Re-evaluating Group Robustness via Adaptive Class-Specific Scaling

Seonguk Seo[1] Bohyung Han[1,2]
[1]ECE & [2]IPAI, Seoul National University
{seonguk, bhhan}@snu.ac.kr

Abstract

Group distributionally robust optimization, which aims to improve robust accuracies—worst-group and unbiased accuracies—is a prominent algorithm used to mitigate spurious correlations and address dataset bias. Although existing approaches have reported improvements in robust accuracies, these gains often come at the cost of average accuracy due to inherent trade-offs. To control this trade-off flexibly and efficiently, we propose a simple class-specific scaling strategy, directly applicable to existing debiasing algorithms with no additional training. We further develop an instance-wise adaptive scaling technique to alleviate this trade-off, even leading to improvements in both robust and average accuracies. Our approach reveals that a naïve ERM baseline matches or even outperforms the recent debiasing methods by simply adopting the class-specific scaling technique. Additionally, we introduce a novel unified metric that quantifies the trade-off between the two accuracies as a scalar value, allowing for a comprehensive evaluation of existing algorithms. By tackling the inherent trade-off and offering a performance landscape, our approach provides valuable insights into robust techniques beyond just robust accuracy. We validate the effectiveness of our framework through experiments across datasets in computer vision and natural language processing domains.

1. Introduction

Machine learning models achieve remarkable performance across various tasks via empirical risk minimization (ERM). However, they are often vulnerable to spurious correlations and dataset biases, resulting in poor classification performance for minority groups despite high average accuracy. For example, in the Colored MNIST dataset [2, 3], a strong correlation exists between digit labels and foreground colors. Consequently, trained models tend to rely on these unintended patterns, resulting in significant performance degradation when classifying digits with rare color associations that are underrepresented in the training data.

Since spurious correlations are well-known to degrade generalization performance in minority groups, group distributionally robust optimization [24] has been widely adopted to address algorithmic bias. Numerous approaches [10, 17, 18, 21, 24, 27, 28] have achieved high robust accuracies, such as worst-group or unbiased accuracies, across various tasks and datasets. However, despite these improvements, they often come at the expense of average accuracy, and little effort has been made to comprehensively evaluate the robust and average accuracies together. Figure 1 demonstrates the trade-offs of existing algorithms.

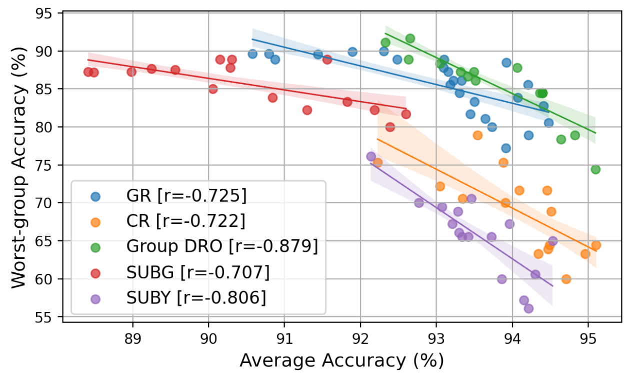
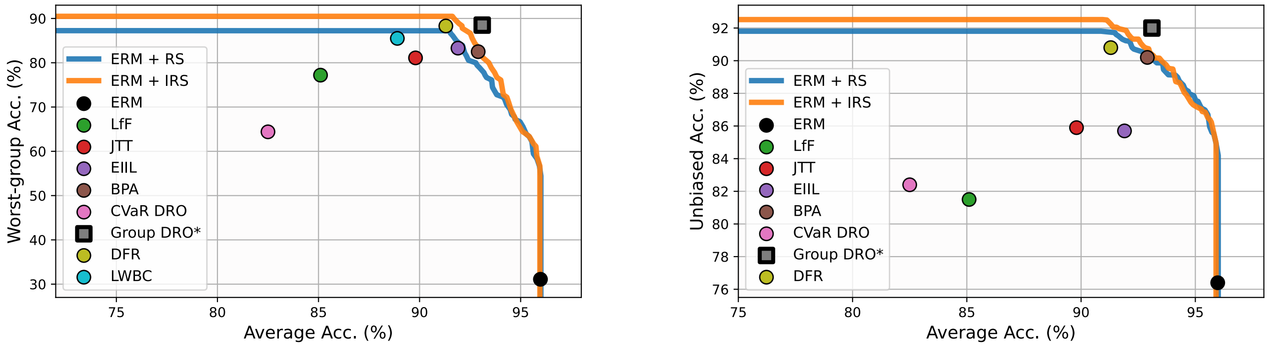


Figure 1. The scatter plots illustrate trade-offs between robust and average accuracies of existing algorithms with ResNet-18 on CelebA. We visualize the results from multiple runs of each algorithm and present the relationship between the two accuracies. The lines denote the linear regression results of individual algorithms and r in the legend indicates the Pearson coefficient correlation.

This paper addresses the limitations of current research trends by introducing a simple post-processing technique, *robust scaling*, which efficiently performs class-specific scaling on prediction scores and conveniently controls the trade-off between robust and average accuracies at test time. It allows us to identify any desired performance points under various metrics such as average accuracy, unbiased accuracy, worst-group accuracy, or balanced accuracy, along the accuracy trade-off curve derived from a pretrained model with negligible extra computational overhead. The proposed robust-scaling method can be seamlessly plugged



(a) Worst-group accuracy

(b) Unbiased accuracy

Figure 2. Comparison between the baseline ERM and existing debiasing approaches with ResNet-50 on CelebA. Existing works have improved robust accuracy substantially compared to ERM, but our robust scaling strategies such as RS and IRS enable ERM to catch up with or even outperform them without further training.

into various existing debiasing algorithms to improve target objectives within the trade-off.

