more nuanced interactions of modern finetuned models with group robustness than was previously known. Our code is available at https://github.com/tmlabonte/revisiting-finetuning.

1 Introduction

Classification performance in machine learning is sensitive to *spurious correlations*: patterns which are predictive of the target class in the training dataset but not at test time. For example, in computer vision tasks, neural networks are known to utilize the backgrounds of images as proxies for their content (Beery et al., 2018; Sagawa et al., 2020a; Xiao et al., 2021). Beyond simple settings, spurious correlations have been identified in high-consequence applications such as criminal justice (Chouldechova, 2016), medicine (Zech et al., 2018), and facial recognition (Z. Liu et al., 2015). In particular, a model's reliance on spurious correlations disproportionately affects its accuracy on *minority groups* which are under-represented in the training dataset; we therefore desire maximizing the model's *group robustness*, quantified by its minimum accuracy on any group (Sagawa et al., 2020a).

The standard workflow in modern machine learning involves initializing from a pretrained model and finetuning on the downstream dataset using empirical risk minimization (ERM) (Vapnik, 1998), which minimizes the average training loss. When *group annotations* are available in the training dataset, practitioners utilize a rich literature of techniques to improve worst-group accuracy (WGA) (Sagawa et al., 2020a; Nam et al., 2022; Kirichenko et al., 2023). However, group annotations are often unknown or problematic to obtain (*e.g.,* due to financial, privacy, or fairness concerns). While group robustness methods have been adapted to work without group annotations (E. Z. Liu et al., 2021; M. Zhang et al., 2022; Qiu et al., 2023; LaBonte et al., 2023), they remain complex variants on the standard finetuning procedure. Hence, it is often unclear to what extent the WGA dynamics of these methods are attributable to details of model finetuning.

In this paper, we take an orthogonal approach to the literature focused on methodology to improve group robustness, and instead pursue a comprehensive understanding of the *fundamental properties* of model

* Corresponding author. Email: tlabonte@gatech.edu.
† Work done at Google DeepMind.

finetuning on four well-established group robustness benchmarks across vision and language tasks. We focus especially on the conjunction of *model scaling* and *class-balancing* — which was recently shown to greatly improve robustness on some datasets (Idrissi et al., 2022) — on the worst-group accuracy of the ERM baseline. These considerations enable us to isolate the impact of group disparities on worst-group accuracy, thereby revealing far more nuanced behaviors of finetuned models than previously known. In particular, we challenge overarching narratives that "overparameterization helps or hurts distributional robustness" and show striking differences in finetuning performance depending on the class-balancing methodology.

In more detail, our main contributions include:

- Identifying two *failure modes* of common class-balancing techniques during finetuning: (1) mini-batch upsampling and loss upweighting experience catastrophic collapse with standard hyperparameters on benchmark datasets, and (2) removing data to create a class-balanced subset can harm WGA for certain datasets.
- Proposing a *mixture balancing* method which combines the advantages of two class-balancing techniques and can improve baseline WGA beyond either method.
- Showing that while overparameterization can harm WGA in certain cases, model scaling is generally beneficial for robustness when applied *in conjunction* with appropriate pretraining and class-balancing.
- Identifying a *spectral imbalance* in the top eigenvalues of the group covariance matrices — even when the classes are balanced — and showing that minority group covariance matrices consistently have larger spectral norm once conditioned on the classes.

1.1 Related work

Here we provide a brief summary of related work along three axes. Throughout the paper, we also provide detailed contextualizations of our results with the most closely related work.

Spurious correlations. The proclivity of empirical risk minimization (ERM) to rely on spurious correlations has been widely studied (Geirhos et al., 2020; Moayeri et al., 2023). Rectifying this weakness is an important challenge for real-world deployment of machine learning algorithms, as spurious correlations can exacerbate unintended bias against demographic minorities (Hovy and Søgaard, 2015; Blodgett et al., 2016; Tatman, 2017; Hashimoto et al., 2018; Buolamwini and Gebru, 2018) or cause failure in high-consequence applications (Z. Liu et al., 2015; Chouldechova, 2016; Zech et al., 2018; Oakden-Rayner et al., 2019). Reliance on spurious correlations manifests in image datasets as the usage of visual shortcuts including background (Beery et al., 2018; Sagawa et al., 2020a; Xiao et al., 2021), texture (Geirhos et al., 2019), and secondary objects (Rosenfeld et al., 2018; Shetty et al., 2019; Singla and Feizi, 2022), and in text datasets as the usage of syntactic or statistical heuristics as a substitute for semantic understanding (Gururangan et al., 2018; Niven and Kao, 2019; McCoy et al., 2019).

Class-balancing and group robustness. *Group-balancing*, or training with an equal number of samples from each group, has been proposed as a simple yet effective method to improve robustness to spurious correlations (Hashimoto et al., 2018; Sagawa et al., 2020b; Chatterji et al., 2023; Stromberg et al., 2024). However, group-balancing requires group annotations, which are often unknown or problematic to obtain (E. Z. Liu et al., 2021; M. Zhang et al., 2022; Qiu et al., 2023; LaBonte et al., 2023). On the other hand, *class-balancing*, or training with an equal number of samples from each class, is a well-studied method in long-tailed classification (Japkowicz and Stephen, 2002; Haixiang et al., 2017; Buda et al., 2018). Recent work has shown that class-balancing is a surprisingly powerful method for improving worst-group accuracy which does not require group annotations (Idrissi et al., 2022; LaBonte et al., 2023; Chaudhuri et al., 2023; Shwartz-Ziv

et al., 2023). In particular, Idrissi et al., 2022 study the WGA dynamics of two common class-balancing methods: removing data from the larger classes (which we call *subsetting*) and upsampling the smaller classes (which we call *upsampling*). Our results complement those of Idrissi et al., 2022 and show more nuanced effects of class-balancing than previously known; we provide additional contextualization with Idrissi et al., 2022 in Section 3.1. We show similar nuanced behavior of *upweighting* smaller classes in the loss function, a popular method in the group-balancing setting (E. Z. Liu et al., 2021; Qiu et al., 2023; Stromberg et al., 2024) which Idrissi et al., 2022 did not study.

Overparameterization and distributional robustness. While the accepted empirical wisdom is that overparameterization improves *in-distribution* test accuracy (Neyshabur et al., 2014; C. Zhang et al., 2021), the relationship between overparameterization and robustness is only incompletely understood. Sagawa et al., 2020b considered a special class of ResNet-18 architectures and showed that increasing the *width* of the ResNet-18 model reduces worst-group accuracy on the Waterbirds and CelebA datasets when trained with ERM (without any class-balancing) — this contrasts with the improvement in average accuracy with width that is widely observed in practice (see, *e.g.,* Nakkiran et al., 2021). Conversely, Hendrycks et al., 2021 unearthed a benefit of overparameterization in robustness to "natural" covariate shifts induced in computer vision, which are quite different from the distribution shifts induced by unequal representations in groups (Koh et al., 2021). On the mathematical front, Tripuraneni et al., 2021; Maity et al., 2022 showed that overparameterization in random feature models that are trained to completion improves robustness to a wide class of covariate shifts, including those induced by shifts in proportions of groups. However, both the optimization trajectory and statistical properties of random features are very different from those of neural networks (see, *e.g.,* Ghorbani et al., 2019). Closely related to our work, Pham et al., 2021 investigated pretrained ResNet and VGG models for vision, and BERT for language, and showed that in fact overparameterization can either improve WGA or leave it unchanged. Our results complement those of Pham et al., 2021 with a richer setup and show that class-balancing — which they do not study — can greatly impact model scaling behavior. We provide additional contextualization with Pham et al., 2021 in Section 4.

2 Preliminaries

Setting. We consider classification tasks with input domain $\mathbb{R}^n$ and target classes $\mathcal{Y} \subset \mathbb{N}$. Suppose $\mathcal{S}$ is a set of *spurious features* such that each example $\boldsymbol{x} \in \mathbb{R}^n$ is associated with exactly one feature $s(\boldsymbol{x}) \in \mathcal{S}$. The dataset is then partitioned into *groups* $\mathcal{G}$, defined by the Cartesian product of classes and spurious features $\mathcal{G} = \mathcal{Y} \times \mathcal{S}$. Given a dataset of m training examples, we define the set of indices of examples which belong to some group $g \in \mathcal{G}$ or class $y \in \mathcal{Y}$ by $\Omega_g \subseteq \{1, \ldots, m\}$ and $\Omega_y \subseteq \{1, \ldots, m\}$, respectively. Then, the *majority group(s)* is defined by the group(s) that maximize $|\Omega_g|$. All other groups are designated as *minority groups*. Further, the *worst group(s)*[1] is defined by the group(s) which incur minimal test accuracy. We define majority and minority classes similarly. It is worth noting that, because groups are defined by the Cartesian product of classes and spurious features, that all training examples in a particular group are identically labeled, and therefore *a group is a subset of a class*. We will frequently invoke this terminology throughout the paper.

We desire a model which, despite group imbalance in the training dataset, enjoys roughly uniform performance over $\mathcal{G}$. Therefore, we evaluate *worst-group accuracy* (WGA), *i.e.,* the minimum accuracy among all groups (Sagawa et al., 2020a). We will also be interested in the relative performance on groups *within the same class*, and we thereby define the *majority group within a class* $y \in \mathcal{Y}$ as the group which maximizes $|\Omega_g|$ over all $g \in \{g \in \mathcal{G} : y \in g\}$. Other groups are designated as the minority groups within that class. For example, referring to the Waterbirds section of Table 2, groups 1 and 2 are the minority groups within classes 0 and 1, respectively.

