A. Comparisons

Below is a brief introduction of the comparisons used in our experiments.

ERM Given a loss function $\ell(\cdot)$, the objective of empirical risk minimization is optimizing the following loss over training data:

$$\min_{\theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) \Big\}. \tag{5}$$

Class reweighting (CR) To mitigate the class imbalance issue, we can simply reweight the samples based on the inverse of class frequency in the training split,

$$\min_{\theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i \ell(f_\theta(x_i), y_i) \Big\} \text{ where } \omega_i = \frac{n}{\sum_j \mathbb{1}(y_j = y_i)}. \tag{6}$$

LfF Motivated by the observation that bias-aligned samples are more easily learned, LfF [21] simultaneously trains a pair of neural network (f_B, f_D). The biased model f_B is trained with generalized cross-entropy loss which intends to amplify bias, while the debiased model f_D is trained with a standard cross-entropy loss, where each sample (x_i, y_i) is reweighted by the following relative difficulty score:

$$\omega_i = \frac{\ell(f_\theta^B(x_i), y_i)}{\ell(f_\theta^B(x_i), y_i) + \ell(f_\theta^D(x_i), y_i)}. \tag{7}$$

JTT JTT [18] consists of two-stage procedures. In the first stage, JTT trains a standard ERM model $\hat{f}(\cdot)$ for several epochs and identifies an error set E of training examples that are misclassified:

$$E := \{(x_i, y_i) \text{ s.t. } \hat{f}(x_i) \neq y_i\}. \tag{8}$$

Next, they train a final model $f_\theta(\cdot)$ by upweighting the examples in the error set E as

$$\min_{\theta} \Big\{ \lambda_{\text{up}} \sum_{(x,y)\in E} \ell(f_\theta(x), y) + \sum_{(x,y)\notin E} \ell(f_\theta(x), y) \Big\}. \tag{9}$$

Group DRO Group DRO [24] aims to minimize the empirical worst-group loss formulated as:

$$\min_{\theta} \Big\{ \max_{g\in\mathcal{G}} \frac{1}{n_g} \sum_{i|g_i=g}^{n_g} \ell(f_\theta(x_i), y_i) \Big\} \tag{10}$$

where n_g is the number of samples assigned to g^{th} group. Unlike previous approaches, group DRO requires group annotations $g = (y, a)$ on the training split.

Group reweighting (GR) Using group annotations, we can extend class reweighting method to group reweighting one based on the inverse of group frequency in the training split, *i.e.*,

$$\min_{\theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i \ell(f_\theta(x_i), y_i) \Big\} \quad \text{where } \omega_i = \frac{n}{\sum_j \mathbb{1}(y_j = y_i, a_j = a_i)} \tag{11}$$

SUBY/SUBG To mitigate the data imbalance issue, SUBY subsample majority classes, so all classes have the same size with the smallest class on the training dataset, as in [12]. Similarly, SUBG subsample majority groups.

Table A6. Results of the attribute-specific robust scaling (ARS) on the CelebA and Waterbirds datasets with the average of three runs (standard deviations in parenthesis), where ARS is applied to maximize each target metric independently. Note that our post-processing strategy, ARS, allows ERM to achieve competitive performance to Group DRO that utilizes the group supervision during training.

Dataset	Method	Robust Coverage		Accuracy (%)		
		Worst.	Unbiased	Worst.	Unbiased	Average
CelebA	ERM	-	-	34.5 (6.1)	77.7 (1.8)	95.5 (0.4)
	ERM + ARS	**87.6 (1.0)**	**89.0 (0.2)**	**88.5 (1.8)**	91.9 (0.3)	**95.8 (0.1)**
	Group DRO	87.3 (0.2)	88.3 (0.2)	88.4 (2.3)	**92.0 (0.4)**	93.2 (0.8)
Waterbirds	ERM	-	-	76.3 (0.8)	89.4 (0.6)	97.2 (0.2)
	ERM + ARS	**84.4 (1.9)**	**87.8 (1.7)**	**89.3 (0.4)**	**92.5 (0.4)**	**97.5 (1.0)**
	Group DRO	83.4 (1.1)	87.4 (2.3)	88.0 (1.0)	**92.5 (0.9)**	95.8 (1.8)

