LACER: LOSS-AWARE CLUSTERING FOR EFFECTIVE REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks trained with Empirical Risk Minimization (ERM) are prone to rely on simple spurious features—features that are correlated with the target but are not causally related to it. To mitigate this over-reliance, Deep Feature Reweighting (DFR) has emerged as an efficient approach, which works by retraining the last layer of an ERM model on a small *reweighting* dataset. While effective, DFR requires group annotations to create the reweighting dataset, which may be challenging and costly to obtain. Though subsequent works have proposed ways to alleviate this constraint, existing methods still largely rely on group annotations for hyperparameter tuning to achieve robust performance. In this paper, we present **LACER**, a method that improves group robustness without requiring explicit group annotations for either training or model selection. **LACER** operates in two stages: first estimating group labels through a loss-weighted clustering formulation that effectively identifies clusters corresponding to underrepresented groups in the validation set, then leveraging these estimated labels for last-layer retraining. Our results provide the empirical evidence that combining semantic feature information with loss values enables effective group label estimation. We validate **LACER** across multiple vision spurious correlations benchmarks, demonstrating performance comparable to oracle last-layer retraining methods that utilize ground-truth group annotations.

1 INTRODUCTION

Deep learning classifiers trained with empirical risk minimization (ERM) are known to rely heavily on *spurious features*—attributes that exhibit correlation with the target class in the training data but are not causally related to the true underlying predictive function. Consequently, these models perform poorly on groups where the spurious correlations do not hold, leading to low worst-group accuracy (WGA) (Beery et al., 2018; Geirhos et al., 2020). This behavior has been attributed to ERM's procedure of minimizing the average training loss. Due to the underrepresentation of minority groups in training data, coupled with the simplicity bias of SGD-based optimization algorithms (Shah et al., 2020), models tend to exploit simple spurious correlations that dominate the majority of training examples rather than learning robust features for classification.

Distributionally Robust Optimization (DRO) techniques enable learning group-robust models by minimizing the worst-case loss over a set of pre-defined groups (Sagawa et al., 2020), rather than the average loss as in ERM. However, the effectiveness of DRO methods is fundamentally limited by their reliance on explicit group annotations during training. Deep Feature Reweighting (DFR) (Kirichenko et al., 2023) has emerged as an effective alternative that improves group robustness by retraining only the last layer of an ERM trained model on a group-balanced reweighting dataset, requiring group annotations for just a small held-out subset. Subsequent work (Qiu et al., 2023; LaBonte et al., 2023) have further relaxed this constraint by eliminating the need for group annotations during last-layer retraining, but still rely on a group-annotated validation set for hyperparameter tuning. Although this reduces the annotation burden, for some domains collecting group labels is significantly costly or challenging.

In this work, we propose **LACER** (**L**oss-**A**ware **C**lustering for **E**ffective **R**eweighting), a practical approach based on last-layer retraining (LLR) that improves group robustness while requiring only knowledge of the total number of groups present in the dataset[1]—a significant relaxation compared to prior approaches. **LACER** operates in two stages: first, we employ a novel loss-weighted clustering technique to partition the feature space of a held-out set, effectively identifying clusters that correspond to underlying groups. Next, we utilize these cluster assignments as proxy group labels to construct a group-balanced reweighting dataset for last-layer retraining.

One of the key insights in our work is that we can reduce the amount of required prior knowledge about the data by leveraging empirical observations from prior group robustness works. Specifically, our method builds upon an observation noted in several prior works (Sohoni et al., 2022; Kirichenko et al., 2023; Izmailov et al., 2022; Zhang et al., 2022; Yang et al., 2024, among others): representations of examples within the same group tend to cluster more closely together, and groups with performance gaps are typically separable in the model's feature embedding space. Moreover, minority groups under ERM exhibit systematically higher loss values (e.g., (Liu et al., 2021) and Qiu et al. (2023) directly leverage this idea to de-bias models). By incorporating both ideas and leveraging them in our clustering approach, **LACER** effectively discovers underlying group structure without requiring explicit group annotations.