[1] Note that, as is standard in the empirical literature on distributional robustness, majority, minority and worst groups are defined with respect to the empirical training distribution, as this is all that we have access to. Moreover, test accuracy is typically maximized by the majority group and minimized by a minority group, though this is not always the case.

Class-balancing. A dataset is considered to be *class-balanced* if it is composed of an equal number of training examples from each class in $\mathcal{Y}$ in expectation over the sampling probabilities. We primarily compare three class-balancing techniques: *subsetting*, *upsampling*, and *upweighting*. We describe each briefly below:

- In *subsetting*, every class is set to the same size as the smallest class by removing the appropriate amount of data from each larger class uniformly at random. This procedure is performed only once, and the subset is fixed prior to training.
- In *upsampling*, the entire dataset is utilized for training with a typical stochastic optimization algorithm, but the sampling probabilities of each class are adjusted so that mini-batches are class-balanced in expectation. Specifically, to draw a single example, we first sample $y \sim \text{Unif}(\mathcal{Y})$, then sample $\boldsymbol{x} \sim \hat{p}(\cdot \mid y)$ where $\hat{p}$ is the *empirical* distribution on training examples.
- In *upweighting*, the minority class samples are directly upweighted in the loss function according to the ratio of majority class data to minority class data, called the *class-imbalance ratio*. Specifically, if the loss function is $\ell(f(\boldsymbol{x}), y)$ for model f, example $\boldsymbol{x}$, and class label y, the upweighted loss function is $\gamma\ell(f(\boldsymbol{x}), y)$ where γ is defined as the class-imbalance ratio for minority class data and 1 for majority class data. It is worth noting that upweighting is equivalent to upsampling in expectation over the sampling probabilities.

Note that the terminology for these class-balancing techniques is not consistent across the literature. For example, Idrissi et al., 2022 call subsetting *subsampling* (denoted SUBY) and upsampling *reweighting* (denoted RWY). On the other hand, Stromberg et al., 2024 call (group-wise) subsetting *downsampling* and use *upweighting* to describe increasing the weight of minority group samples in the loss function. We have chosen our terminology to be as descriptive as possible without conflating commonly overloaded terms such as "sampling" and "weighting".

Datasets and models. We study four classification datasets, two in the vision domain and two in the language domain, which are well-established as benchmarks for group robustness. We summarize each dataset below and provide additional numerical details in Appendix A.

- *Waterbirds* (Welinder et al., 2010; Wah et al., 2011; Sagawa et al., 2020a) is an image dataset where birds are classified as land species ("landbirds") or water species ("waterbirds"). The spurious feature is the image background: more landbirds are present on land backgrounds and vice versa.[2]
- *CelebA* (Z. Liu et al., 2015; Sagawa et al., 2020a) is an image dataset where celebrities are classified as blond or non-blond. The spurious feature is gender, with 16× more blond women than blond men in the training dataset.
- *CivilComments* (Borkan et al., 2019; Koh et al., 2021) is a language dataset where online comments are classified as toxic or non-toxic. The spurious feature is the presence of one of the following categories: male, female, LGBT, black, white, Christian, Muslim, or other religion.[3] More toxic comments contain one of these categories than non-toxic comments, and vice versa.
- *MultiNLI* (Williams et al., 2018; Sagawa et al., 2020a) is a language dataset where pairs of sentences are classified as 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.

[2] We note that the Waterbirds dataset is known to contain incorrect labels (Taghanaki et al., 2022). We report results on the original, un-corrected version as is standard in the literature.

[3] This version of CivilComments has four groups, used in this work and by Sagawa et al., 2020a; Idrissi et al., 2022; Izmailov et al., 2022; Kirichenko et al., 2023; LaBonte et al., 2023. There is another version where the identity categories are not collapsed into one spurious feature; that version is used by E. Z. Liu et al., 2021 and M. Zhang et al., 2022. Both versions use the WILDS split (Koh et al., 2021).

Waterbirds is class-imbalanced with a majority/minority class ratio of 3.31:1, CelebA a ratio of 5.71:1, and CivilComments a ratio of 7.85:1. MultiNLI is class-balanced *a priori*. Since the Waterbirds dataset has a shift in group proportion from train to test, we weight the group accuracies by their proportions in the training set when reporting the test average accuracy (Sagawa et al., 2020a).

We utilize a ResNet (He et al., 2016) and ConvNeXt-V2 (Woo et al., 2023) model 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 the AdamW optimizer (Loshchilov and Hutter, 2019) for finetuning on three independent seeds, randomizing both mini-batch order and any other stochastic procedure such as subsetting, and we report error bars corresponding to one standard deviation. We do not utilize early-stopping: instead, to consider the impact of overparameterization in a holistic way, we train models to completion to properly measure the overfitting effect.[4] This can result in longer training than commonly seen in the literature (*e.g.,* we finetune on CelebA for about 3× more gradient steps than is standard). See Appendix B for further training details.

3 Nuanced effects of class-balancing on group robustness

We now present our first set of results, which shows that the choice of class balancing method greatly impacts the group robustness of the ERM baseline.

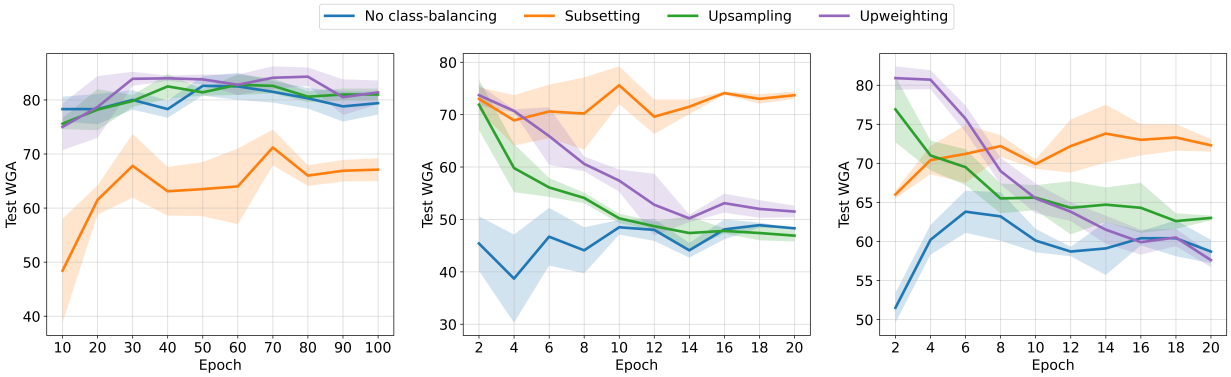
3.1 Catastrophic collapse of class-balanced upsampling and upweighting

In a recent paper, LaBonte et al., 2023 observed that contrary to the central hypothesis underlying the Just Train Twice method (E. Z. Liu et al., 2021), the worst-group accuracy of ERM decreases dramatically with training epochs on CelebA and CivilComments; however, they provide no explanation for this phenomenon. In this section, we show that this degradation of WGA is due to their choice of class-balancing method (*i.e.,* upsampling). Specifically, ERM finetuned with upsampling experiences a *catastrophic collapse* in test WGA over the course of training, a phenomenon that was previously only noticed in synthetic datasets with a linear classifier (Idrissi et al., 2022). Moreover, while Idrissi et al., 2022 state that class-balanced subsetting is not recommended in practice, we show that it can in fact improve WGA conditional on the presence of a small minority group *within the majority class*. Finally, we show that class-balanced upweighting — a popular technique which Idrissi et al., 2022 do not study — experiences a similar WGA collapse as upsampling.

We finetune a ConvNeXt-V2 Base on Waterbirds and CelebA and a BERT Base on CivilComments, and we compare the subsetting, upsampling, and upweighting techniques to a class-imbalanced baseline. The results of our experiments are displayed in Figure 1. On CelebA and CivilComments, the more class-imbalanced datasets, upsampling and upweighting both experience catastrophic collapse over the course of training and reduces WGA to no better than without class-balancing. We believe this collapse is caused by overfitting to the minority group data within the minority class, as any individual point from this group is sampled far more often during upsampling and weighted far more heavily during upweighting, causing overfitting during long training runs. In fact, upsampling does even worse on CelebA than observed in LaBonte et al., 2023 because we train 3× longer to ensure convergence. With that said, optimally tuned early-stopping appears to mitigate the collapse (as previously noticed by Idrissi et al., 2022 in a toy setting). While we focus on training models to completion, optimal early-stopping could be an interesting direction for future research.

Our experiments also highlight a previously unnoticed disadvantage of class-balanced subsetting: if there is a small minority group present in the majority class, subsetting will further reduce the minority group proportion and cause WGA to decrease. For example, in the Waterbirds dataset, the type of bird (landbirds/waterbirds) is the class label and the background (land/water) is the spurious feature. In this dataset, *landbirds/water* is a small minority group within the majority class (*landbirds*). When the landbirds class is cut by 3.31×, the landbirds/water group greatly suffers, harming WGA. On the other hand, in the CelebA dataset, the hair color (non-blond/blond) is the class label and the gender (female/male) is

[4] To be more specific, we finetune ConvNeXt-V2 Base roughly to a training loss of 10^{-4} on Waterbirds and 10^{-3} on CelebA, and BERT Base roughly to a training loss of 10^{-3} on CivilComments and 10^{-2} on MultiNLI.