B. Attribute-specific Robust Scaling with Group Supervision

If the supervision of group (spurious-attribute) information can be utilized during our robust scaling, it will provide flexibility to further improve the performance. To this end, we first partition the examples based on the values of spurious attributes and find the optimal scaling factors for each partition separately. Like as the original robust scaling procedure, we obtain the optimal scaling factors for each partition in the validation split and apply them to the test split. However, this partition-wise scaling is basically unavailable because we do not know the spurious attribute values of the examples in the test split and thus cannot partition them, In other words, we need to estimate the spurious-attribute values in the test split for partitioning. To conduct attribute-specific robust scaling (ARS), we follow a simple algorithm described below:

1. Partition the examples in the validation split by the values of the spurious attribute.
2. Find the optimal scaling factors for each partition in the validation split.
3. Train an independent estimator model to classify spurious attribute.
4. Estimate the spurious attribute values of the examples in the test split using the estimator, and partition the test samples according to their estimated spurious attribute values.
5. For each sample in the test split, apply the optimal scaling factors obtained in step 2 based on its partition.

To find a set of scale factors corresponding to each partition, we adopt a naïve greedy algorithm that performed in one partition at a time. This attribute-specific robust scaling further increases the robust accuracy compared to the original robust scaling, and also improves the robust coverage, as shown in Table A6. Note that our attribute-specific scaling strategy allows ERM to match the supervised state-of-the-art approach, Group DRO [24].

One limitation is that it requires the supervision of spurious attribute information to train the estimator model in step 3. However, we notice that only a very few examples with the supervision is enough to train the spurious-attribute estimator, because it is much easier to learn as the word "spurious correlation" suggests. To determine how much the group-labeled data is needed, we train several spurious-attribute estimators by varying the number of group-labeled examples, and conduct ARS using the estimators. Table A7 validates that, compared to the overall training dataset size, a very small amount of group-labeled examples is enough to achieve high robust accuracy.

C. Experimental Details

C.1. Datasets

CelebA [20] is a large-scale dataset for face image recognition, consisting of 202,599 celebrity images, with 40 attributes labeled on each image. Among the attributes, we primarily examine *hair color* and *gender* attributes as a target and spurious attributes, respectively. We follow the original train-validation-test split [20] for all experiments in the paper. Waterbirds [24] is a synthesized dataset, which are created by combining bird images in the CUB dataset [29] and background images from the Places dataset [31], consisting of 4,795 training examples. The two attributes—one is the type of bird, {waterbird, landbird} and the other is background places, {water, land}, are used for the experiments with this dataset. CivilComments-WILDS [16] is a large-scale text dataset, which has 269,038 training comments, 45,180 validation comments, and 133,782 test comments.

Table A7. Effects of the size of group-labeled examples on the attribute-specific robust scaling on the CelebA dataset. Group-labeled size denotes a ratio of group-labeled samples among all training examples for training estimators. Spurious accuracy indicates the average accuracy of spurious-attribute classification using the estimators on the test split.

	Accuracy (%)	Accuracy (%)			Robust Coverage	
Group-labeled size	Spurious	Worst-group	Unbiased	Average	Worst-group	Unbiased
100%	98.4	89.1 (3.0)	92.4 (1.1)	93.1 (1.2)	87.6 (1.0)	89.0 (0.5)
10%	97.7	88.5 (1.8)	91.9 (0.3)	92.8 (0.6)	86.8 (0.4)	89.0 (0.2)
1%	95.8	88.5 (1.8)	91.9 (0.3)	92.9 (0.6)	87.1 (0.3)	89.0 (0.2)
0.1%	92.6	88.4 (2.1)	91.8 (0.5)	92.4 (0.8)	87.1 (0.3)	89.0 (0.2)

Table A8. Realized robust coverage results on the Waterbirds and CelebA datasets with the average of three runs (standard deviations in parenthesis).