An interesting observation is that, by adopting the proposed robust scaling, even the ERM baseline accomplishes competitive performance without extra training compared to the recent group distributionally robust optimization approaches [5,14,15,17,18,21,24,27,30], as illustrated in Figure 2. Furthermore, we propose an advanced robust scaling algorithm that adaptively applies scaling to individual examples adaptively based on their cluster membership at test time to maximize performance. This instance-wise adaptive scaling strategy effectively mitigates the trade-off and delivers performance improvements in both robust and average accuracies.

By taking advantage of the robust scaling technique, we develop a novel comprehensive evaluation metric that consolidates insights into the trade-off of group robustness algorithms, providing a unique perspective on group distributionally robust optimization. We argue that assessing robust accuracy in isolation, without accounting for average accuracy, provides an incomplete picture and a unified evaluation of debiasing algorithms is required. For a comprehensive performance evaluation, we introduce *robust coverage*, a new measure that effectively captures the trade-off between average and robust accuracies from a Pareto optimal perspective, summarizing each algorithm's performance with a single scalar value.

Contribution We propose a simple yet effective approach for group robustness by analyzing the trade-off between robust and average accuracies. Our framework captures the complete landscape of robust-average accuracy trade-offs, facilitates understanding the behavior of existing debiasing techniques, and enables optimization of arbitrary objectives along the trade-off curve without additional training. We emphasize that our framework does not solely focus on performance improvement in robust accuracy; more importantly, **our method not only highlights the inherent trade-offs in existing debiasing approaches but also facilitates the identification of desired performance points based on target objectives, paving the way for accurate, fair, and comprehensive evaluations of group robustness.** Our main contributions are summarized as follows.

- We propose a training-free class-specific scaling strategy to capture and control the trade-off between robust and average accuracy with negligible computational cost. This approach allows us to optimize a debiasing algorithm towards arbitrary objectives within the trade-off, building on top of any existing models.
- We develop an instance-wise robust scaling algorithm by extending the original class-specific scaling with joint consideration of feature clusters. This technique is effective to alleviate the trade-off and improve both robust and average accuracy.
- We introduce a novel comprehensive and unified performance evaluation metric based on the robust scaling method, which summarizes the trade-off as a scalar value from the Pareto optimal perspective.
- The extensive experiments analyze the characteristics of existing methods and validate the effectiveness of our frameworks on the multiple standard benchmarks.

2. Related Works

Mitigating spurious correlation has emerged as an important problem in many areas in machine learning. Many algorithms are based on the practical assumption that training examples are provided in groups, and that a test distribution is represented as a mixture of these groups. Existing approaches can be categorized into the following three main groups.

Sample reweighting The most popular approaches involve assigning different training weights to each sample to promote minority groups, with the weights determined by either group frequency or loss. Group DRO [24] minimizes the worst-group loss by reweighting samples based on the average loss per group. Although Group DRO achieves robust results against group distribution shifts, it requires training examples with group supervision. To handle this limitation, several unsupervised approaches have been proposed. George [28] and BPA [27] extend Group DRO to an unsupervised setting by initially training an ERM model and subsequently inferring pseudo-groups through feature clustering. CVaR DRO [17] minimizes the worst loss over all α-sized subpopulations, effectively providing an upper bound on the worst-group loss for unknown groups. LfF [21] simultaneously trains two models, one is with generalized cross-entropy and the other is with the standard cross-entropy loss, and reweights the examples based on their relative difficulty score. JTT [18] conducts a two-stage procedure, which upweights the examples that are misclassified by the first-stage model. Idrissi *et al*. [12] analyze simple data subsampling and reweighting baselines based on group or class frequency to handle dataset imbalance issues. LWBC [14] employs an auxiliary module to identify bias-conflicted data and assigns large weights to them.

Representation learning Some approaches aim to learn debiased representations to mitigate spurious correlations directly. ReBias [3] employs the Hilbert-Schmidt independence criterion [7] to ensure feature representations remain independent of predefined biased representations. Cobias [26] measures bias through conditional mutual information between feature representations and group labels and incorporates this metric as a debiasing regularizer. IRM [2] learns invariant representations across diverse environments, where the environment variable is treated as equivalent to the group. While IRM requires supervision for the environment variable, unsupervised alternatives such as EIIL [5] and PGI [1] infer environments by assigning each training example to groups that violate the IRM objective.

Post-processing While most existing approaches focus on in-processing techniques, such as feature representation learning or sample reweighting during training to improve group robustness, our framework stands apart by addressing group robust optimization through a simple post-processing method based on class-specific score scaling, which requires no additional training. Although post-processing techniques like temperature scaling [8] or Platt scaling [23] are popular in confidence calibration, they are unsuitable for our task since they scale prediction scores uniformly across classes and do not alter label predictions. Recently, post-hoc methods have been proposed to retrain the model?s last layer using a group-balanced dataset [15] or adjust the final logits [19], but these approaches still involve additional training, differentiating them from our framework.

3. Proposed Approach

This section first presents our class-specific scaling technique, which captures the trade-off landscape and identifies the optimal performance points for desired objectives along the trade-off curve. We also propose an instance-wise class-specific scaling approach to overcome the trade-off and further improve the performance. Based on the proposed scaling strategy, we introduce a novel and intuitive measure for evaluating the group robustness of an algorithm with consideration of the trade-off.