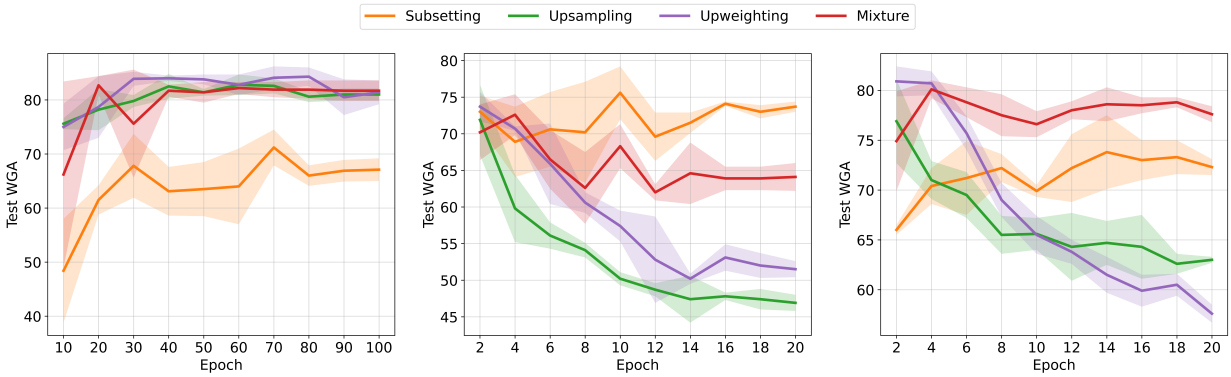
(a) Waterbirds (b) CelebA (c) CivilComments

Figure 1: **Class-balanced upsampling and upweighting experience catastrophic collapse in more imbalanced datasets.** We compare *subsetting*, where data is removed to set every class to the same size as the smallest class, *upsampling*, where the sampling probabilities of each class are adjusted so that the mini-batches are class-balanced in expectation, and *upweighting*, where the loss for the smaller classes is scaled by the class-imbalance ratio. We observe a catastrophic collapse over the course of training of upsampling and upweighting on CelebA and CivilComments, the more class-imbalanced datasets. On the other hand, subsetting reduces WGA on Waterbirds because it removes data from the small minority group within the majority class. MultiNLI is class-balanced *a priori*, so we do not include it here.

the spurious feature. In this dataset, the only small minority group is *blond/male*, while the majority and minority groups are nearly balanced in the majority class. In this case, subsetting preserves all blond/male examples and greatly increases their proportion, helping WGA.

Finally, it is interesting to note that while upsampling and upweighting have similar WGA dynamics, both differ greatly from subsetting. The similarity between upsampling and upweighting is to be expected, as they are equivalent in expectation *only* over the randomness in the sampling mechanism for upsampling on any training dataset. Recently, Stromberg et al., 2024 proved a broader theoretical equivalence between subsetting and upsampling of the *groups* in what is commonly called the *population* setting in optimization theory, *i.e.,* assuming access to the training data distribution or, equivalently, an infinite number of training examples. The further equivalence of upsampling and upweighting that we have outlined above would then imply that all three balanced objectives are optimized by the same solution. Our results suggest that a similar equivalence may not hold in the real-world *empirical* setting, where subsetting has distinctly different behavior from upsampling and upweighting, and the number of training examples may in fact be comparable or even smaller than the number of model parameters. As previously mentioned, this may be due to overfitting to a small number of minority class data repeated often during training; we believe theoretically investigating this discrepancy is an interesting future direction.

Contextualization with previous work. Our observations explain the decrease in WGA of CelebA and CivilComments noticed by LaBonte et al., 2023, a phenomenon which they left unresolved. Our result implies that group robustness methods which assume that WGA increases during training, such as Just Train Twice (E. Z. Liu et al., 2021), may only be justified when appropriate class-balancing is applied. Idrissi et al., 2022 show that upsampling can cause catastrophic collapse in WGA, but only in a synthetic dataset with a linear classifier. In realistic datasets, Idrissi et al., 2022 perform extensive hyperparameter tuning (using group labels, which may not always be available) to achieve good results with upsampling, while we show that catastrophic collapse can occur in the same datasets when standard hyperparameters are used. Moreover, Idrissi et al., 2022 state that class-balanced subsetting is not recommended in practice, but we show that subsetting can be effective except when there is a small minority group within the majority class,



(a) Waterbirds (b) CelebA (c) CivilComments

Figure 2: **Mixture balancing mitigates catastrophic collapse of upsampling and upweighting.** We propose a class-balanced *mixture method*, which combines subsetting and upsampling by first drawing a class-imbalanced subset uniformly at random from the dataset, then adjusting sampling probabilities so that mini-batches are balanced in expectation. Our method increases exposure to majority class data without over-sampling the minority class. Remarkably, mixture balancing outperforms all three class-balancing methods on Waterbirds and CivilComments, and while it does not outperform subsetting on CelebA, it significantly alleviates the WGA collapse experienced by upsampling and upweighting. MultiNLI is class-balanced *a priori*, so we do not include it here.

a previously unnoticed nuance. Finally, we show that subsetting experiences different WGA dynamics from upsampling and upweighting in the empirical setting, suggesting additional complexity compared to the population setting equivalence results of Stromberg et al., 2024.

Without extensive hyperparameter tuning, class-balanced *upsampling* and *upweighting* can induce WGA no better than without class-balancing. While class-balanced *subsetting* can improve WGA, practitioners should use caution if a small minority group is present within the majority class.

3.2 Mixture balancing: interpolating between subsetting and upsampling

To mitigate the catastrophic collapse of class-balanced upsampling and upweighting, we propose a simple *mixture method* which interpolates between subsetting and upsampling. Our method increases exposure to majority class data without over-sampling the minority class, which can improve WGA and prevent overfitting to the minority group. In our mixture method, we first create a data subset with a specified *class-imbalance ratio* by removing data from the larger classes uniformly at random until the desired (smaller) ratio is achieved. Next, we perform ERM finetuning on this subset by adjusting sampling probabilities so that mini-batches are balanced in expectation. Clearly, using a class-imbalance ratio of 1:1 reduces to subsetting, and using the original class-imbalance ratio reduces to upsampling.

We finetune ConvNeXt-V2 Base on Waterbirds and CelebA and BERT Base on CivilComments, and we compare our class-balanced mixture method to the subsetting, upsampling, and upweighting techniques. The results of our experiments are displayed in Figure 2. We plot the performance of our mixture method with the best class-imbalance ratio during validation; an ablation study varying the ratio is included in Appendix C. Remarkably, mixture balancing outperforms both all three class-balancing methods on Waterbirds and CivilComments, and while it does not outperform subsetting on CelebA, it significantly alleviates the WGA collapse experienced by upsampling.

Next, we perform an ablation of the necessity of subsetting in mixture balancing. We compare our method with an alternative implementation that eschews subsetting entirely, instead adjusting the sampling

Table 1: **Mixture balancing is robust to model selection without group annotations.** We compare the best class-balancing method during validation with and without group annotations in the validation set. We show that both worst-class accuracy (Yang et al., 2023) and the bias-unsupervised validation score of Tsirigotis et al., 2023 are effective for model selection without group annotations, often selecting the same method or mixture ratio as worst-group accuracy (WGA) validation. For each dataset, we list the method which maximizes each validation metric as well as its average WGA over three seeds.

Validation Metric	Group Anns	Waterbirds	CelebA	CivilComments
Bias-unsupervised Score	✗	Upsampling (79.9)	Subsetting (74.1)	Mixture 3:1 (77.6)
Worst-class Accuracy	✗	Mixture 2:1 (81.1)	Subsetting (74.1)	Mixture 3:1 (77.6)
Worst-group Accuracy	✓	Mixture 2:1 (81.1)	Subsetting (74.1)	Mixture 3:1 (77.6)

probabilities of each class so that the mini-batches have a particular class ratio in expectation. For example, instead of performing upsampling on a 2:1 class-imbalanced subset, we simply upsample the majority class by a ratio of 2:1 on the entire dataset. The results of our ablation are included in Appendix C; our mixture method outperforms this alternative because the alternative incompletely corrects for class imbalance, therefore suffering from from similar collapse behavior as upsampling and upweighting.

Note on validation. In Figure 2, we plot the best class-imbalance ratio achieved using validation on a group annotated held-out set. While this is a common assumption in the literature (Sagawa et al., 2020a; E. Z. Liu et al., 2021; Izmailov et al., 2022; Kirichenko et al., 2023), it is nevertheless unrealistic when the training set does not have any group annotations. Therefore, we compare with both worst-class accuracy (Yang et al., 2023) and the *bias-unsupervised validation score* of Tsirigotis et al., 2023, which do not use any group annotations for model selection, and in Table 1 we list the method which maximizes each validation metric as well as its average WGA over three seeds.

