Adaptive Group Robust Ensemble Knowledge Distillation

Anonymous authors
Paper under double-blind review

Abstract

Neural networks can learn spurious correlations in the data, often leading to performance degradation for underrepresented subgroups. Studies have demonstrated that the disparity is amplified when knowledge is distilled from a complex teacher model to a relatively "simple" student model. Prior work has shown that ensemble deep learning methods can improve the performance of the worst-case subgroups; however, it is unclear if this advantage carries over when distilling knowledge from an ensemble of teachers, especially when the teacher models are debiased. This study demonstrates that traditional ensemble knowledge distillation can significantly drop the performance of the worst-case subgroups in the distilled student model even when the teacher models are debiased. To overcome this, we propose Adaptive Group Robust Ensemble Knowledge Distillation (AGRE-KD), a simple ensembling strategy to ensure that the student model receives knowledge beneficial for unknown underrepresented subgroups. Leveraging an additional biased model, our method selectively chooses teachers whose knowledge would better improve the worst-performing subgroups by upweighting the teachers with gradient directions deviating from the biased model. Our experiments on several datasets demonstrate the superiority of the proposed ensemble distillation technique and show that it can even outperform classic model ensembles based on majority voting.

1 Introduction

When trained with empirical risk minimization (ERM), neural networks are susceptible to capturing spurious correlations in the data (Tiwari & Shenoy, 2023), which are features that correlate with but not causally related to the class label (Qiu et al., 2023). In particular, the class label might spuriously correlate with patterns in the data that are easier to learn than the intended pattern. For example, in the Waterbirds dataset (Sagawa et al., 2019), which contains images of landbirds and waterbirds, most landbirds are located in a land background, and waterbirds in a water background. Instead of predicting the actual bird species, the model trained with ERM can achieve high accuracy by relying on the background to make predictions. This results in a significantly higher error for underrepresented subgroups that do not exhibit the spurious correlation (e.g., landbird on water background and waterbird on land background). Several works have shown that model compression methods such as pruning (Hooker et al., 2020), and knowledge distillation (Lukasik et al., 2023; Lee & Lee, 2023; Wang et al., 2023) can exacerbate the performance disparities between different subgroups. In knowledge distillation (KD), a network with a smaller capacity (student) is trained using the output of a higher capacity network (teacher) (Hinton et al., 2015). While KD can improve the student model's average performance, the gain is not uniform across subgroups (Lukasik et al., 2023).

On the other hand, deep ensemble models have been shown to enhance generalization performance compared to individual models (Ganaie et al., 2022), and simple deep ensemble models with the same architecture, objective, and optimization settings can attenuate this shortcoming and improve the worst-case group performance (Ko et al., 2023; Kenfack et al., 2021). However, evaluating several models at test time can be computationally expensive, making them less practical for deployment on edge devices. To address this issue, ensemble knowledge distillation involves distilling the knowledge of multiple teachers to a single student model (You et al., 2017; Radwan et al., 2024), and it remains unclear whether distilling from an ensemble of teachers improves the student's worst-group performance.

This paper studies how knowledge distillation from multiple teachers impacts underrepresented subgroups. We investigate whether the subgroup performance gains of deep ensemble models apply to ensemble knowledge distillation. Focusing on logit distillation, we consider ERM-trained (biased) and debiased teachers obtained by deep feature reweighting (DFR). DFR is a method proposed by Kirichenko et al. (2022), showing that simply retraining the last layer of the ERM-trained model using a small held-out validation set of group-balanced data can mitigate the spurious correlation. In this work, we also investigate whether the student model can learn debiased representations from the teacher's output, whose only the last layer was retrained to mitigate bias.