3.1. Problem Setup

Consider a triplet (x, y, a) with an input feature $x \in \mathcal{X}$, a target label $y \in \mathcal{Y}$, and an attribute $a \in \mathcal{A}$. We define groups based on the pair of a target label and an attribute, such that $g := (y, a) \in \mathcal{Y} \times \mathcal{A} =: \mathcal{G}$. Suppose that the training set consists of n examples without attribute annotations, *e.g.*, $\{(x_1, y_1), ..., (x_n, y_n)\}$, while the validation set includes m examples with group annotations, *e.g.*, $\{(x_1, y_1, a_1), ..., (x_m, y_m, a_m)\}$, for selecting scaling parameters. This assumption is known to be essential for model selection or hyperparameter tuning [12, 18, 21, 24] although not desirable for the practicality of algorithms. However, we will show that our algorithm works well with only a few examples with attribute annotations in the validation set; considering such marginal labeling cost, our approach is a meaningful step to deal with notorious bias problems in datasets and models.

Our goal is to learn a model $f_\theta(\cdot) : \mathcal{X} \to \mathcal{Y}$ that is robust to group distribution shifts. To measure the group robustness, we typically refer to the robust accuracy such as unbiased accuracy (UA) and worst-group accuracy (WA). The definitions of UA and WA require the group-wise accuracy (GA), which is formally given by

$$\mathrm{GA}_{(r)} := \frac{\sum_i \mathbb{1}(f_\theta(\mathbf{x}_i) = y_i, g_i = r)}{\sum_i \mathbb{1}(g_i = r)}, \qquad (1)$$

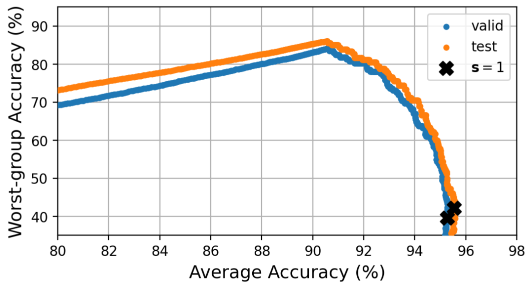
where $\mathbb{1}(\cdot)$ denotes an indicator function and $\mathrm{GA}_{(r)}$ is the accuracy of the r^{th} group samples. Then, the robust accuracies are defined by

$$\mathrm{UA} := \frac{1}{|\mathcal{G}|} \sum_{r \in \mathcal{G}} \mathrm{GA}_{(r)} \quad \text{and} \quad \mathrm{WA} := \min_{r \in \mathcal{G}} \mathrm{GA}_{(r)}. \qquad (2)$$

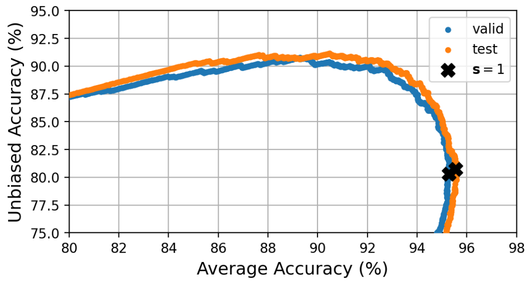
The goal of the group robust optimization is to ensure robust performance in terms of UA or WA regardless of the group membership of a sample.

3.2. Class-Specific Robust Scaling

As illustrated in Figure 1, all algorithms exhibit a clear trade-off between robust accuracy and average accuracy. To



(a) Worst-group accuracy



(b) Unbiased accuracy

Figure 3. The relation between the robust and average accuracies obtained by varying the class-specific scaling factor $\mathbf{s}$ with ERM on CelebA. The black marker denotes the original point, where the uniform scaling is applied.

analyze this behavior more closely, we propose a simple non-uniform scaling method for adjusting the scores associated with individual classes. This approach can influence the final decision of the classifier: by upweighting the prediction scores for minority classes, a sample may be classified into a minority class even if its initial score is low. Consequently, this adjustment can enhance the worst-group accuracy at the cost of a slight reduction in average accuracy, yielding a favorable trade-off for achieving group robustness. Formally, the prediction with the class-specific scaling is given by

$$\arg\max_{c} \ (\mathbf{s} \odot \hat{\mathbf{y}})_c, \tag{3}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$ is a prediction score vector over C classes, $\mathbf{s} \in \mathbb{R}^C$ is a C-dimensional scaling coefficient vector, and $\odot$ denotes the element-wise product operator.

Based on the ERM model, we obtain a set of the average and robust accuracy pairs using a wide range of the class-specific scaling factors and illustrate their relations in Figure 3. The black markers indicate the point with a uniform scaling, *i.e.*, $\mathbf{s} = (1, \ldots, 1) \in \mathbb{R}^C$. The graphs show that a simple class-specific scaling effectively represents the landscape of the trade-off between the two accuracies. This validates the ability to identify the desired Pareto optimal points between the robust and average accuracies in the test set by following two simple steps: 1) finding the optimal class-specific scaling factors that maximize the target objective (UA, WA, or AA) in the validation set, and 2) apply the scaling factors to the test set[1]. We refer to this scaling strategy for robust prediction as *robust scaling*.

To identify the optimal scaling factor $\mathbf{s}$, we perform a greedy search, where we first identify the best scaling factor for a class and then determine the optimal factors of the remaining ones sequentially conditioned on the previously estimated scaling factors. The greedy search is sufficient for finding good scaling factors partly because there are many different near-optimal solutions. Thanks to the simplicity of the process, the entire procedure takes negligible time even in large-scale datasets with multiple classes. It is worth noting that, as a post-processing method, robust scaling can be seamlessly applied to existing robust optimization techniques without requiring additional training. Our approach enables the identification of any desired performance point on the trade-off envelope using a pretrained model. For example, even when dealing with multiple tasks, our robust scaling approach is flexible enough to handle the situation; we only need to apply a scaling factor optimized for each target objective, leaving the trained model unchanged. Meanwhile, existing robust optimization methods have limited flexibility and require to training separate models for each target objective.