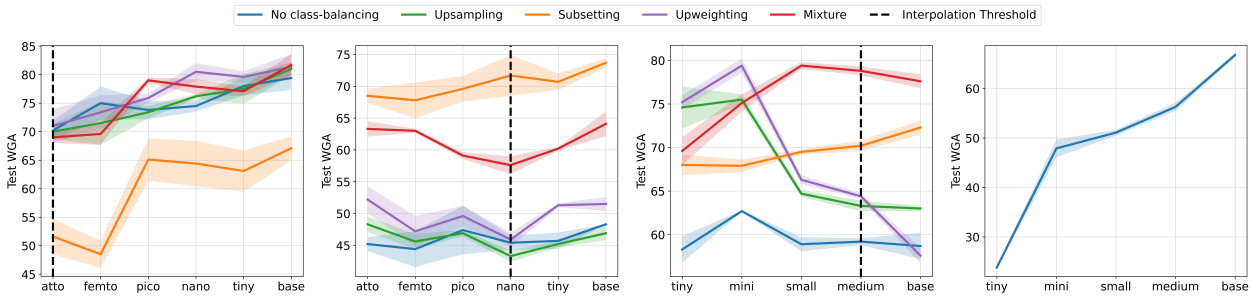
The bias-unsupervised validation score uses a biased model (in our case, the original ERM model finetuned without class-balancing) to construct a confusion matrix whose average accuracy is a proxy for worst-group accuracy. Let K denote the number of classes and y denote the true class label for a particular datapoint x. Let $\hat{y}_{\text{bias}}$ represent the biased model prediction while $\hat{y}$ represents the class-balanced model prediction. Denote by $S_{y,\hat{y}_{\text{bias}}}$ the validation datapoints which have true label y and biased model prediction $\hat{y}_{\text{bias}}$. Then, we select the model with the highest bias-unsupervised validation score:

$$\widehat{\text{ValScore}} := \frac{1}{K^2} \sum_{y,\hat{y}_{\text{bias}}} \frac{1}{|S_{y,\hat{y}_{\text{bias}}}|} \sum_{(x_i,y_i)\in S_{y,\hat{y}_{\text{bias}}}} \mathbb{1}(y_i = \hat{y}_i). \tag{1}$$

Overall, both worst-class accuracy and the bias-unsupervised validation score perform well for our use case and enable tuning of the mixture ratio without any group annotations in the validation set.

Contextualization with previous work. The concept of increasing exposure to majority class data without over-sampling the minority class was previously explored by Kirichenko et al., 2023, who propose drawing ten independent class-balanced subsets, training a logistic regression model on each, and averaging the weights. However, this method only works for *linear* models — as nonlinear models cannot be naively averaged — and it requires training an additional model for each subset, which is computationally infeasible for neural networks. In comparison, our mixture method is a simple and efficient alternative which extends easily to nonlinear models.

The catastrophic collapse of class-balanced upsampling and upweighting can be mitigated by a *mixture method* combining subsetting and upsampling. It increases exposure to majority class data without over-sampling the minority class and can improve baseline WGA beyond either technique.



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 3: **Scaling class-balanced pretrained models can improve worst-group accuracy.** We finetune each model size starting from pretrained checkpoints and plot the test worst-group accuracy (WGA) as well as the interpolation threshold, above which the model reaches 100% training accuracy. We find that model scaling is generally beneficial for WGA *only in conjuction* with appropriate class-balancing, and in fact scaling on imbalanced datasets or with the wrong balancing technique can harm robustness. Note that MultiNLI is class-balanced *a priori* and does not interpolate at any size. See Appendix C for training accuracy plots.

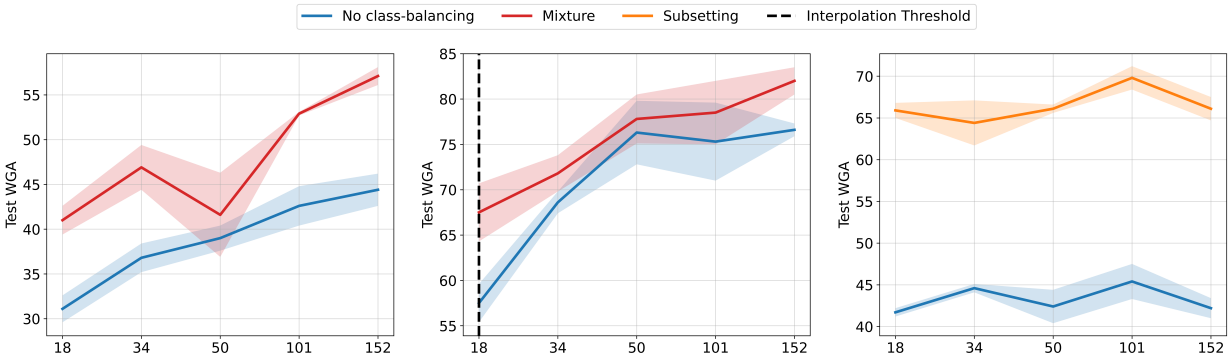
4 Model scaling improves WGA of class-balanced finetuning

The relationship between overparameterization and group robustness has been well-studied in the literature, with often conflicting conclusions depending on the setting (Sagawa et al., 2020b; Tripuraneni et al., 2021). In this section, we study the impact of model scaling on worst-group accuracy in a new setting — finetuning pretrained models — which more closely resembles practical use-cases. Importantly, we evaluate the impact of model scaling in conjunction with class-balancing to obtain a more reflective picture of the impact of group inequities alone on WGA as a function of model size. We find that with appropriate class-balancing, overparameterization can in fact significantly improve WGA over a very wide range of parameter scales, including before and after the interpolation threshold. On the other hand, scaling on imbalanced datasets or with the wrong balancing technique can harm robustness.

We take advantage of advancements in efficient architectures (Turc et al., 2019; Woo et al., 2023) to finetune pretrained models at vastly different parameter scales, ranging from 3.4M to 101M parameters. We study six different sizes of ImageNet1K-pretrained ConvNeXt-V2 and five different sizes of Book Corpus/English Wikipedia pretrained BERT; specifications for each model size are included in Appendix B. We utilize each class-balancing method proposed in Section 3 and display our results in Figure 3.

We find that model scaling is beneficial for group robustness in conjunction with appropriate class-balancing, with improvements of up to 12% WGA for interpolating models and 40% WGA for non-interpolating models. This comes in stark contrast to scaling on class-imbalanced datasets or with the wrong class-balancing technique, which shows either a neutral trend or decrease in WGA — the most severe examples being upsampling and upweighting on CivilComments. Returning to the class-balanced case, we notice some subtle nuances in performance across datasets. In particular, CivilComments WGA decreases slightly after the interpolation threshold, while Waterbirds and CelebA WGA continue to increase well beyond interpolation. On the other hand, BERT does not interpolate MultiNLI at any model size, greatly improving robustness at scale. It is unclear why Waterbirds and CelebA experience different behavior from CivilComments after the interpolation threshold; the toy linear model of Sagawa et al., 2020b suggests the presence of a benign "spurious-core information ratio", but a complete understanding is left to future theoretical investigation.

The most closely related work to ours is Pham et al., 2021, who also study the impact of scaling pretrained models on group robustness. Specifically, they use different ResNet (He et al., 2016) sizes ranging from 11.2M to 58.1M parameters, and they either train a linear probe on Waterbirds or finetune the entire model on CelebA. However, because their experiments do not employ any form of class-balancing, their conclusions about WGA scaling may be overly pessimistic. We replicate their experiments with our hyperparameters



(a) Waterbirds (last layer only) (b) Waterbirds (finetuning) (c) CelebA (finetuning)

Figure 4: **Class-balancing greatly affects ResNet scaling results of Pham et al., 2021.** We contrast the ResNet scaling behavior of Pham et al., 2021 — who do not use class-balancing — to the scaling of class-balanced ResNets using the best method from Section 3. We finetune each model size starting from pretrained checkpoints and plot the test worst-group accuracy (WGA) as well as the interpolation threshold, where the model reaches 100% training accuracy. On Waterbirds, we find that class-balancing enables a much more beneficial trend during model scaling. On CelebA, class-balancing greatly increases baseline WGA but does not affect scaling behavior (in contrast to the ConvNeXt-V2 plots in Figure 3). We use SGD for last-layer training and AdamW for full finetuning. See Appendix C for training accuracy plots.

and contrast with our results using class-balancing in Figure 4 to illustrate the performance disparity. We find that class-balancing greatly affects their results: on Waterbirds, class-balancing enables a much more beneficial trend during model scaling regardless of whether a linear probe or the entire model is trained. On the other hand, class-balancing greatly increases baseline WGA on CelebA but does not affect scaling behavior, whereas we observe a more positive WGA trend when scaling ConvNeXt-V2 in Figure 3.

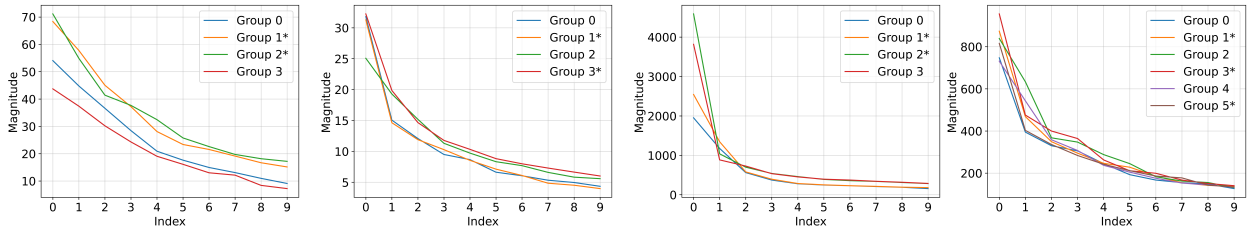
Contextualization with previous work. While previous work has primarily studied either linear probing of pretrained weights or training small models from scratch (Sagawa et al., 2020b; Tripuraneni et al., 2021), we study full finetuning of large-scale pretrained models and show that class-balancing can have a major impact on scaling behavior. We compare directly with the most closely related work, Pham et al., 2021, and show that class-balancing can either induce strikingly different scaling behavior or greatly increase baseline WGA. Overall, training with class-balancing allows us to isolate the impact of group inequities on robustness and more precisely observe the often-beneficial trend of model scaling for worst-group accuracy.