The rationale behind focusing on logit distillation for investigating bias in ensemble knowledge is also motivated by the study by Izmailov et al. (2022), where the authors demonstrated that several debiasing methods do not learn better feature representation but correctly weight core features in the last classification layer, which reduce the reliance on spurious features for predictions. Our results reveal that underrepresented subgroups can be negatively impacted when training the student model using the aggregated outputs of multiple teachers. Other ensemble distillation approaches have been designed to boost the student's performance by modelling a better aggregation of the teachers' knowledge (Du et al., 2020; Zhang et al., 2022). In particular, the ensemble distillation method proposed by Du et al. (2020) aims to find a better compromise when teachers have conflicting predictions, and the method by Zhang et al. (2022) ensures distillation is done only using teachers with confident predictions. Our results show that these ensemble distillation methods do not effectively fix the performance disparity across groups in the student model.

We therefore propose an Adaptive Group Robust Ensemble Knowledge Distillation (AGRE-KD) method to encourage the student to improve the performance of unknown worst-case subgroups. Specifically, our method relies on a model that has captured the spurious correlation (i.e., a biased model) to guide the teachers' outputs aggregation process and ensure the student model does not capture biased knowledge from teachers. Prior work has relied on a reference classifier to target and upweight samples from unknown worst-case, using the errors of the reference classifier (Liu et al., 2021; Nam et al., 2020) or its per-sample gradient magnitude (Ahn et al., 2022). In contrast, our proposal uses the gradient direction of the biased model to select and weigh teachers' outputs adaptively during the training.

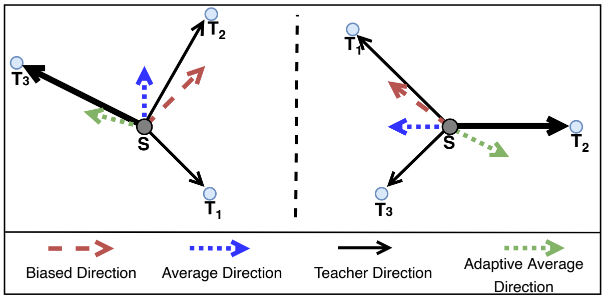


Figure 1: Illustration of our adaptive weighting method based on gradient direction. The bolder lines indicate the teacher's higher weight in the aggregated output.

Intuitively, training the student model solely following the gradient direction that minimizes its KD loss with a biased teacher can result in local/global minima with a higher loss for the underrepresented groups (Lukasik et al., 2021). Our results suggest this behaviour can be exacerbated in ensemble knowledge distillation since the inherent consensus from the teachers' gradient direction can be dominated by a direction that minimizes the average error at the expense of the worst-group error. Our proposed methods compare the gradient direction of student loss with the biased model and each teacher in the ensemble, and upweight teachers whose gradient direction deviates the most from the biased model.Figure 1 illustrates the gradient directions of the student loss with the biased model and three teachers and shows how classic averaging yields a direction that mostly aligns with the biased teachers, while our adaptive weight prioritizes the least biased models. As the gradient magnitude can be very noisy, we compare gradient directions by computing the dot product of their normalized vectors (i.e., the cosine similarity), which removes the influence of the gradient magnitude. The contributions of this paper can be summarized as follows:

- We demonstrate empirically that ensemble knowledge distillation amplifies the performance drop for the worst-case group, contrary to deep ensemble models. We attribute this to the reduced capacity of the student network by showing that ensemble self-distillation using models with the same capacity improves robustness to spurious correlation and achieves comparable performance with the teachers.

- We show that distilling knowledge from teacher models with only their last layer retrained to mitigate bias provides significantly more robust knowledge to the student model.
- We propose a novel gradient-based weighting scheme to ensure the distillation process minimizes the student's loss towards better worst-case group error. The proposed method utilizes a model that has learned the spurious correlation (biased model) to orchestrate the distillation process in the gradient space.
- We perform intensive experiments on three well-known benchmarks, and the results demonstrate the superiority of the proposed method.