3.3. Instance-wise Robust Scaling

The optimal scaling factor can be adaptively applied to each test example, enabling instance-specific scaling to potentially overcome the trade-off and further improve accuracy. Previous approaches [27, 28] have demonstrated the ability to identify hidden spurious attributes by clustering in the feature space for debiased representation learning. Similarly, we take advantage of feature clustering for adaptive robust scaling; we obtain the optimal class-specific scaling factors based on the cluster membership of each sample. The overall algorithm of our instance-wise robust scaling (IRS) is outlined as follows.

1. Perform clustering with the validation dataset on the feature space and store the cluster centroids.
2. Find the optimal scaling factor for each cluster.
3. Apply the estimated scaling factor to each test example based on its cluster membership.

In step 1, we use a simple K-means clustering algorithm. Empirically, when K is sufficiently large, *i.e.*, $K > 10$, IRS achieves stable and superior results, compared to the original class-specific scaling.

[1]Refer to our supplementary document for the coherency of robust scaling in the validation and test sets.

Table 1. Experimental results of the robust scaling (RS) and instance-wise robust scaling (IRS) on the CelebA dataset using ResNet-18 with the average of three runs (standard deviations in parenthesis), where RS and IRS are applied to maximize each target metric independently. *Gain* indicates the average (standard deviations) of performance improvement for each run. On top of all existing approaches, RS can maximize all target metrics consistently and IRS further boosts the performance.

Method	Group Supervision	Robust Coverage		Accuracy (%)					
		Worst-group	Unbiased	Worst-group	(Gain)	Unbiased	(Gain)	Average	(Gain)
ERM		-	-	34.5 (6.1)	-	77.7 (1.8)	-	95.5 (0.4)	-
ERM + RS		83.0 (0.8)	88.1 (0.6)	82.8 (3.3)	+47.7 (7.8)	91.2 (0.5)	+13.3 (2.0)	**95.8 (0.2)**	+0.4 (0.2)
ERM + IRS		**83.4 (0.1)**	**88.4 (0.4)**	**87.2 (2.0)**	+52.7 (3.3)	**91.7 (0.2)**	+13.8 (1.6)	**95.8 (0.1)**	+0.4 (0.3)
CR		-	-	70.6 (6.0)		88.7 (1.2)		94.2 (0.7)	
CR + RS		82.9 (0.5)	88.2 (0.3)	82.7 (5.2)	+12.2 (7.5)	91.0 (1.0)	+2.2 (1.3)	95.4 (0.5)	+1.3 (0.4)
CR + IRS		**83.6 (1.1)**	**88.6 (0.5)**	**84.8 (1.5)**	+14.2 (5.2)	**91.3 (0.4)**	+2.5 (1.4)	**95.5 (0.1)**	+1.3 (0.3)
SUBY [12]		-	-	65.7 (3.9)	-	87.5 (0.9)	-	94.5 (0.7)	-
SUBY + RS		81.5 (1.0)	87.4 (0.1)	80.8 (2.9)	+15.1 (3.0)	90.5 (0.8)	+3.0 (0.9)	95.3 (0.6)	+0.8 (0.6)
SUBY + IRS		**82.3 (1.1)**	**87.8 (0.2)**	**82.3 (2.0)**	+16.5 (4.1)	**90.8 (0.8)**	+3.3 (1.1)	**95.5 (0.3)**	+1.1 (0.4)
LfF [21]		-	-	55.6 (6.6)	-	81.5 (2.8)	-	92.4 (0.8)	-
LfF + RS		74.1 (3.5)	79.7 (2.6)	78.7 (4.1)	+23.2 (2.5)	85.4 (2.4)	+4.0 (0.8)	**93.4 (0.7)**	+1.0 (0.2)
LfF + IRS		**74.6 (4.1)**	**79.8 (3.1)**	**78.9 (5.3)**	+23.4 (4.1)	**86.0 (2.2)**	+4.6 (1.5)	93.1 (1.5)	+0.7 (0.5)
JTT [18]		-	-	75.1 (3.6)	-	85.9 (1.4)	-	89.8 (0.8)	-
JTT + RS		77.3 (0.7)	81.9 (0.7)	82.9 (2.3)	+7.8 (3.0)	87.6 (0.5)	+1.7 (0.4)	90.3 (1.3)	+0.6 (0.1)
JTT + IRS		**78.9 (2.1)**	**82.1 (1.5)**	**84.9 (4.5)**	+9.8 (3.7)	**88.5 (0.8)**	+2.5 (0.8)	**91.0 (1.8)**	+1.2 (0.5)
GR		-	-	88.6 (1.9)	-	92.0 (0.4)	-	92.9 (0.8)	-
GR + RS	✓	86.9 (0.4)	88.4 (0.2)	90.0 (1.6)	+1.4 (1.1)	92.4 (0.5)	+0.5 (0.4)	93.8 (0.4)	+0.8 (0.5)
GR + IRS		**87.0 (0.2)**	**88.6 (0.2)**	**90.0 (2.3)**	+1.4 (1.8)	**92.6 (0.6)**	+0.6 (0.4)	**94.2 (0.3)**	+1.3 (1.0)
SUBG [12]		-	- -	87.8 (1.2)	-	90.4 (1.2)	-	91.9 (0.3)	-
SUBG + RS	✓	83.6 (1.6)	87.5 (0.7)	88.3 (0.7)	+0.5 (0.4)	90.9 (0.5)	+0.5 (0.5)	93.9 (0.2)	+1.9 (0.6)
SUBG + IRS		**84.5 (0.8)**	**87.9 (0.1)**	**88.7 (0.6)**	+0.8 (0.7)	**91.0 (0.3)**	+0.6 (0.9)	**94.0 (0.2)**	+2.1 (1.0)
Group DRO [24]		-	-	88.4 (2.3)	-	92.0 (0.4)	-	93.2 (0.8)	-
Group DRO + RS	✓	87.3 (0.2)	88.3 (0.2)	89.7 (1.2)	+1.4 (1.0)	92.3 (0.1)	+0.4 (0.2)	93.9 (0.3)	+0.7 (0.5)
Group DRO + IRS		**87.5 (0.4)**	**88.4 (0.2)**	**90.0 (2.3)**	+2.6 (1.8)	**92.6 (0.6)**	+0.6 (0.4)	**94.7 (0.3)**	+1.5 (1.1)