While overparameterization can sometimes harm WGA, pretraining and appropriate class-balancing make scaling generally beneficial. Moreover, modern language datasets are complex enough that standard models *do not interpolate*, greatly improving robustness at scale.

5 Spectral imbalance may exacerbate group disparities

In a recent paper, Kaushik et al., 2024 propose *spectral imbalance* of class covariance matrices, or differences in their eigenspectrum, as a potential source of disparities in accuracy *across classes* even when they are balanced *a priori*. Here, we examine whether similar insights hold in the group robustness setting. Our observations reveal surprising nuances in the behavior of *group-wise spectral imbalance*; nevertheless, we conclude that spectral imbalance may play a similar role in modulating worst-group accuracy after class-balancing is applied.

Let us denote by $\boldsymbol{z}_i$ the feature vector corresponding to a sample $\boldsymbol{x}_i$ (*i.e.,* the vectorized output of the



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 5: **Group disparities are visible in the top eigenvalues of the group covariance matrices.** We visualize the mean, across 3 experimental trials, of the top 10 eigenvalues of the group covariance matrices for a ConvNeXt-V2 Nano finetuned on Waterbirds and CelebA and a BERT Small finetuned on CivilComments and MultiNLI. The standard deviations are omitted for clarity. The models are finetuned using the best class-balancing method from Section 3 for each dataset. The group numbers are detailed in Table 2 and the minority groups within each class are denoted with an asterisk. The largest λ_1 in each case belongs to a minority group, though it may not be the *worst* group, and minority group eigenvalues are overall larger than majority group eigenvalues within the same class.

penultimate layer). Recall from Section 2 that Ω_g is the set of indices of samples which belong to group g. We further define $\bar{\boldsymbol{z}}_g$ to be the empirical mean of features with group g. To obtain the estimated eigenspectrum, we first compute the empirical covariance matrix for group $g \in \mathcal{G}$ by

$$\boldsymbol{\Sigma}_g = \frac{1}{|\Omega_g|} \sum_{i \in \Omega_g} (\boldsymbol{z}_i - \bar{\boldsymbol{z}}_g)(\boldsymbol{z}_i - \bar{\boldsymbol{z}}_g)^\top. \tag{2}$$

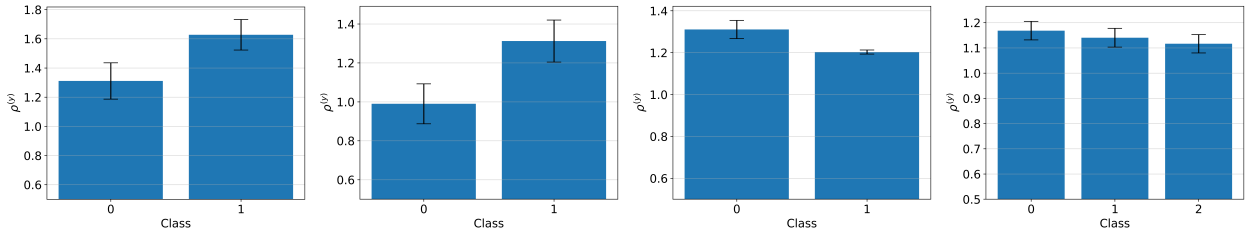
We then compute the eigenvalue decomposition $\boldsymbol{\Sigma}_g = \mathbf{V}_g \boldsymbol{\Lambda}_g \mathbf{V}_g^{-1}$, where $\boldsymbol{\Lambda}_g$ is a diagonal matrix with non-negative entries $\lambda_i^{(g)}$ and the columns of $\mathbf{V}_g$ are the eigenvectors of $\boldsymbol{\Sigma}_g$. Without loss of generality, we assume $\lambda_1^{(g)} \geq \lambda_2^{(g)} \geq \cdots \geq \lambda_m^{(g)}$ where m is the rank of $\boldsymbol{\Sigma}_g$.

We compute the group covariance matrices using a ConvNeXt-V2 Nano model for Waterbirds and CelebA, and a BERT Small model for CivilComments and MultiNLI. The models are finetuned using the best class-balancing method from Section 3 for each dataset. We plot the top 10 eigenvalues of each group covariance matrix in Figure 5. Even though we finetune with class-balancing, disparities in eigenvalues across groups are clearly visualized in Figure 5, especially for the largest eigenvalues. We include an extension to the top 50 eigenvalues, as well as results for class covariance matrices, in Appendix C.

Close observation of Figure 5 yields interesting findings. First, the group g^* that maximizes $\lambda_1^{(g)}$ in each case belongs to a minority group; though, importantly, it may not belong to the *worst* group. This is different from the findings of Kaushik et al., 2024, who showed in their setting that the largest eigenvalues typically belong to the *worst-performing* class, which usually depended on the model architecture in their setting as the training dataset was in fact class-balanced. Second, we find that minority group eigenvalues are overall larger than majority group eigenvalues, but only when *conditioned on the class*. In other words, a majority group belonging to one class may have larger eigenvalues than a minority group belonging to another class, but there exists a consistent spectral imbalance between majority and minority groups within the same class.[5]

To quantify this group-wise spectral imbalance, we introduce a new metric called the *intra-class spectral norm ratio*. Suppose $g_{\text{min}}(y)$ and $g_{\text{maj}}(y)$ are the minority and majority groups within a particular class $y \in \mathcal{Y}$. Then, we define the intra-class spectral norm ratio by $\rho^{(y)} := \lambda_1^{(g_{\text{min}}(y))} / \lambda_1^{(g_{\text{maj}}(y))}$. We note that while $\rho^{(y)}$ only considers the top eigenvalue and not the entire spectrum, the *absolute* magnitude of individual

[5] For example, in Figure 5c, the spectrum for group 3 (the majority group within class 1) is larger than the spectrum for group 1 (the minority group within class 0). However, conditioning on the class, we find that the spectrum for group 2 (the minority group within class 1) is larger than that of group 3, and the spectrum of group 1 is larger than that of group 0 (the majority group within class 0).



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 6: **Group-wise spectral imbalance is apparent once conditioned on the classes.** We plot the mean and standard deviation, across 3 experimental trials, of the intra-class spectral norm ratio $\rho^{(y)}$, or the ratio of the top eigenvalues of the minority and majority group covariance matrices, for each class $y \in \mathcal{Y}$. We compute this metric using a finetuned ConvNeXt-V2 Nano on Waterbirds and CelebA and a finetuned BERT Small on CivilComments and MultiNLI, each using the best class-balancing method from Section 3 for each dataset. The key observation is that $\rho^{(y)}$ is at least one for all classes $y \in \mathcal{Y}$ (except a single seed for class 0 on CelebA), illustrating a group disparity captured by the eigenspectrum once we condition on the classes.

eigenvalues was found in Kaushik et al., 2024 to correlate better with worst-class accuracy than *relative* metrics such as the "decay rate" of the eigenvalues. We plot the intra-class spectral norm ratios for each dataset in Figure 6; notably, they are always at least one (except for a single seed for class 0 on CelebA), showing more clearly the group disparity captured by the eigenspectrum.

Finally, in Table 5, we compare the class with the largest $\rho^{(y)}$ to the class with the largest disparity in group test accuracies, *i.e.,* $\text{Acc}(g_{\text{maj}}(y)) - \text{Acc}(g_{\text{min}}(y))$. We see that in most cases these classes correspond, suggesting an *explanatory power* of the intra-class spectral norm ratio. In particular, this correspondence is consistent throughout all trials of CelebA and CivilComments, the most class-imbalanced datasets we study.

Contextualization with previous work. Our spectral analysis of the group covariance matrices is inspired by Kaushik et al., 2024. We both study class-balanced settings, with the key difference that they study *class* disparities instead of group disparities. However, we show a more nuanced impact of spectral imbalance across both classes and groups, *i.e.,* spectral imbalance is more prevalent *between majority and minority groups within to the same class*, rather than across groups globally.

Spectral imbalance in the group covariance matrices may exacerbate group disparities even when the classes are balanced. While the worst-group covariance may not have largest spectral norm, the minority group spectra are consistently larger *conditioned on the class.*

6 Discussion

In this paper, we identified nuanced impacts of class-balancing and model scaling on worst-group accuracy, as well as a spectral imbalance in the group covariance matrices. Overall, our work calls for a more thorough investigation of generalization in the presence of spurious correlations to unify the sometimes contradictory perspectives in the literature. We hope that, as the community continues to develop group robustness methods with increasing performance and complexity, researchers and practitioners alike remain cognizant of the disproportionate impact of the details.

Acknowledgments. We thank Google Cloud for the gift of compute credits, Jacob Abernethy for the additional compute assistance, and Chiraag Kaushik for helpful discussions. T.L. acknowledges support from the DoD NDSEG Fellowship. V.M. acknowledges support from the NSF (awards CCF-223915 and IIS-2212182), Google Research, Adobe Research and Amazon Research.

References

Beery, Sara, Grant van Horn, and Pietro Perona (2018). "Recognition in Terra Incognita". In: *European Conference on Computer Vision (ECCV)* (cit. on pp. 1, 2).

Blodgett, Su Lin, Lisa Green, and Brendan O'Connor (2016). "Demographic Dialectal Variation in Social Media: A Case Study of African-American English". In: *Empirical Methods in Natural Language Processing (EMNLP)* (cit. on p. 2).

Borkan, Daniel, Lucas Dixon, Jeffrey Sorenson, Nithium Thain, and Lucy Vasserman (2019). "Nuanced metrics for measuring unintended bias with real data for text classification". In: *World Wide Web (WWW)* (cit. on p. 4).