We empirically validate **LACER** across three diverse image classification tasks, demonstrating improvements in worst-group accuracy compared to existing baselines. Through extensive ablations on varying degrees of group imbalance in validation data, we show that **LACER** has particularly strong advantage in scenarios with high group imbalance—a critical advantage given that validation sets obtained by setting aside a subset of training data would likely exhibit such imbalances.

2 Background

In this section, we formalize the problem of group robustness (§2.1) and review existing approaches based on last-layer retraining, with a particular focus on Deep Feature Reweighting (§2.2).

2.1 Group Robustness

We consider the group robustness setting (Sagawa et al., 2020), where each input datapoint $x \in \mathcal{X}$ is associated with a class label $y \in \mathcal{Y}$ and a spurious attribute $s \in \mathcal{S}$. The groups $g \in \mathcal{G}$ are defined by combinations of class label and spurious attribute ($\mathcal{G} = \mathcal{Y} \times \mathcal{S}$). We consider scenarios with inherent group imbalance in the training data distribution ($\mathcal{D}_{\text{train}}$), where certain groups are highly represented (*majority groups*) while others are significantly underrepresented (*minority groups*). Our focus is to build classification models that maintain high accuracy across all groups, which we evaluate using *worst-group accuracy*—the minimum accuracy across all groups $\mathcal{G}$.

2.2 Last-Layer Retraining

Deep Feature Reweighting (DFR) (Kirichenko et al., 2023) demonstrated that group robustness can be improved by retraining just the last layer of an ERM-trained model on a group-balanced reweighting dataset. This approach was motivated by a key observation: while ERM models may rely heavily on spurious features for classification, they still learn meaningful representations of the core predictive features. Formally, given a model $m_\theta = (f_\phi, f_\psi)$ trained using standard ERM, where f_ϕ is the feature extractor and f_ψ represents the last classifier layer, DFR freezes f_ϕ and retrains f_ψ on a group-balanced reweighting dataset. While effective and efficient, DFR requires access to group annotations to construct the reweighting dataset.

Subsequent works have explored various approaches to relax DFR's requirement of group annotations. One line of work proposes using class-balanced reweighting datasets for last-layer retraining, though this approach shows reduced effectiveness when the held-out data exhibits high group imbalance (LaBonte et al., 2023). Automatic Feature Reweighting (AFR) method introduces a weighted loss for last-layer retraining which is designed to emphasize examples where the ERM model performs poorly, thereby implicitly upweighting minority groups (Qiu et al., 2023). While this method eliminates the need for group annotations during training, it still relies a group-annotated validation set for hyperparameter tuning to achieve robust performance—a requirement that can be challenging to satisfy in certain domains.

[1]For example, 2 groups per class in standard Waterbirds dataset.

Algorithm 1 LACER: Loss Aware Clustering for Effective Reweighting

1: **Input:** Training set $\mathcal{D}_{\text{train}}$, held out set $\mathcal{D}_{\text{val}}$, a classifier decomposed as $m_\theta = f_\phi \circ f_\psi$, the number of groups for each class c denoted as g_c.
2: **Output:** Model $m_{\hat{\theta}} = f_\phi \circ f_{\hat{\psi}}$ with improved group robustness compared to ERM baseline.
3: **Stage 0: ERM** Model checkpoint $\theta = (\phi, \psi)$ trained using ERM until convergence on $\mathcal{D}_{\text{train}}$.
4: **Stage 1: Estimate group labels using loss-weighted clustering**
5: **for** each class c, g_c **do**
6: $\quad E_c \leftarrow \{f_\psi(x) \mid (x, y) \in \mathcal{D}_{\text{val}}, y = c\}$ {Extract feature embeddings for class c.}
7: $\quad$ Weight each data point x_i with $w_i = \exp(-\gamma_{y_i}\, p_i)$ where p_i is the softmax probability for the correct class y_i and γ_{y_i} is the class dependent upweighting parameter.
8: $\quad$ Cluster E_c into g_c clusters with *weighted k-means clustering* with weights w_i.
9: $\quad$ Use the cluster labels z_i as the estimated group labels for LLR.
10: **end for**
11: **Stage 2: Last layer retraining**
Perform LLR with ℓ_1 regularization and use the estimated group labels z_i to construct a group-balanced reweighting dataset.