3.4. Robust Coverage

Although robust scaling identifies a desired performance point on the trade-off curve, it captures only a single point, overlooking the other Pareto-optimal solutions. To enable a more comprehensive evaluation of an algorithm, we propose a convenient scalar measure that summarizes the robust-average accuracy trade-off. Formally, we define the *robust coverage* as

$$\text{(Robust coverage)} := \int_{c=0}^{1} \max_{\mathbf{s}} \left\{ \text{RA}^{\mathbf{s}} | \text{AA}^{\mathbf{s}} \geq c \right\} dc$$
$$\approx \sum_{d=0}^{D-1} \frac{1}{D} \max_{\mathbf{s}} \left\{ \text{RA}^{\mathbf{s}} | \text{AA}^{\mathbf{s}} \geq \frac{d}{D} \right\}, \quad (4)$$

where $\text{RA}^{\mathbf{s}}$ and $\text{AA}^{\mathbf{s}}$ denote the robust and average accuracies, respectively, and $D = 10^3$ is the number of slices used for discretization. The robust coverage measures the area under the Pareto frontier of the robust-average accuracy trade-off curve, where the maximum operation in (4) identifies the Pareto optimum for each threshold. Depending on the target objective of robust coverage in (3), we use either WA or UA as the measure of RA.

4. Experiments

4.1. Experimental Setup

Implementation details Following prior works, we adopt ResNet-18, ResNet-50 [9], and DenseNet-121 [11], pre-trained on ImageNet [6], as our backbone networks for the CelebA, Waterbirds, and FMoW-WILDS datasets, respectively. For the text classification dataset, CivilComments-WILDS, we use DistillBert [25]. We employ the standard K-means clustering for IRS, where the number of clusters is set to 20, *i.e.*, $K = 20$, for all experiments. We select the final model with the scaling factor that gives the best unbiased coverage in the validation split. Our implementations are based on the Pytorch [22] framework and all experiments are conducted on a single NVIDIA Titan XP GPU. Please refer to our supplementary file for the details about the dataset usage.

Evaluation metrics We evaluate all algorithms in terms of the proposed unbiased and worst-group coverages for comprehensive evaluation, and additionally use the average, unbiased, and worst-group accuracies for comparisons. Following previous works [18,24], we report the adjusted average accuracy instead of the naïve version for the Waterbirds dataset due to its dataset imbalance issue; we first calculate the accuracy for each group and then report the weighted average, where the weights are given by the relative portion

Table 2. Experimental results of RS and IRS on the Waterbirds dataset using ResNet-50 with the average of three runs (standard deviations in parenthesis), where RS and IRS are applied to maximize each target metric independently.

Method	Group Supervision	Robust Coverage Worst-group	Robust Coverage Unbiased	Accuracy (%) Worst-group	(Gain)	Unbiased	(Gain)	Average	(Gain)
ERM		-	-	76.3 (0.8)	-	89.4 (0.6)	-	97.2 (0.2)	-
ERM + RS		76.1 (1.4)	82.6 (1.3)	81.6 (1.9)	+5.3 (1.3)	89.8 (0.5)	+0.4 (0.4)	97.5 (0.1)	+0.4 (0.2)
ERM + IRS		**83.4 (1.1)**	**86.9 (0.4)**	**89.3 (0.5)**	+13.0 (0.9)	**92.7 (0.4)**	+3.3 (0.7)	**97.5 (0.3)**	+0.3 (0.4)
CR		-	-	76.1 (0.7)	-	89.1 (0.7)	-	97.1 (0.3)	-
CR + RS		73.6 (2.3)	82.0 (1.5)	79.4 (2.4)	+3.4 (1.8)	89.4 (1.0)	+0.3 (0.4)	**97.5 (0.3)**	+0.4 (0.1)
CR + IRS		**84.2 (2.5)**	**88.3 (1.0)**	**88.2 (2.7)**	+12.2 (2.1)	**92.1 (0.7)**	+3.1 (0.1)	97.4 (0.2)	+0.3 (0.2)
SUBY [12]		-	-	72.8 (4.1)	-	84.9 (0.4)	-	93.8 (1.5)	-
SUBY + RS		72.5 (1.0)	81.2 (1.4)	75.9 (4.4)	+3.4 (1.8)	86.3 (0.9)	+2.3 (0.9)	95.5 (0.2)	+1.7 (1.1)
SUBY + IRS		**78.8 (2.7)**	**85.9 (1.0)**	**82.1 (4.0)**	+9.3 (1.1)	**89.1 (0.9)**	+4.2 (1.0)	**96.2 (0.6)**	+2.4 (1.4)
GR		-	-	86.1 (1.3)	-	89.3 (0.9)	-	95.1 (1.3)	-
GR + RS	✓	83.7 (0.3)	86.8 (0.7)	**89.3 (1.3)**	+3.2 (2.0)	92.0 (0.7)	+2.7 (1.3)	95.4 (1.3)	+0.4 (0.2)
GR + IRS		**84.8 (1.7)**	**87.4 (0.4)**	89.1 (0.8)	+3.0 (1.6)	**92.2 (1.0)**	+2.9 (1.6)	**95.6 (0.8)**	+0.6 (0.3)
SUBG [12]		-	-	86.5 (0.9)	-	88.2 (1.2)	-	87.3 (1.1)	-
SUBG + RS	✓	80.6 (2.0)	82.3 (2.0)	87.1 (0.7)	+0.6 (0.5)	**88.5 (1.2)**	+0.3 (0.3)	91.3 (0.4)	+4.0 (0.9)
SUBG + IRS		**82.2 (0.8)**	**84.1 (0.8)**	**87.3 (1.3)**	+0.8 (0.6)	88.2 (1.2)	+0.0 (0.2)	**93.5 (0.4)**	+6.2 (1.5)
Group DRO [24]		-	-	88.0 (1.0)	-	92.5 (0.9)	-	95.8 (1.8)	-
Group DRO + RS	✓	83.4 (1.1)	87.4 (1.4)	89.1 (1.7)	+1.1 (0.8)	92.7 (0.8)	+0.2 (0.1)	96.4 (1.5)	+0.5 (0.5)
Group DRO + IRS		**86.3 (2.3)**	**90.1 (2.6)**	**90.8 (1.3)**	+2.8 (1.5)	**93.9 (0.2)**	+1.4 (0.9)	**97.1 (0.4)**	+1.2 (0.8)