Buda, Mateusz, Atsuto Maki, and Maciej A Mazurowski (2018). "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106.1, pp. 249–259 (cit. on p. 2).

Buolamwini, Joy and Timnit Gebru (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)* (cit. on p. 2).

Chatterji, Niladri S, Saminul Haque, and Tatsunori Hashimoto (2023). "Undersampling is a Minimax Optimal Robustness Intervention in Nonparametric Classification". In: *Transactions on Machine Learning Research (TMLR)* (cit. on p. 2).

Chaudhuri, Kamalika, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz (2023). "Why does throwing away data improve worst-group error?" In: *International Conference on Machine Learning (ICML)* (cit. on p. 2).

Chouldechova, Alexandra (2016). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)* (cit. on pp. 1, 2).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *North American Association for Computational Linguistics (NAACL)* (cit. on pp. 5, 19).

Falcon, William and the PyTorch Lightning maintainers and contributors (2019). "PyTorch Lightning". In: *GitHub* (cit. on p. 19).

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann (2020). "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2 (1), pp. 665–673 (cit. on p. 2).

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel (2019). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *International Conference on Learning Representations (ICLR)* (cit. on p. 2).

Ghorbani, Behrooz, Song Mei, Theodor Misiakiewicz, and Andrea Montanari (2019). "Limitations of lazy training of two-layers neural network". In: *Advances in Neural Information Processing Systems* 32 (cit. on p. 3).

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (2018). "Annotation Artifacts in Natural Language Inference Data". In: *North American Association for Computational Linguistics (NAACL)* (cit. on p. 2).

Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing (2017). "Learning from class-imbalanced data: Review of methods and applications". In: *Expert Systems with Applications* 73.1, pp. 220–239 (cit. on p. 2).

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant (2020). "Array programming with NumPy". In: *Nature* 585.1, pp. 357–362 (cit. on p. 19).

Hashimoto, Tatsunori B., Megha Srivastava, Hongseok Namkoong, and Percy Liang (2018). "Fairness Without Demographics in Repeated Loss Minimization". In: *International Conference on Machine Learning (ICML)* (cit. on p. 2).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 5, 9, 19).

Hendrycks, Dan, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer (2021). "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization". In: *International Conference on Computer Vision (ICCV)* (cit. on p. 3).

Hovy, Dirk and Anders Søgaard (2015). "Tagging Performance Correlates with Author Age". In: *Association for Computational Linguistics (ACL)* (cit. on p. 2).

Hunter, John D. (2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95 (cit. on p. 19).

Idrissi, Badr Youbi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz (2022). "Simple data balancing achieves competitive worst-group-accuracy". In: *Conference on Causal Learning and Reasoning (CLeaR)* (cit. on pp. 2–6, 19).

Izmailov, Pavel, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson (2022). "On Feature Learning in the Presence of Spurious Correlations". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on pp. 4, 8, 19).

Japkowicz, Nathalie and Shaju Stephen (2002). "The class imbalance problem: A systematic study". In: *Intelligent Data Analysis* 6.5, pp. 429–449 (cit. on p. 2).

Kaushik, Chiraag, Ran Liu, Chi-Heng Lin, Amrit Khera, Matthew Y Jin, Wenrui Ma, Vidya Muthukumar, and Eva L Dyer (2024). "Balanced Data, Imbalanced Spectra: Unveiling Class Disparities with Spectral Imbalance". In: *International Conference on Machine Learning (ICML)* (cit. on pp. 10–12, 22, 23).

Kirichenko, Polina, Pavel Izmailov, and Andrew Gordon Wilson (2023). "Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations". In: *International Conference on Learning Representations (ICLR)* (cit. on pp. 1, 4, 8).

Koh, Pang Wei, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang (2021). "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: *International Conference on Machine Learning (ICML)* (cit. on pp. 3, 4).

LaBonte, Tyler (2023). *Milkshake: Quick and extendable experimentation with classification models*. http://github.com/tmlabonte/milkshake (cit. on p. 19).

LaBonte, Tyler, Vidya Muthukumar, and Abhishek Kumar (2023). "Towards Last-layer Retraining for Group Robustness with Fewer Annotations". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on pp. 1, 2, 4–6).

Liu, Evan Zheran, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn (2021). "Just Train Twice: Improving Group Robustness without Training Group Information". In: *International Conference on Machine Learning (ICML)* (cit. on pp. 1–6, 8).

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang (2015). "Deep Learning Face Attributes in the Wild". In: *International Conference on Computer Vision (ICCV)* (cit. on pp. 1, 2, 4).

Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations (ICLR)* (cit. on p. 5).

Maity, Subha, Saptarshi Roy, Songkai Xue, Mikhail Yurochkin, and Yuekai Sun (2022). "How does over-parametrization affect performance on minority groups?" In: *arXiv preprint arXiv:2206.03515* (cit. on p. 3).

McCoy, Tom, Ellie Pavlick, and Tal Linzen (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference". In: *Association for Computational Linguistics (ACL)* (cit. on p. 2).

Moayeri, Mazda, Wenxiao Wang, Sahil Singla, and Soheil Feizi (2023). "Spuriosity Rankings: Sorting Data to Measure and Mitigate Biases". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on p. 2).

Nakkiran, Preetum, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever (2021). "Deep double descent: Where bigger models and more data hurt". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12, p. 124003 (cit. on p. 3).

Nam, Junhyun, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin (2022). "Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation". In: *International Conference on Learning Representations (ICLR)* (cit. on p. 1).

Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro (2014). "In search of the real inductive bias: On the role of implicit regularization in deep learning". In: *arXiv preprint arXiv:1412.6614* (cit. on p. 3).

Niven, Timothy and Hung-Yu Kao (2019). "Probing Neural Network Comprehension of Natural Language Arguments". In: *Association for Computational Linguistics (ACL)* (cit. on p. 2).

Oakden-Rayner, Luke, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré (2019). "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* (cit. on p. 2).

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer (2017). "Automatic differentiation in PyTorch". In: *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Automatic Differentiation* (cit. on p. 19).

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on p. 19).

Pham, Alan, Eunice Chan, Vikranth Srivatsa, Dhruba Ghosh, Yaoqing Yang, Yaodong Yu, Ruiqi Zhong, Joseph E Gonzalez, and Jacob Steinhardt (2021). "The effect of model size on worst-group generalization". In: *arXiv preprint arXiv:2112.04094* (cit. on pp. 3, 9, 10, 22).

Qiu, Shikai, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson (2023). "Simple and Fast Group Robustness by Automatic Feature Reweighting". In: *International Conference on Machine Learning (ICML)* (cit. on pp. 1–3).

Rosenfeld, Amir, Richard Zemel, and John K. Tsotsos (2018). "The Elephant in the Room". In: *arXiv preprint 1808.03305* (cit. on p. 2).

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.1, pp. 211–252 (cit. on pp. 5, 19).

Sagawa, Shiori, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang (2020a). "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization". In: *International Conference on Learning Representations (ICLR)* (cit. on pp. 1–5, 8, 19).

Sagawa, Shiori, Aditi Raghunathan, Pang Wei Koh, and Percy Liang (2020b). "An investigation of why overparameterization exacerbates spurious correlations". In: *International Conference on Machine Learning (ICML)* (cit. on pp. 2, 3, 9, 10).

Shetty, Rakshith, Bernt Schiele, and Mario Fritz (2019). "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)* (cit. on p. 2).

Shwartz-Ziv, Ravid, Micah Goldblum, Yucen Lily Li, C Bayan Bruss, and Andrew Gordon Wilson (2023). "Simplifying Neural Network Training Under Class Imbalance". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on p. 2).

Singla, Sahil and Soheil Feizi (2022). "Salient ImageNet: How to discover spurious features in Deep Learning?" In: *International Conference on Learning Representations (ICLR)* (cit. on p. 2).

Stromberg, Nathan, Rohan Ayyagari, Monica Welfert, Sanmi Koyejo, and Lalitha Sankar (2024). "Robustness to Subpopulation Shift with Domain Label Noise via Regularized Annotation of Domains". In: *arXiv preprint arXiv:2402.11039* (cit. on pp. 2–4, 6, 7).

Taghanaki, Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Trana, Ali Mahdavi-Amiri, and Ghassan Hamarneh (2022). "MaskTune: Mitigating Spurious Correlations by Forcing to Explore". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on p. 4).

Tatman, Rachael (2017). "Gender and Dialect Bias in YouTube's Automatic Captions". In: *Association for Computational Linguistics (ACL) Workshop on Ethics in Natural Language Processing* (cit. on p. 2).

TorchVision maintainers and contributors (2016). "TorchVision: PyTorch's Computer Vision library". In: *GitHub* (cit. on p. 19).

Tripuraneni, Nilesh, Ben Adlam, and Jeffrey Pennington (2021). "Overparameterization Improves Robustness to Covariate Shift in High Dimensions". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on pp. 3, 9, 10).

Tsirigotis, Christos, Joao Monteiro, Pau Rodriguez, David Vazquez, and Aaron Courville (2023). "Group Robust Classification Without Any Group Information". In: *Conference on Neural Information Processing Systems (NeurIPS)* (cit. on p. 8).

Turc, Iulia, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models". In: *arXiv preprint 1908.08962* (cit. on pp. 9, 19).

Vapnik, Vladimir (1998). *Statistical Learning Theory*. Wiley (cit. on p. 1).

Wah, Catherine, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. California Institute of Technology (cit. on p. 4).