3 LACER: LOSS AWARE CLUSTERING FOR EFFECTIVE REWEIGHTING

We introduce **LACER**, a method for improving group robustness in scenarios where explicit group annotations are unavailable. Our approach is built on two key empirical observations: (a) groups that exhibit significant performance gaps are separable in the neural network's feature space, and (b) minority group samples tend to have higher loss values under the final ERM model. **LACER** leverages these insights through a two-stage framework: (1) first estimating group labels using a novel loss-weighted clustering approach (3.1), and (2) then using these estimated labels to retrain the last layer f_ψ of an ERM model (3.2). The pseudocode for our algorithm is presented in Algorithm 1.

3.1 LOSS-WEIGHTED CLUSTERING

The first stage of our approach involves estimating the group labels for the held-out validation dataset $\mathcal{D}_{\text{val}}$. Our approach builds on the observation that groups within a class exhibiting significant performance discrepancy must have distinguishable feature representations (Sohoni et al., 2022)—otherwise, the classifier would achieve similar accuracy across groups. While this suggests that groups should be separable in the feature space, standard k-means clustering—which minimizes average reconstruction error—often fails to identify clusters corresponding to minority groups due to their underrepresentation in the validation set ($\mathcal{D}_{\text{val}}$).

To address this limitation, we propose a loss-weighted clustering approach that leverages loss values from the ERM checkpoint to inform cluster assignments. Specifically, our formulation upweights samples with higher loss values, enabling better identification of clusters that correspond to minority groups—which typically exhibit higher losses under the ERM model.

Specifically, for each class c, we first extract the feature embedding E_c using the frozen feature extractor f_ϕ for all validation samples belonging to that class. We then assign weights to each datapoint x_i based on its loss under the ERM model:

$$w_i = \exp(-\gamma_{y_i}\, p_i). \tag{1}$$

Here p_i is the softmax probability assigned to the correct class y_i by the ERM checkpoint, and γ_{y_i} is a class-dependent parameter that determines the degree to which points with higher loss are upweighted. We adopt this weight formulation in equation 1 from AFR (Qiu et al., 2023) which uses it to reweight examples directly for last-layer retraining, while we use these weights to modify k-means clustering. While AFR sometimes relies on validation group labels to tune the hyperparameter γ, we utilize the silhouette score (Rousseeuw, 1987) to automatically tune the hyper-parameter in an unsupervised way. This automated tuning is motivated by the observation that the optimal value of γ_{y_i} should result in well-separated clusters that correspond to minority and majority groups within each class.

Table 1: **Comparison of last-layer retraining methods on Waterbirds & Urbancars.** We report the average *worst-group accuracy* over 5 independent runs, with columns showing different imbalance ratios (minority size to majority size) in the validation set $\mathcal{D}_{\text{val}}$. DFR represents the oracle performance with access to ground truth group annotations, while other methods operate without explicit group labels. AFR results are shown for three fixed values of γ, the upweighting parameter. Our method, **LACER**, achieves competitive performance across all settings while only requiring metadata knowledge of the number of groups present in data, demonstrating effective improvement in group robustness with minimal supervision.

Method	Waterbirds				Urbancars			
	0.1	0.15	0.2	1.0	0.1	0.15	0.2	1.0
ERM	73.3	73.3	73.3	73.3	25.9	25.9	25.9	25.9
DFR (Kirichenko et al., 2023)	80.1	85.2	88.2	92.3	81.1	80.9	84.6	84.8
CB LLR (LaBonte et al., 2023)	69.9	76.2	82.0	92.5	70.2	73.1	74.5	85.2
AFR ($\gamma = 1.0$) (Qiu et al., 2023)	76.8	80.1	83.7	92.3	52.0	57.7	64.6	86.4
AFR ($\gamma = 2.0$) (Qiu et al., 2023)	78.7	82.3	86.8	91.6	72.6	76.0	80.6	85.2
AFR ($\gamma = 3.0$) (Qiu et al., 2023)	80.3	83.4	88.2	85.6	78.5	78.7	81.4	69.2
k-Means Clustering + LLR	66.6	79.3	83.9	92.3	70.7	72.8	72.8	84.9
LACER (Ours)	83.7	87.4	89.1	92.6	73.9	77.7	78.2	84.3

In practice, prior to cluster the datapoints, we apply UMAP dimensionality reduction (McInnes et al., 2018). The dimensionality reduction hyperparameters are decided automatically using silhouette scores, with details provided in the Appendix B.