		Robust Coverage		Realized Robust Coverage	
Dataset	Method	Worst-group	Unbiased	Worst-group	Unbiased
Waterbirds	ERM	70.3 (1.3)	79.4 (0.7)	69.0 (1.5)	78.7 (0.8)
Waterbirds	CR	68.9 (1.1)	78.5 (0.5)	67.8 (1.2)	77.9 (0.4)
Waterbirds	Group DRO	80.8 (0.6)	85.2 (0.1)	78.6 (1.0)	83.8 (0.4)
Waterbirds	GR	78.8 (5.6)	83.7 (0.7)	77.9 (1.4)	82.8 (0.8)
CelebA	ERM	78.9 (1.7)	86.0 (0.6)	75.9 (2.2)	85.4 (0.7)
CelebA	CR	77.2 (2.8)	85.6 (0.9)	71.8 (1.3)	85.0 (0.6)
CelebA	Group DRO	84.2 (0.6)	86.7 (0.5)	81.0 (1.7)	86.1 (0.2)
CelebA	GR	84.2 (0.5)	87.5 (0.3)	81.2 (1.6)	87.0 (0.5)

This task is to classify whether an online comment is toxic or not, which is spuriously correlated to demographic identities (*male, female, White, Black, LGBTQ, Muslim, Christian, and other religion*). FMoW-WILDS [16] is based on the Functional Map of the World dataset [4], comprising high-resolution satellite images from over 200 countries and over the years 2002-2018. The label is one of 62 building or land use categories, and the attribute represents both the year and geographical regions (*Africa, the Americas, Oceania, Asia, or Europe*). It consists of 76,863 training images from the years 2002-2013, 19,915 validation images from the years 2013-2016, and 22,108 test images from the years 2016-2018.

C.2. Class-specific Scaling

To identify the optimal points, we obtain a set of the average and robust accuracy pairs using a wide range of the class-specific scaling factors, *i.e.*, $\mathbf{s}_i = (1.05)^n$ for $-200 \leq n \leq 200$ for i^{th} class. Note that we search for the scaling factor of each class in a greedy manner, as stated in Section 3.2.

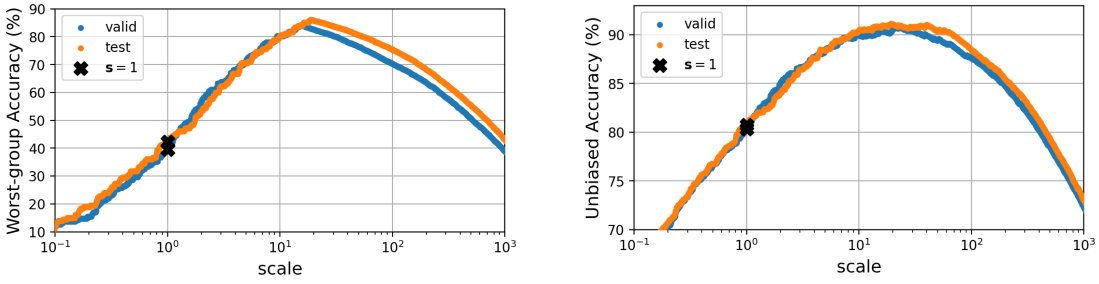
C.3. Hyperparameter Tuning

We tune the learning rate in $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and the weight decay in $\{1.0, 0.1, 10^{-2}, 10^{-4}\}$ for all baselines on all datasets. We used 0.5 of q for LfF. For JTT, we searched λ_{up} in $\{20, 50, 100\}$ and updated the error set every epoch for CelebA dataset and every 60 epochs for Waterbirds dataset. For Group DRO, we tuned C in $\{0, 1, 2, 3, 4\}$, and used 0.1 of η.