Welinder, Peter, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona (2010). *Caltech-UCSD Birds 200*. Tech. rep. California Institute of Technology (cit. on p. 4).

Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). "A broad-coverage challenge corpus for sentence understanding through inference". In: *North American Association for Computational Linguistics (NAACL)* (cit. on pp. 4, 19).

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP) System Demonstrations* (cit. on p. 19).

Woo, Sanghyun, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie (2023). "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 5, 9, 19).

Xiao, Kai Yuanqing, Logan Engstrom, Andrew Ilyas, and Aleksander Mądry (2021). "Noise or Signal: The Role of Image Backgrounds in Object Recognition". In: *International Conference on Learning Representations (ICLR)* (cit. on pp. 1, 2).

Yang, Yuzhe, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi (2023). "Change is Hard: A Closer Look at Subpopulation Shift". In: *International Conference on Machine Learning* (cit. on p. 8).

Zech, John R., Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann (2018). "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". In: *PLoS Medicine* 15 (11), e1002683 (cit. on pp. 1, 2).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2021). "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3, pp. 107–115 (cit. on p. 3).

Zhang, Michael, Nimit S. Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré (2022). "Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations". In: *International Conference on Machine Learning (ICML)* (cit. on pp. 1, 2, 4).

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *International Conference on Computer Vision (ICCV)* (cit. on pp. 5, 19).

A Dataset composition

Table 2: **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 and MultiNLI is the only dataset which is class-balanced *a priori*. The minority groups within each class are denoted by an asterisk in the "Num" column. Probabilities may not sum to 1 due to rounding.

Dataset	Group g			Training distribution $\hat{p}$			Data quantity		
	Num	Class y	Spurious s	$\hat{p}(y)$	$\hat{p}(g)$	$\hat{p}(y\|s)$	Train	Val	Test
Waterbirds	0	landbird	land	.768	.730	.984	3498	467	2225
	1*	landbird	water		.038	.148	184	466	2225
	2*	waterbird	land	.232	.012	.016	56	133	642
	3	waterbird	water		.220	.852	1057	133	642
CelebA	0	non-blond	female	.851	.440	.758	71629	8535	9767
	1*	non-blond	male		.411	.980	66874	8276	7535
	2	blond	female	.149	.141	.242	22880	2874	2480
	3*	blond	male		.009	.020	1387	182	180
CivilComments	0	neutral	no identity	.887	.551	.921	148186	25159	74780
	1*	neutral	identity		.336	.836	90337	14966	43778
	2*	toxic	no identity	.113	.047	.079	12731	2111	6455
	3	toxic	identity		.066	.164	17784	2944	8769
MultiNLI	0	contradiction	no negation	.333	.279	.300	57498	22814	34597
	1*	contradiction	negation		.054	.761	11158	4634	6655
	2	entailment	no negation	.334	.327	.352	67376	26949	40496
	3*	entailment	negation		.007	.104	1521	613	886
	4	neither	no negation	.333	.323	.348	66630	26655	39930
	5*	neither	negation		.010	.136	1992	797	1148

B Training details

We utilize a ResNet (He et al., 2016) and ConvNeXt-V2 (Woo et al., 2023) model 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. These pretrained models are used as the initialization for ERM finetuning under the cross-entropy loss. We use standard ImageNet normalization with standard flip and crop data augmentation for the vision tasks and BERT tokenization for the language tasks (Izmailov et al., 2022). Our implementation uses the following packages: `NumPy` (Harris et al., 2020), `PyTorch` (Paszke et al., 2017; Paszke et al., 2019), `Lightning` (Falcon and the PyTorch Lightning maintainers and contributors, 2019), `TorchVision` (TorchVision maintainers and contributors, 2016), `Matplotlib` (Hunter, 2007), `Transformers` (Wolf et al., 2020), and `Milkshake` (LaBonte, 2023).

To our knowledge, the licenses of Waterbirds and CelebA are unknown. CivilComments is released under the CC0 license, and information about MultiNLI's license may be found in Williams et al., 2018.

Our experiments were conducted on four Google Cloud Platform (GCP) 16GB Nvidia Tesla P100 GPUs and two local 24GB Nvidia RTX A5000 GPUs. The spectral imbalance experiments in Section 5 were conducted on a GCP system with a 16-core CPU and 128GB of RAM. We believe our work could be reproduced for under $5000 in GCP compute credits, with a majority of that compute going towards running experiments over multiple random seeds.

We list model scaling parameters in Table 3 and hyperparameters used for each dataset in Table 4. The smaller BERT models were introduced by Turc et al., 2019. We perform model selection only for our mixture balancing method (see Table 1) and not for the ERM finetuning hyperparameters, most of which are standard in the literature (Sagawa et al., 2020a; Idrissi et al., 2022; Izmailov et al., 2022). For the last-layer training experiments in Figure 4a and Figure 9a, we use SGD with learning rate 10^{-3} and train for 20 epochs. Different from previous work, we train CelebA for about 3× more gradient steps than usual to ensure convergence, and we double the batch size for CivilComments and MultiNLI to increase training stability (we also double the epochs to hold the number of gradient steps constant).

Table 3: **Model scaling parameters.** ConvNeXt-V2 and ResNet are composed of four separate "stages", and we list the depths of these stages individually. All of these configurations are standard in the literature.

(a) **ConvNeXt-V2 parameters.**

Size	Width	Depth (4 stages)	Params
Atto	40	(2, 2, 6, 2)	3.4M
Femto	48	(2, 2, 6, 2)	4.8M
Pico	64	(2, 2, 6, 2)	8.6M
Nano	80	(2, 2, 8, 2)	15.0M
Tiny	96	(3, 3, 9, 3)	27.9M
Base	128	(3, 3, 27, 3)	87.7M

(b) **BERT parameters.**

Size	Width	Depth	Params
Tiny	2	128	4.4M
Mini	4	256	11.2M
Small	4	512	28.8M
Medium	8	512	41.4M
Base	12	768	109M

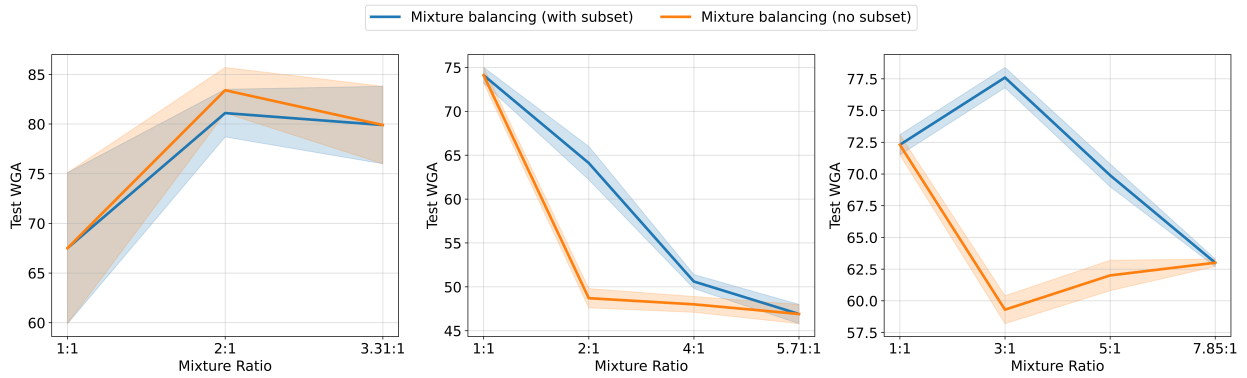
(c) **ResNet parameters.**

Size	Width (4 stages)	Depth (4 stages)	Params
18	(64, 128, 256, 512)	(2, 2, 2, 2)	11.2M
34	(64, 128, 256, 512)	(3, 4, 6, 3)	21.3M
50	(256, 512, 1024, 2048)	(3, 4, 6, 3)	23.5M
101	(256, 512, 1024, 2048)	(3, 4, 23, 3)	42.5M
152	(256, 512, 1024, 2048)	(3, 8, 36, 3)	58.1M

Table 4: **ERM finetuning hyperparameters.**

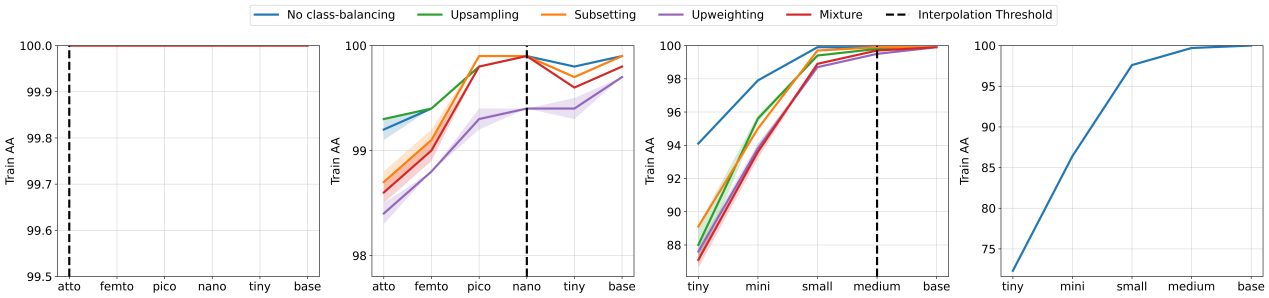
Dataset	Optimizer	Initial LR	LR schedule	Batch size	Weight decay	Epochs
Waterbirds	AdamW	1×10^{-5}	Cosine	32	1×10^{-4}	100
CelebA	AdamW	1×10^{-5}	Cosine	32	1×10^{-4}	20
CivilComments	AdamW	1×10^{-5}	Linear	32	1×10^{-4}	20
MultiNLI	AdamW	1×10^{-5}	Linear	32	1×10^{-4}	20