Cluster-averaged silhouette score. The standard silhouette score (Rousseeuw, 1987) measures how well each datapoint is clustered with similar samples, calculated as the average silhouette coefficient across all datapoints. In practice, we observe that this aggregate metric tends to overlook the clustering quality of smaller clusters, which typically correspond to minority groups underrepresented in the held-out dataset. To address this limitation, we propose a *cluster-averaged silhouette score* (CAS) that first computes the silhouette score for each cluster independently and then averages these scores across clusters. This modification ensures equal importance to all clusters regardless of their size, making it particularly suitable for scenarios with significant group imbalance. Formally, given clusters $\{C_1, \ldots, C_k\}$, we define the silhouette score SIL_{C_i} for cluster i as:

$$\text{SIL}_{C_i} = \frac{1}{|C_i|} \sum_{j \in C_i} \text{SIL}(x_j), \tag{2}$$

where $\text{SIL}(x_j)$ is the silhouette coefficient for datapoint x_j. The *cluster-averaged silhouette score* SIL_{CAS} is then computed as:

$$\text{SIL}_{CAS} = \frac{1}{k} \sum_{i=1}^{k} \text{SIL}_{C_i}, \tag{3}$$

where k is the number of clusters (i.e., the number of groups in the dataset).

3.2 Step 2: Last Layer Retraining

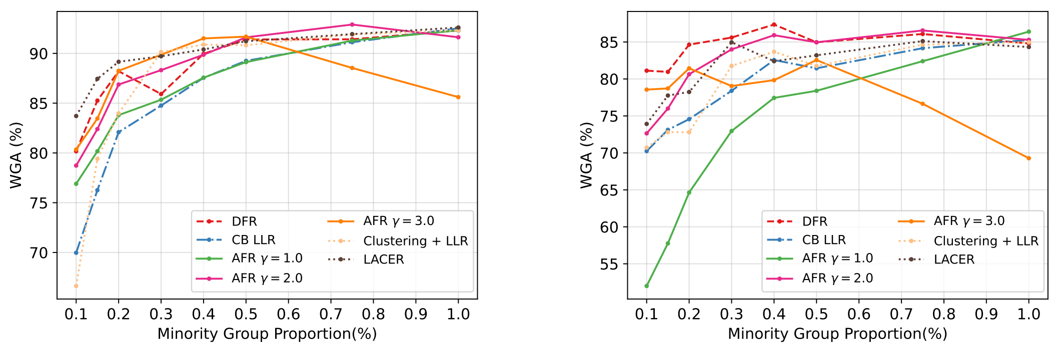
Last-layer retraining (LLR) has proven effective in improving group robustness of ERM-trained models. In the second stage of our algorithm, we leverage the estimated group labels from the first stage (§3.1) to construct a group-balanced reweighting dataset for LLR. For the retraining process, we follow the standard hyperparameter settings from DFR (Kirichenko et al., 2023), using ℓ_1-regularization (λ as the regularization strength), with details provided in the Appendix A.

4 Experiments

In this section, we evaluate the effectiveness of our proposed algorithm on standard benchmarks for group robustness.

Table 2: **Comparison of last-layer retraining methods on CelebA using ResNet and ConvNeXT architectures.** We report the average *worst-group accuracy* (WGA) across give different runs. *K-Means* refers to standard k-means clustering for group label estimation followed by LLR. While **LACER** improves the WGA of the original ERM model on ResNet, it does not match the performance of other baseline approaches. However, with the stronger ConvNeXT feature extractor, **LACER** achieves performance comparable to the oracle DFR.