2 Background

We consider a multiclass classification task using a given the training $D = \{(x_i, y_i, a_i)\}_{i=1}^m$, where $x_i \in \mathcal{X}$ is the input feature, $y_i \in \mathcal{Y}$ the target variable with $c = |\mathcal{Y}|$ classes, and $a_i \in \mathcal{A}$ an unknown spurious feature. We aim to build a classifier $h(\cdot)$ that accurately predicts the class of the unlabelled test dataset. The classifier uses mapping function $f : \mathcal{X} \rightarrow \mathbb{R}^c$, that assigns scores $[\sigma_1(z_1), \ldots, \sigma_c(z_c)]$, such that z is the output logits of the a given sample x and $\sigma_y(z)$ the softmax function defined as $\sigma_y(z) = \frac{\exp(z_y)}{\sum_{j\in[c]} \exp(z_j)}$, $\forall y \in [c]$.
The classifier is derived by predicting the class label that maximizes the softmax, $h(x) = \text{argmax}_{j\in[c]}\, \sigma_j(z)$. We evaluate the classifier's performance during training using a loss function, such as the softmax cross-entropy loss function, which measures how accurately samples are classified. When the attribute a spuriously correlates with the target y, ERM-trained models can strongly rely on the spurious feature and fail to generalize on the subgroups where the spurious correlation does not hold (Ye et al., 2024)

Knowledge Distillation (KD). In KD, a student network f^s aims to achieve performance close to the higher-capacity network by mimicking the teacher model f^t Hinton et al. (2015). In practice, we train the student model to mimic the teacher's output by minimizing the Kullback-Leibler (KL) divergence between their outputs, defined as follows:

$$\mathcal{L}_{\text{KD}} = \tau^2 \cdot \text{KL}(\sigma(\frac{z^s}{\tau})\ ,\ \sigma(\frac{z^t}{\tau})) \tag{1}$$

where z^s and z^t are the student and the teacher logits, respectively. τ is the temperature parameter controlling the smoothness of the probability distribution for more fine-grained information. The student loss can be combined with the classification loss on the ground truth label using the following double loss,

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{KD}} + (1 - \alpha) \cdot \mathcal{L}_{\text{cls}} \tag{2}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss (e.g., cross-entropy loss) between the student's output and the ground truth label (y), and α is a hyperparameter controlling the classification loss and knowledge distillation loss. In other KD techniques, the student network is enforced to mimic the teacher's internal representation instead of only outputs (Romero et al., 2014). Transferring knowledge from intermediary representation (feature-level) can provide more fine-grained information and boost the students' performance Romero et al. (2014). On the other hand, recent studies have shown that models trained with ERM still learn core features while spurious features are only amplified in the last classifier layer (Kirichenko et al., 2022; Qiu et al., 2023; Izmailov et al., 2022). In this regard, we restrict ourselves to logit distillation and leave feature distillation for future exploration.

Deep Ensemble. Ensemble learning is a widely used technique to boost the generalization performance of a model (Allen-Zhu & Li, 2020). Studies have shown that training several independent models and averaging their predictions at test time can yield a model that outperforms each individual model in the ensemble (Ganaie et al., 2022). However, evaluating multiple models for predictions at test time limits the practical use of deep ensembles due to computational overhead. Knowledge distillation can address this issue by enforcing the student network to mimic the ensemble's output.

Ensemble KD. Like in ensemble learning, distilling knowledge from multiple teachers instead of a single one is expected to improve the student model (You et al., 2017). In ensemble KD, we train the student model using the averaged softened output or the averaged knowledge distillation losses of M teachers[1] as follows:

$$\mathcal{L}_{\text{ensKD}} = \tau^2 \cdot \text{KL}(\sigma(\frac{z^s}{\tau}) \ , \ \frac{1}{M}\sum_{m=1}^{M} \sigma(\frac{z^m}{\tau})) \tag{3}$$