D. Additional Results

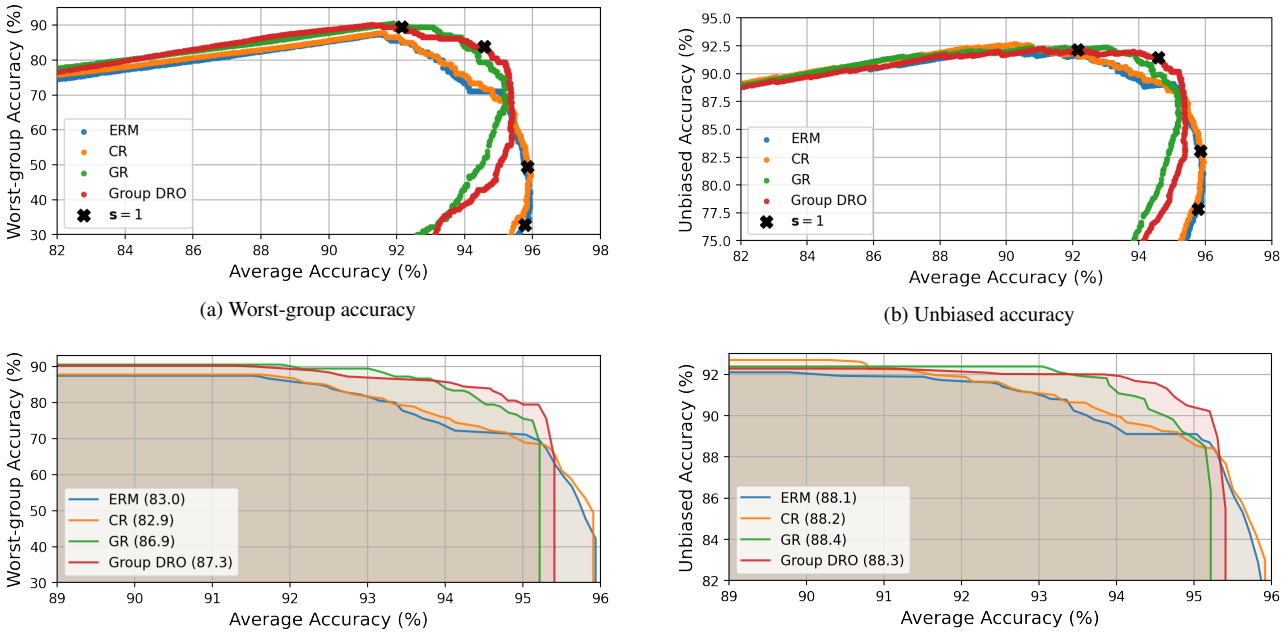
Feasibility We visualize the relationship between scaling factors and robust accuracies in Figure A6, where the curves are constructed based on validation and test splits are sufficiently well-aligned to each other. This implies that the optimal scaling factor identified in the validation set can be used in the test set to get the final robust prediction.

Robust coverage curve Figure A7a and A7b are robust-average accuracy trade-off curves while Figure A7c and A7d are their corresponding robust coverage curves, which represent the Pareto frontiers of Figure A7a and A7b, respectively. The area under the curve in Figure A7c and A7d indicates the robust coverage of each algorithm.



(a) Worst-group accuracy (b) Unbiased accuracy

Figure A6. Effects of varying the class-specific scaling factors on the robust accuracy using ERM model on the CelebA dataset. Since this experiment is based on the binary classifier, a single scaling factor is varied with the other fixed to one. These results show that the optimal scaling factor identified in the validation set can be used in the test set to get the final robust prediction.



(a) Worst-group accuracy (b) Unbiased accuracy

(c) Worst-group accuracy (d) Unbiased accuracy

Figure A7. The robust-average accuracy trade-off curves ((a), (b)) and their corresponding robust coverage curves ((c), (d)), respectively, on the CelebA dataset. The curves in (c) and (d) represent the Pareto frontiers of the curves in (a) and (b), respectively. In (c) and (d), the numbers in the legend denote the robust coverage, which measures the area under the curve.

Table A9. Variations of robust scaling methods and their performances tested on the FairFace dataset.

Method	Cost	Worst-group	Unbiased	Average
ERM	–	15.8	47.0	54.1
+ RS (2 super classes)	$\mathcal{O}(n)$	18.6	51.8	52.9
+ RS (greedy search)	$\mathcal{O}(n)$	**19.2**	52.3	**53.3**
+ RS (full grid search)	$\mathcal{O}(n^9)$	19.0	**52.8**	53.1

Scalability As mentioned in Section 3.2, we search for the scaling factor of each class in a greedy manner. Hence, the time complexity increases linearly with respect to the number of classes instead of the exponential growth with the full grid search; even with 1,000 classes, the whole process takes less than a few minutes in practice, which is negligible compared to the model training time. Moreover, we can reduce the computational cost even further by introducing the superclass concept

Table A10. Results of our robust scaling methods on top of various baselines on the CelebA dataset, which supplement Table 1. Blue color denotes the target metric that the robust scaling aims to maximize. Compared to RS, IRS improves the overall trade-off.