C Supplemental experiments



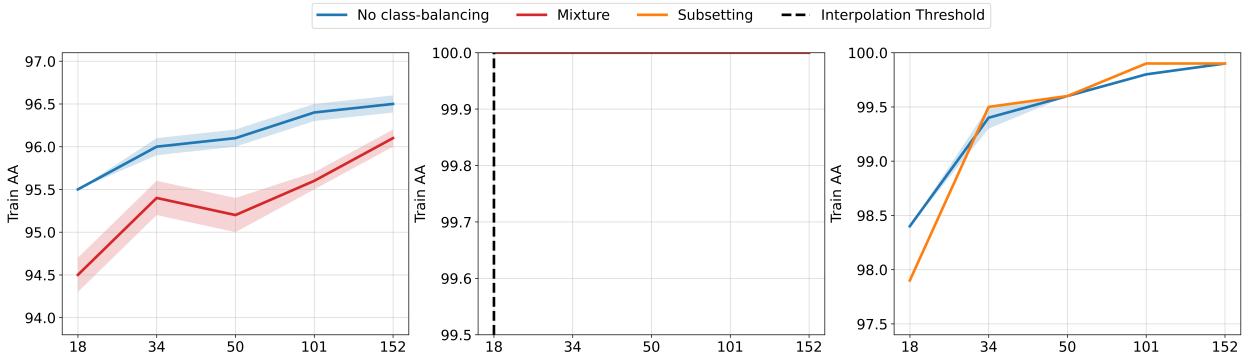
(a) Waterbirds (b) CelebA (c) CivilComments

Figure 7: **Mixture balancing ablation studies.** We perform two ablation studies on our mixture balancing method. First, we vary the class-imbalance ratio across the x axis. On the left-hand side, using a class-imbalance ratio of 1:1 reduces to the subsetting technique; on the right-hand side, using the original class-imbalance ratio in the dataset reduces to upsampling. Second, we perform an ablation of whether subsetting is essential in mixture balancing. We plot our proposed method (which takes a subset of data based on the class-imbalance ratio, then performs upsampling) against a similar method which does not use subsetting, instead adjusting the class probabilities on the entire dataset as specified by the class-imbalance ratio. MultiNLI is class-balanced *a priori*, so we do not include it here.



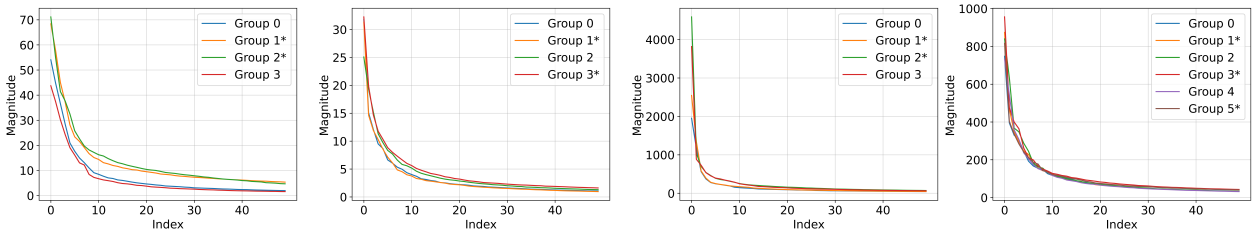
(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 8: **Average accuracy of scaled models.** We finetune each model size starting from pretrained checkpoints and plot the train average accuracy (AA) as well as the interpolation threshold, where *at least one seed* of the non-class-balanced model reaches 100% training accuracy. (For example, CelebA does not interpolate with all three seeds). Average accuracy consistently increases with model size regardless of class-balancing, implying the scaling dynamics for AA and WGA are starkly different. Note that MultiNLI is class-balanced *a priori* and does not interpolate at any size.



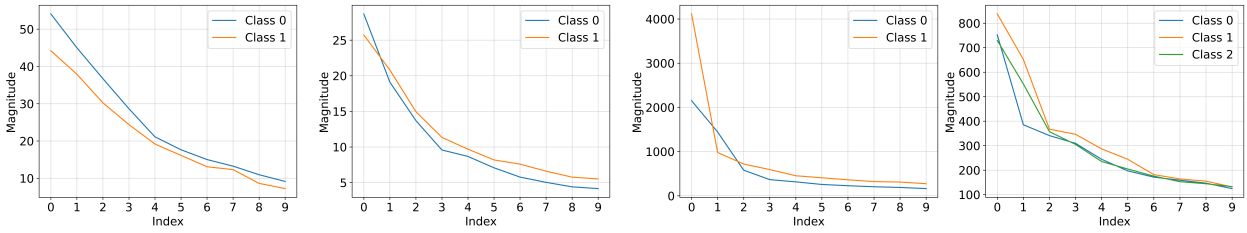
(a) Waterbirds (last layer only) (b) Waterbirds (finetuning) (c) CelebA (finetuning)

Figure 9: **Average accuracy of scaled ResNets.** We contrast the ResNet scaling behavior of Pham et al., 2021 — who do not use class-balancing — to the scaling of class-balanced ResNets using the best method from Section 3. We finetune each model size starting from pretrained checkpoints and plot the train average accuracy (AA) as well as the interpolation threshold, where the model reaches 100% training accuracy. Similarly to Figure 8, average accuracy consistently increases with model size. We use SGD for last-layer training and AdamW for full finetuning.



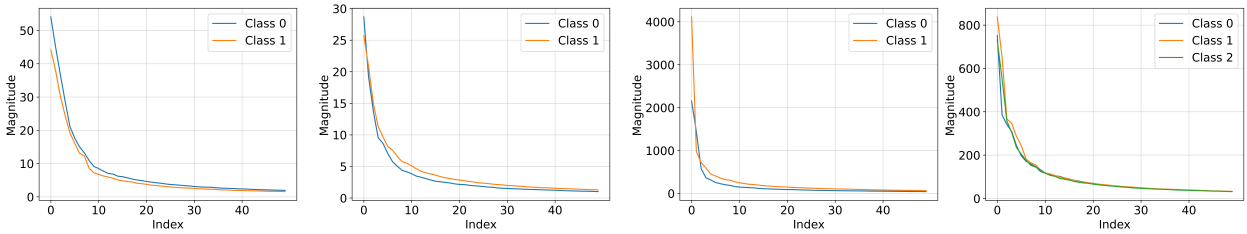
(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 10: **Additional eigenvalues of the group covariance matrices.** In contrast to Figure 5, we visualize the top 50 eigenvalues of the group covariance matrices for a ConvNeXt-V2 Nano finetuned on Waterbirds and CelebA and a BERT Small finetuned on CivilComments and MultiNLI. The models are finetuned using the best class-balancing method from Section 3 for each dataset. The group numbers are detailed in Table 2 and minority groups are marked with an asterisk. It becomes difficult to distinguish patterns between the groups in the lower eigenvalues, which is why we focus only on local properties of the top eigenvalues (*e.g.,* the spectral norm and the relative ordering of the groups). With that said, it would be interesting to explore power-law decay metrics (Kaushik et al., 2024), which characterize relatively global properties of the eigenspectrum, in future work.



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 11: **Class disparities are visible in the top eigenvalues of the class covariance matrices.** We visualize the mean, across 3 experimental trials, of the top 10 eigenvalues of the class covariance matrices for a ConvNeXt-V2 Nano finetuned on Waterbirds and CelebA and a BERT Small finetuned on CivilComments and MultiNLI. The standard deviations are omitted for clarity. The models are finetuned using the best class-balancing method from Section 3 for each dataset. The class numbers are detailed in Table 2. The minority class eigenvalues for CelebA and CivilComments are overall larger, while the reverse is true for Waterbirds, a slightly different conclusion than Kaushik et al., 2024.



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 12: **Additional eigenvalues of the class covariance matrices.** In contrast to Figure 11, we visualize the top 50 eigenvalues of the class covariance matrices for a ConvNeXt-V2 Nano finetuned on Waterbirds and CelebA and a BERT Small finetuned on CivilComments and MultiNLI. The models are finetuned using the best class-balancing method from Section 3 for each dataset. The class numbers are detailed in Table 2. Similar to the groups, it becomes difficult to distinguish patterns between the classes in the lower eigenvalues, which is why we again focus only on local properties of the top eigenvalues (*e.g.,* the spectral norm and the relative ordering of the classes).

Table 5: **Correspondence between $\rho^{(y)}$ and intra-class group accuracy disparity.** We compare $\rho^{(y)}$, the intra-class spectral norm ratio, to the difference in intra-class group accuracy. Each row represents a different experimental seed. Each cell contains a tuple with the class label for the class with largest value of $\rho^{(y)}$ paired with the class label for the class with the largest intra-class group test accuracy disparity, *i.e.,* $\text{Acc}(g_{\text{maj}}(y)) - \text{Acc}(g_{\text{min}}(y))$. We see that in most cases these classes correspond, suggesting an *explanatory power* of the spectral norm ratio. In particular, this correspondence is consistent throughout all trials of CelebA and CivilComments, the most class-imbalanced datasets we study.

Seed	Waterbirds	CelebA	CivilComments	MultiNLI
1	(1, 1)	(1, 1)	(0, 0)	(0, 0)
2	(1, 1)	(1, 1)	(0, 0)	(0, 1)
3	(0, 1)	(1, 1)	(0, 0)	(2, 0)