of each group in the training set. We ran the experiments three times for each algorithm and report their average and standard deviation.

4.2. Results

CelebA Table 1 presents the experimental results of our robust scaling methods (RS and IRS) on top of the existing approaches including CR, SUBY, LfF, JTT, Group DRO*, GR*, and SUBG*[2] on the CelebA dataset, where '$*$' indicates the method that requires the group supervision in training sets. In this evaluation, RS and IRS choose scaling factors to maximize individual target metrics—worst-group, unbiased, and average accuracies[3]. As shown in the table, our robust scaling strategies consistently improve the performance for all target metrics. In terms of the robust coverage and robust accuracy after scaling, LfF and JTT are not superior to ERM on the CelebA dataset although their robust accuracies without scaling are much higher than ERM. The methods that leverage group supervision such as Group DRO and GR achieve better robust coverage results than the others, which verifies that group supervision helps to improve overall performance. For the group-supervised methods, our scaling technique achieves relatively small performance gains in robust accuracy since the gaps between robust and average accuracies are small and the original results are already close to the optimal robust accuracy. Note that, compared to RS, IRS further boosts the robust coverage and all types of accuracies consistently in all algorithms.

Waterbirds Table 2 demonstrates the outstanding performance of our approaches with all baselines on the Waterbirds dataset. Among the compared algorithms, GR and SUBG are reweighting and subsampling methods based on group frequency, respectively. Although the two baseline approaches exhibit competitive robust accuracy, the average accuracy of SUBG is far below than GR (87.3% vs. 95.1%). This is mainly because SUBG drops a large portion of training samples (95%) to make all groups have the same size, resulting in the significant loss of average accuracy. Subsampling generally helps to achieve high robust accuracy, but it degrades the overall trade-off as well as the average accuracy, consequently hindering the benefits of robust scaling. This observation is coherent to our main claim; the optimization towards the robust accuracy is incomplete and more comprehensive evaluation criteria are required to understand the exact behavior of debiasing algorithms. Note that GR outperforms SUBG in terms of all accuracies after adopting the proposed RS or IRS.

CivilComments-WILDS We also validate the effectiveness of the proposed approach in a large-scale text classification dataset, CivilComments-WILDS [16], which has 8 attribute groups. As shown in Table 3, our robust scaling strategies still achieve meaningful performance improvements for all baselines on this dataset. Although group-supervised baselines such as GR and Group DRO accomplish higher robust accuracies than the ERM without scaling, ERM benefits from RS and IRS greatly. ERM+IRS outperforms both Group DRO and GR in average accuracy

[2] A brief introduction to these methods is provided in the supplementary document.

[3] Since our robust scaling strategy is a simple post-processing method, we do not need to retrain models for each target measure and the cost is negligible, taking only a few seconds for each target metric.

Table 3. Experimental results on the CivilComments-WILDS dataset using a DistilBert architecture with the average of 3 runs.

Method	Group Supervision	Robust Coverage Worst-group	Robust Coverage Unbiased	Accuracy (%) Worst-group	(Gain)	Unbiased	(Gain)	Average	(Gain)
ERM		-	-	54.5 (6.8)	-	75.0 (1.2)	-	92.3 (0.4)	-
ERM + RS		57.2 (5.1)	70.9 (1.5)	65.5 (1.2)	+11.0 (2.5)	78.6 (1.5)	+3.7 (2.4)	**92.5 (0.3)**	+0.2 (0.1)
ERM + IRS		**59.2 (5.3)**	**71.2 (2.3)**	**67.0 (2.3)**	+12.5 (2.7)	**78.8 (1.1)**	+3.8 (1.7)	**92.5 (0.3)**	+0.2 (0.1)
GR		-	-	64.7 (1.1)	-	78.4 (0.2)	-	87.2 (1.0)	-
GR + RS	✓	59.0 (2.8)	69.8 (1.0)	66.0 (0.5)	+1.3 (0.6)	78.5 (0.1)	+0.1 (0.1)	87.9 (0.8)	+0.7 (0.3)
GR + IRS		**59.7 (1.6)**	**70.1 (0.7)**	**66.2 (0.4)**	+1.6 (0.7)	**78.6 (0.1)**	+0.2 (0.2)	**88.4 (0.6)**	+1.2 (0.6)
Group DRO		-	-	67.7 (0.6)	-	78.4 (0.6)	-	90.0 (0.1)	-
Group DRO + RS	✓	60.6 (0.6)	71.5 (0.3)	68.8 (0.7)	+1.1 (0.5)	**78.8 (0.4)**	+0.4 (0.3)	90.5 (0.2)	+0.5 (0.3)
Group DRO + IRS		**62.1 (0.7)**	**71.9 (0.2)**	**69.6 (0.4)**	+1.9 (0.6)	**78.8 (0.5)**	+0.4 (0.6)	**90.8 (0.3)**	+0.8 (0.3)

Table 4. Experimental results on the FMoW-WILDS dataset using a DenseNet-121 architecture with the average of 3 runs.