This ensemble KD loss can be plugged in equation 2 for training the student with KD from an ensemble of M teachers. In contrast to these methods, our method is fully unsupervised, both based in terms of ground truth class label and group information. Related works more aligned with our approach are unsupervised methods that leverage teachers' diverse knowledge to improve students' performance. (You et al., 2017; Du et al., 2020; Fukuda et al., 2017; Zhang et al., 2022; Kwon et al., 2020). For example, Fukuda et al. (2017) suggest that randomly selecting a teacher during mini-batch training can allow the student model to capture complementary knowledge of teachers. Other authors argue that simply averaging teachers' softmax outputs can mislead the student model, particularly when there is competition or contradictions between teachers. In this regard, Zhang et al. (2022) proposes a sample-wise weighting for teacher loss based on the confidence of the teacher's prediction compared to the ground label. Other methods consider label-free weighing schemes by comparing teachers in the gradient space. For instance, Du et al. (2020) use a multi-objective optimization approach in the gradient space to find teachers that agree the most in the gradient direction that minimizes their loss with the student model. Similarly, Zhou et al. (2024) sample-wise teachers selection by only averaging the majority of teachers with the same gradient directions. Our method also operates within gradient space to enhance students' resilience to spurious correlation. Our experiments demonstrate that adhering to the majority of teachers' opinions does not always benefit the underrepresented subgroups.

3 Related Work

Bias mitigation without group label. We consider settings where samples in the dataset D are associated with *unknown* group labels that spuriously correlate with the class label. For example, in the CelebA dataset, the class 'hair color' (blond, non-blond) correlates with the gender (male, female) since most images with blond hair belong to the female group. Neural networks can capture this correlation, resulting in poor performance for certain subgroups (e.g., blond males) (Sagawa et al., 2019). We aim to ensure that the model does not capture spurious correlations in the data and accurately classifies samples from all subgroups. In particular, we measure the model's bias using the *worst-case group accuracy* (WGA), i.e., group of samples where the spurious correlation does not apply. Several methods have been proposed to mitigate these biases in individual models (Kenfack et al., 2024a; Ye et al., 2024). When the group information is known, Sagawa et al. (2019) propose Group Distributionally Robust Optimization (DRO) that minimizes the loss of the group experiencing the maximum loss. However, group information can be costly to collect or unavailable due to privacy restrictions (Kenfack et al., 2024b). Several methods have been proposed in this setting to improve the worst-case group performance without group labels (Ye et al., 2024). For instance, Liu et al. (2021) and Nam et al. (2020) rely on a reference classifier to target and upweight worst-performing subgroups. These methods use the mistakes of the reference classifier to improve group robustness by up-weighting misclassified samples (Liu et al., 2021). The reference classifier is generally trained to amplify the misclassification of samples from the unknown worst-performing group (Nam et al., 2020). While these methods do not use group labels during the training, they require access to a small validation set with group labels for model selection and intensive hyperparameter tuning (Kenfack et al., 2024a). Kirichenko et al. (2022) proposed *Deep Feature Reweighting* (DRF) for training group robust model by retraining the last layer of ERM model with a small subset of a held-out group-balanced dataset, which achieves state-of-the-art worst-case group performance. However, all debiasing methods use neural networks of high capacity, and they are less effective on models with smaller capacity Izmailov et al. (2022). Furthermore, the effect of distilling these debiased models on the student model's robustness remains underexplored. This work builds on Kirichenko et al. (2022) for debiasing the teacher models by retraining their last layers using a held-out group balanced with

[1] Du et al. (2020) showed that the averaged softened output is equal to averaged KL loss.