	Robust Coverage		Accuracy (%)			Accuracy (%)		
Method	Worst-group	Unbiased	Worst-group	Unbiased	Average	Worst-group	Unbiased	Average
ERM	-	-	34.5 (6.1)	77.7 (1.8)	**95.5 (0.4)**	34.5 (6.1)	77.7 (1.8)	95.5 (0.4)
ERM + RS	83.0 (0.7)	88.1 (0.5)	82.1 (3.7)	91.1 (0.6)	92.2 (1.3)	**45.0 (7.4)**	**81.7 (1.8)**	**95.8 (0.2)**
ERM + IRS	**83.4 (0.1)**	**88.4 (0.4)**	**87.2 (2.0)**	**91.7 (0.2)**	91.5 (0.8)	44.1 (4.2)	81.3(0.8)	**95.8 (0.1)**
CR	-	-	70.6 (6.0)	88.7 (1.2)	**94.2 (0.7)**	**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)	91.0 (1.0)	91.7 (1.3)	48.5 (8.9)	82.5 (2.2)	**95.8 (0.1)**
CR + IRS	**83.6 (1.1)**	**88.6 (0.5)**	**84.8 (1.5)**	**91.3 (0.4)**	90.7 (1.3)	48.8 (9.1)	82.7 (2.4)	**95.8 (0.1)**
SUBY	-	-	65.7 (3.9)	87.5 (0.9)	**94.5 (0.7)**	**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)	90.5 (0.8)	91.1 (1.7)	45.4 (6.7)	81.4 (2.0)	**95.5 (0.0)**
SUBY + IRS	**82.3 (1.1)**	**87.8 (0.2)**	**82.3 (2.0)**	**90.8 (0.8)**	90.7 (1.9)	46.0 (6.9)	81.5 (2.1)	**95.5 (0.1)**
SUBG	-	-	87.8 (1.2)	90.4 (1.2)	**91.9 (0.3)**	**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)	90.9 (0.5)	90.6 (1.0)	67.8 (6.5)	85.2 (2.0)	93.9 (0.2)
SUBG + IRS	**84.5 (0.8)**	**87.9 (0.1)**	**88.7 (0.6)**	**91.0 (0.3)**	90.6 (0.8)	68.5 (6.5)	85.5 (1.9)	**94.0 (0.2)**
GR	-	-	88.6 (1.9)	92.0 (0.4)	**92.9 (0.8)**	**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)**	92.4 (0.5)	92.5 (0.5)	66.5 (0.3)	85.4 (0.4)	93.8 (0.4)
GR + IRS	**87.0 (0.2)**	**88.6 (0.2)**	**90.0 (2.3)**	**92.6 (0.6)**	92.5 (0.4)	62.0 (5.3)	84.5 (0.7)	**94.2 (0.3)**
GroupDRO	-	-	88.4 (2.3)	92.0 (0.4)	93.2 (0.8)	**88.4 (2.3)**	**92.0 (0.4)**	93.2 (0.8)
GroupDRO + RS	87.3 (0.2)	88.3 (0.2)	89.7 (1.2)	92.3 (0.1)	**93.7 (0.5)**	64.9 (3.3)	85.1 (0.7)	93.9 (0.3)
GroupDRO + IRS	**87.5 (0.4)**	**88.4 (0.2)**	**90.0 (2.3)**	**92.6 (0.6)**	93.5 (0.4)	60.4 (5.4)	84.4 (0.6)	**94.7 (0.3)**

Table A11. Results of our robust scaling methods on top of various baselines on the Waterbirds dataset, which supplement Table 2. Blue color denotes the target metric that the robust scaling aims to maximize. Compared to RS, IRS improves the overall trade-off.