Method	Group Supervision	Robust Coverage Worst-group	Robust Coverage Unbiased	Accuracy (%) Worst-group	(Gain)	Unbiased	(Gain)	Average	(Gain)
ERM		-	-	34.5 (1.4)	-	51.7 (0.5)	-	52.6 (0.8)	-
ERM + RS		32.9 (0.4)	39.4 (1.3)	35.7 (1.6)	+1.2 (0.4)	52.3 (0.3)	+0.6 (0.3)	53.1 (0.8)	+0.6 (0.3)
ERM + IRS		**35.1 (0.2)**	**40.2 (1.1)**	**36.2 (1.4)**	+1.7 (0.3)	**52.4 (0.2)**	+0.7 (0.4)	**53.4 (0.9)**	+0.8 (0.4)
GR		-	-	31.4 (1.1)	-	49.0 (0.9)	-	50.1 (1.3)	-
GR + RS	✓	30.2 (1.2)	37.7 (0.6)	35.5 (0.4)	+4.2 (0.7)	49.8 (0.7)	+0.8 (0.3)	50.7 (1.2)	+0.6 (0.1)
GR + IRS		**31.7 (1.0)**	**38.9 (2.1)**	**35.7 (0.9)**	+4.4 (0.4)	**50.1 (0.6)**	+1.1 (0.3)	**50.8 (1.4)**	+0.7 (0.1)
Group DRO		-	-	33.7 (2.0)	-	50.4 (0.7)	-	52.0 (0.4)	-
Group DRO + RS	✓	30.8 (1.8)	38.2 (0.7)	36.0 (2.4)	+2.3 (0.4)	50.9 (0.6)	+0.4 (0.4)	52.4 (0.2)	+0.5 (0.2)
Group DRO + IRS		**34.1 (0.8)**	**40.7 (0.5)**	**36.4 (2.3)**	+2.7 (0.4)	**51.1 (0.3)**	+0.7 (0.5)	**52.7 (0.2)**	+0.7 (0.2)

while achieving competitive worst-group and unbiased accuracies, even without group supervision in training samples and extra training.

FMoW-WILDS FMoW-WILDS [16] is a high-resolution satellite imagery dataset with 65 classes and 5 attribute groups, which involves domain shift issues as train, validation, and test splits come from different years. We report the results from our experiments in Table 4, which shows that GR and Group DRO have inferior performance even compared with ERM. On the other hand, our robust scaling methods do not suffer from any performance degradation and even enhance all kinds of accuracies substantially. This fact supports the strengths and robustness of our framework in more challenging datasets with distribution shifts.

4.3. Analysis

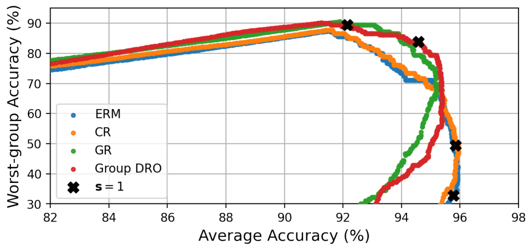
Validation set sizes We analyze the impact of the validation set size on the robustness of our algorithm. Table 5 presents the ERM results on the CelebA dataset by varying the validation set size to $\{100\%, 50\%, 10\%, 1\%\}$ of its full size. Note that other approaches also require validation sets with group annotations for early stopping and hyperparameter tuning, which are essential to achieve high robust accuracy. As shown in the table, with only 10% or 50% of the validation set, both RS and IRS achieve almost equivalent performance to the versions with the entire validation set. Surprisingly, even only 1% of the validation

Table 5. Ablation study on the size of validation set in our robust scaling strategies on CelebA.

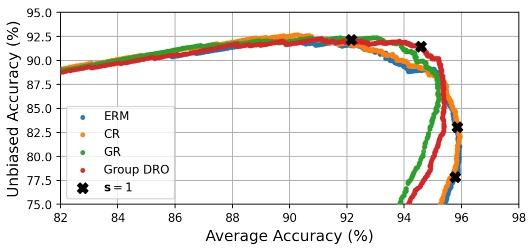
Method	Valid set size	Worst-group	Gain	Unbiased	Gain
ERM	-	34.5 (6.1)	-	77.7 (1.8)	-
+ RS	100%	82.8 (3.3)	**+48.3**	91.2 (0.5)	**+13.5**
+ RS	50%	83.3 (3.7)	**+48.8**	91.5 (0.9)	**+13.8**
+ RS	10%	82.4 (4.3)	**+48.0**	91.4 (0.8)	**+13.7**
+ RS	1%	79.2 (10.3)	**+44.7**	90.8 (2.2)	**+13.1**
+ IRS	100%	88.7 (0.9)	**+54.2**	92.0 (0.3)	**+14.3**
+ IRS	50%	86.9 (2.0)	**+52.4**	91.8 (0.4)	**+14.1**
+ IRS	10%	84.4 (6.3)	**+50.0**	91.4 (1.0)	**+13.7**
+ IRS	1%	60.4 (14.4)	**+25.9**	85.8 (3.2)	**+8.0**

set is enough for RS to gain sufficiently high robust accuracy but inevitably entails a large variance of results. On the other hand, IRS suffers from performance degradation when only 1% of the validation set is available. This is mainly because IRS takes advantage of feature clustering on the validation set, which would need more examples for stable results. In overall, our robust scaling strategies generally improve performance substantially even with a limited number of validation examples with group annotations for all cases.