Model	*ERM*	*DFR*	*CB LLR*	*AFR*			*K-Means*	**LACER**
				$\gamma = 1$	$\gamma = 2$	$\gamma = 3$		
ResNet	43.7	89.5	68.4	79.2	85.2	80.6	72.0	73.3
ConvNeXT	46.3	90.7	73.4	82.5	85.6	78.0	91.4	90.7



(a) Waterbirds (b) UrbanCars

Figure 1: **Comparison of worst-group accuracy (WGA) across varying minority group proportions in the validation set for (a) Waterbirds and (b) UrbanCars datasets.** DFR represents oracle performance with ground truth annotations, while other methods operate without explicit group labels. **LACER** achieves competitive performance with the best-performing AFR variants across all proportions while only requiring metadata knowledge of the number of groups, demonstrating its effectiveness in scenarios where explicit group annotations are not viable.

Datasets. We evaluate **LACER** on three image classification benchmarks: **(a)** *Waterbirds* (Sagawa et al., 2020), **(b)** *CelebA* (Liu et al., 2015) and **(b)** *UrbanCars* (Li et al., 2023). The *Waterbirds* dataset requires classifying birds as either waterbirds or landbirds, where the background (water or land) serves as the spurious feature. *CelebA* involves hair color prediction with gender as the spurious attribute. *UrbanCars* presents a more challenging scenario with multiple spurious features per class: the task is to classify cars as either urban or country vehicles, where both the background and co-occurring objects serve as spurious features.

Setup. Following Kirichenko et al. (2023), we use a ResNet-50 model (He et al., 2016) pretrained on ImageNet-1k (Russakovsky et al., 2015) as our base architecture. For *CelebA*, we additionally experiment with a stronger ConvNeXT model (Liu et al., 2022) pre-trained on ImageNet-22k and fine-tuned on ImageNet-1k. For all experiments, we train the feature extractor f_ϕ using the standard training subset and use $\mathcal{D}_{\text{val}}$ for last-layer retraining across all baselines. We report average worst-group accuracy over five random seeds.

Baselines. We compare **LACER** against several LLR approaches that do not require explicit group annotations: (1) ERM-trained base model, (2) class-balanced LLR (LaBonte et al., 2023), which retrains using a class-balanced reweighting dataset, (3) AFR (Qiu et al., 2023) with different fixed values of γ, and (4) LLR using group labels estimated through standard k-means clustering. We compare against DFR (Kirichenko et al., 2023) as an oracle baseline that uses ground truth annotations for LLR.

Results. In Table 1, we present results on Waterbirds (Sagawa et al., 2020) and UrbanCars (Li et al., 2023), varying the ratio of minority-to-majority examples in validation data $\mathcal{D}_{\text{val}}$. In realistic scenarios where validation data is a random subset of all available data, it is likely that this ratio will be highly skewed. On Waterbirds, **`LACER`** shows the strongest performance out of all methods, outperforming even DFR on highly skewed validation sets. On the more challenging UrbanCars dataset, **`LACER`** demonstrates competitive performance, however, it is outperformed by AFR with $\gamma = 3$. Figure 1 provides a more comprehensive comparison across a wider range of minority-to-majority ratios.

Table 2 presents results comparing **`LACER`** against baselines on CelebA Liu et al. (2015) using both ResNet and ConvNeXT architectures. Unlike Waterbirds and UrbanCars, the validation set in CelebA has natural group imbalance, so we do not vary the group proportions. With ResNet, LACER improves upon the ERM baseline (from 43.7% to 73.3% WGA) but similar to other baselines is not competitive with the oracle DFR. Using the stronger ConvNeXT architecture, LACER achieves performance (90.7%) matching DFR (90.7%), outperforming other baselines.

5 Related Works

DFR (Kirichenko et al., 2023) demonstrated that group robustness can be improved by retraining the last layer on a group-balanced reweighting dataset, but requires group annotations for the entire held-out set. Subsequent works have attempted to relax this annotation requirement through different approaches. A recent work (Qiu et al., 2023) proposed using a weighted loss during last-layer retraining that emphasizes high-loss examples, thereby implicitly upweighting minority groups. Another recent proposal (LaBonte et al., 2023) leverages predictive differences between ERM-trained models and auxiliary regularized models to create balanced reweighting datasets. However, while these approaches eliminate the need for group annotations during training, they still require access to a group-annotated validation set for hyperparameter tuning—a requirement that can be prohibitive in many real-world scenarios.