	Robust Coverage		Accuracy (%)			Accuracy (%)		
Method	Worst-group	Unbiased	Worst-group	Unbiased	Average	Worst-group	Unbiased	Average
ERM	-	-	76.3 (0.8)	89.4 (0.6)	**97.2 (0.2)**	76.3 (0.8)	89.4 (0.6)	97.2 (0.2)
ERM + RS	76.1 (0.4)	82.6 (0.3)	81.6 (1.9)	89.8 (0.5)	**97.2 (0.2)**	**79.1 (2.7)**	**89.7 (0.6)**	97.5 (0.1)
ERM + IRS	**83.4 (1.1)**	**86.9 (0.4)**	**89.3 (0.5)**	**92.7 (0.4)**	94.1 (0.3)	77.6 (7.0)	89.6 (1.1)	**97.5 (0.3)**
CR	-	-	76.1 (0.7)	89.1 (0.7)	**97.1 (0.5)**	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)	89.4 (1.0)	96.8(0.8)	76.4(1.5)	**89.3 (0.8)**	**97.5 (0.3)**
CR + IRS	**84.2 (2.5)**	**88.3 (1.0)**	**88.2 (2.7)**	**92. 1(0.7)**	95.7 (1.1)	**77.3 (4.7)**	88.6(1.2)	97.4 (0.2)
SUBY	-	-	72.8 (4.1)	84.9 (0.4)	93.8 (1.5)	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)	86.3 (0.9)	**95.2 (1.4)**	70.7 (5.8)	85.4 (1.6)	95.5 (0.2)
SUBY + IRS	**78.8 (2.7)**	**85.9 (1.0)**	**82.1 (4.0)**	**89.1 (0.9)**	92.6 (2.2)	**74.1 (4.1)**	**86.3 (0.9)**	**96.2 (0.6)**
SUBG	-	-	86.5 (0.9)	88.2 (1.2)	87.3 (1.1)	**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)	**88.5 (1.2)**	**87.9 (1.1)**	74.0 (5.6)	85.9 (2.8)	91.3 (0.4)
SUBG + IRS	**82.2 (0.8)**	**84.1 (0.8)**	**87.3 (1.3)**	88.2 (1.2)	87.6 (1.2)	70.2 (1.6)	84.5 (1.0)	**93.5 (0.4)**
GR	-	-	86.1 (1.3)	89.3 (0.9)	**95.1 (1.3)**	**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)**	92.0 (0.7)	93.1 (3.2)	82.2 (1.3)	**90.8 (0.5)**	95.4 (1.3)
GR + IRS	**84.8 (1.7)**	**87.4 (0.4)**	89.1 (0.8)	**92.2 (1.0)**	92.9 (2.1)	82.1 (1.4)	90.5 (0.7)	**95.6 (0.8)**
GroupDRO	-	-	88.0 (1.0)	92.5 (0.9)	95.8 (1.8)	**88.0 (1.0)**	**92.5 (0.9)**	95.8 (1.8)
GroupDRO + RS	83.4 (1.1)	87.4 (1.4)	89.1 (1.7)	92.7 (0.8)	**96.4 (1.5)**	80.9 (4.4)	91.3 (1.0)	**97.1 (0.3)**
GroupDRO + IRS	**86.3 (2.3)**	**90.1 (2.6)**	**90.8 (1.3)**	**93.9 (0.2)**	96.0 (0.6)	83.2 (1.7)	91.5 (0.8)	**97.1 (0.4)**

and allocating a single scaling factor for each superclass. We compare three different options—greedy search, superclass-level search, and full grid search—on the FairFace dataset [13] with 9 classes. Table A9 shows that the greedy search is as competitive as the full grid search despite the time complexity reduction by several orders of magnitude and the superclass-level search is also effective to reduce cost with competitive accuracies. Note that the superclasses are identified by the feature similarity of class signatures.

Additional Results Table A11 presents full experimental results on the CelebA and Waterbirds datasets, which supplement Table 1 and 2. We test our robust scaling strategies (RS, IRS) with two scenarios, each aimed at maximizing worst-group or average accuracies, respectively, where each target metric is marked in blue in the tables.

E. Discussion

Limitation Although our framework is simple yet effective for improving target metrics with no extra training, it does not learn debiased representations as it is a post-processing method. However, this suggests that existing training approaches may also not actually learn debiased representations, but rather focus on prediction adjustment for group robustness in terms of robust accuracy. From this point of view, our comprehensive measurement enables a more accurate and fairer evaluation of base algorithms, considering the full landscape of trade-off curve.