Accuracy trade-off Figure 4 depicts the robust-average accuracy trade-offs of several existing algorithms on the CelebA dataset. The black markers denote the points without scaling, implying that there is room for improvement in robust accuracy along the trade-off curve.



(a) With the worst-group accuracy

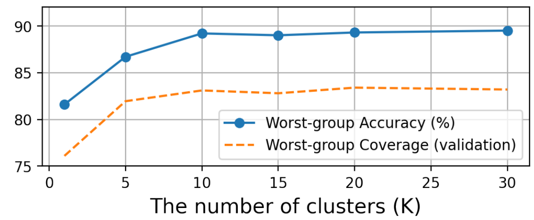


(b) With the unbiased accuracy

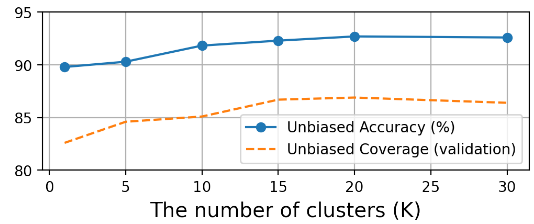
Figure 4. The robust-average accuracy trade-off curves of various baselines on the CelebA dataset. The black marker denotes the original point, where the uniform scaling is applied.

Number of clusters We adjust the number of clusters for feature clustering in IRS on the Waterbirds dataset. Figure 5 illustrates that the worst-group and unbiased accuracies gradually improve as K increases and are stable with a sufficiently large $K(> 10)$. The leftmost point ($K = 1$) denotes RS in each figure. We also plot the robust coverage results in the validation split, which are almost consistent with the robust accuracy measured in the test dataset.

Comparison to reweighting or resampling techniques As mentioned in Section 2, most existing debiasing techniques [12,15,18,21,24,27], in principle, perform reweighting and/or resampling of training data. Our approach has a similar idea, but, instead of giving favor to the examples in minority groups during training and boosting their classification scores indirectly via iterative model updates, we directly adjust their classification scores by class-wise scaling after training, thus it gives similar but clearer effects on the results. As shown in Figure 4, although class reweighting (CR) improves the robust accuracy, this in fact identifies one of the Pareto optimal points on the trade-off curve of ERM obtained by class-specific scaling. However, because class reweighting employs a single fixed reweighting factor during training based on class frequency, it only reflects a single point and has limited flexibility compared to our wide range of scaling search. If CR employs a wide range of reweighting factors, then it can identify additional optimal points and achieve additional performance gains, but it requires training separate models for each factor, which is not realistic. Note that our method can be easily applied to



(a) With the worst-group accuracy



(b) With the unbiased accuracy

Figure 5. Sensitivity analysis with respect to the number of clusters in IRS on Waterbirds. The tendency of the robust coverage in the validation split (orange) is similar with the robust accuracy in the test split (blue).

CR or other methods, which allows us to identify more desirable optimal points on the trade-off curve with negligible computational overhead.

5. Conclusion

We presented a simple but effective post-processing method that provides a novel perspective of group robustness. Our work starts from the observation that there exists a clear trade-off between robust and average accuracies in existing works. From this observation, we first proposed the robust scaling strategy, which captures the full trade-off landscape and identifies any desired performance point on the trade-off curve with no extra training. Moreover, we proposed an instance-wise robust scaling algorithm that is effective to enhance the trade-off itself. Based on these strategies, we introduced a novel convenient measure that summarizes the trade-off from a Pareto optimal perspective for a comprehensive evaluation of group robustness. We believe that our approaches are helpful for analyzing the exact behavior of existing debiasing methods and paving the way in the future research direction.

Acknowledgements This work was partly supported by Samsung Advanced Institute of Technology (SAIT), and by the Institute of Information communications Technology Planning & Evaluation (IITP) grants [No.RS-2022-II220959 (No.2022-0-00959), No.RS-2021-II211343, No.RS-2021-II212068], and the National Research Foundation of Korea (NRF) grant [No.RS-2022-NR070855], funded by the Korean government (MSIT).

References

[1] Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *ICLR*, 2020. 3

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 3

[3] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020. 1, 3

[4] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 13

[5] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML*, 2021. 2, 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[7] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, 2005. 3

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 5

[10] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 1

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 5

[12] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *CLeaR*, 2022. 3, 5, 6, 8, 11

[13] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. In *WACV*, 2021. 15

[14] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. *arXiv preprint arXiv:2206.10843*, 2022. 2, 3

[15] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 2, 3, 8

[16] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 6, 7, 12, 13

[17] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *NeurIPS*, 2020. 1, 2, 3

[18] 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 *ICML*, 2021. 1, 2, 3, 5, 8, 11

[19] Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. 2023. 3

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 12

[21] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *NeurIPS*, 2020. 1, 2, 3, 5, 8, 11

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[23] John Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advanced in Large Margin Classifiers*, 10, 06 2000. 3

[24] 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. In *ICLR*, 2020. 1, 2, 3, 5, 6, 8, 11, 12

[25] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5

[26] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In *AAAI*, 2022. 3

[27] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *CVPR*, 2022. 1, 2, 3, 4, 8

[28] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020. 1, 3, 4

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 12

[30] 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

[31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2017. 12