Our work is also closely related to George (Sohoni et al., 2022), which estimates group labels through clustering and uses these estimates for group DRO (Sagawa et al., 2020). While we build upon this clustering-based approach, **`LACER`** differs in several key aspects. First, unlike George which performs complete model retraining, our approach follows the more efficient last-layer retraining paradigm. Second, George employs over-clustering to discover minority groups—potentially creating multiple clusters corresponding to majority groups—whereas **`LACER`**'s loss-weighted clustering directly enables balanced group discovery. Additionally, while George alternates between using UMAP or loss components depending on the benchmark for clustering, our approach effectively combines both sources of information in its clustering formulation.

6 Discussion

In this work, we presented **`LACER`**, a simple and effective improvement for last-layer retraining for group robustness by combining two key insights: groups with performance disparities are separable in the feature space, and minority examples typically have higher loss values under the final ERM checkpoint. **`LACER`** significantly enhances our ability to build group robust models in domains where group annotations are restricted due to cost, privacy or fairness concerns. Through extensive experiments, we demonstrated our approach's effectiveness in improving group robustness across varying degrees of imbalance, with particularly strong performance in scenarios with high group imbalance in the data distribution.

There are several **important limitations** of our work. First, while **`LACER`** reduces annotation requirements compared to prior approaches, it still relies on the knowledge of the number of groups present in the dataset. This requirement could limit applicability in scenarios without knowledge of the underlying spurious correlations or group structure in the data.

Future work could explore extending our clustering approach to automatically discover spurious correlations and the number of groups. Additionally, while we demonstrated **`LACER`**'s effectiveness on image classification tasks, exploring its applicability to other domains such as text could further broaden its impact.

REFERENCES

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018. [1]

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. [1]

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. [5]

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532, 2022. [2]

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2023. URL `https://arxiv.org/abs/2204.02937`. [1, 2, 4, 5, 6]

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. *ArXiv*, abs/2309.08534, 2023. URL `https://api.semanticscholar.org/CorpusID:261875640`. [1, 2, 4, 5, 6]

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20071–20082, June 2023. [5, 6]

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021. [2]

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL `https://arxiv.org/abs/2201.03545`. [5]

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015. [5, 6]

Ryan McConville, Raul Santos-Rodriguez, Robert J Piechocki, and Ian Craddock. N2d:(not too) deep clustering via clustering the local manifold of an autoencoded embedding. In *2020 25th international conference on pattern recognition (ICPR)*, pp. 5145–5152. IEEE, 2021. [9]

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. [4]

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting, 2023. URL `https://arxiv.org/abs/2306.11074`. [1, 2, 3, 4, 5, 6]

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. [3, 4]

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. [5]

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL `https://arxiv.org/abs/1911.08731`. [1, 2, 5, 6]

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020. [1]

Nimit S. Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems, 2022. URL `https://arxiv.org/abs/2011.12945`. [2, 3, 6]

UMAP Developers. UMAP: Parameters — Read the Docs, 2025. URL `https://umap-learn.readthedocs.io/en/latest/parameters.html`. Accessed: 2025-02-17. [9]

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, pp. 2953–2961. PMLR, 2024. [2]

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. [2]

A Last-Layer Retraining

Last-layer retraining (LLR) has proven effective in improving group robustness of ERM-trained models. In the second stage of our algorithm, we leverage the estimated group labels from the first stage (§3.1) to construct a group-balanced reweighting dataset for LLR.

Following DFR, we apply ℓ_1 regularization to encourage sparse solutions and eliminate irrelevant features. We tune the regularization hyperparameter (λ) by splitting the validation set in half and use one half to tune the regularization strength. After identifying the optimal regularization strength, we train 20 different logistic regression models using distinct group-balanced reweighting datasets and average the weights of the learned models to ensure robust performance.

B Clustering Details

Following the recommendation by McConville et al. (2021), we apply UMAP for dimensionality reduction preceding clustering. We explore two sets of hyperparameters: embedding dimensions of $\{10, 15\}$ and number of neighbors of $\{5, 10\}$. The optimal configuration is selected based on the silhouette score. More details about these parameters can be found here in UMAP Developers (2025).