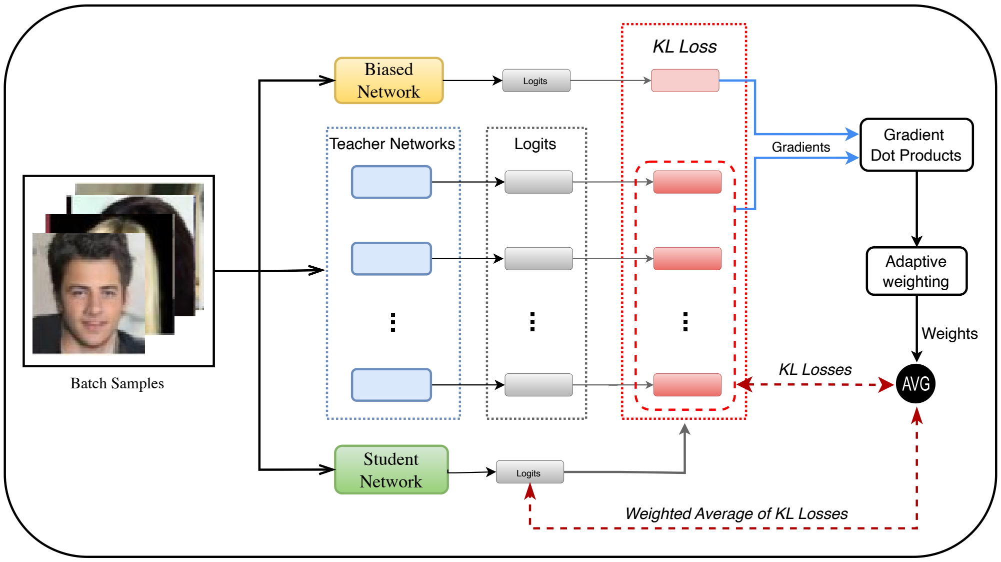


Figure 2: Overview of AGRE-KD.

fewer group annotations, and we investigate whether distilling knowledge from an ensemble of (debiased) teachers can lead to more robust student models.

Bias in Knowledge Distillation. Several works have studied bias in knowledge distillation with a single teacher model (Lukasik et al., 2023; Lee & Lee, 2023; Lukasik et al., 2023; Tiwari et al., 2024; Bassi et al., 2024). In particular, Lukasik et al. (2023) demonstrates that teacher errors can be amplified by the student during distillation and proposed a mitigation strategy that distills only the confident predictions of the teacher. However, their study focuses on worst-class errors and KD with a single teacher, while we study worst-subgroup errors with multiple teachers. Lee & Lee (2023) propose an adapted version of *Simple knowledge distillation* (SimKD (Chen et al., 2022)) that transplants the last layer of the teacher to student and only distills features. With the teacher trained with Group DRO, they show that transplanting the teacher's last layer to the student only improves the worst-case group performance if the feature distillation is performed by upweighting misclassified sample from a reference classifier (Lee & Lee, 2023). However, their method uses the reference classifier by Liu et al. (2021), where the efficiency of sample weighting requires intensive tuning of the number of epochs to train the reference classifier. Similarly, Tiwari et al. (2024) uses the earlier layers of the neural network to train a reference classifier and shows that it improves the recall of worst group samples within the misclassification set, which are upweighted in the KD loss. However, their method also requires class labels to derive the misclassified samples.. In Lukasik et al. (2023), the authors study where it is best to apply the debiasing mechanism (Group DRO) and conclude that applying the robust loss to both the teacher and student model improves the average performance along with the worst-case group performance. In contrast to these prior works, we study bias in knowledge distillation with multiple teachers **without group information and class labels**. We investigate whether the subgroup's performance gain observed in deep ensemble models also applies when the knowledge of the ensemble is distilled to a single model. We aim to achieve better worst-case performance across subgroups when aggregating the outputs of multiple teachers in knowledge distillation. To the best of our knowledge, this represents the first study on bias in ensemble knowledge distillation.

4 Adaptive Group Robust Ensemble Knowledge Distillation (AGRE-KD)

In this work, we considered each teacher in the ensemble to have the same architecture and trained using different random initializations. Prior work has shown ensemble models with different random initializations are diverse enough to improve the performances (Ganaie et al., 2022). The first contribution of this work is showing that, unlike deep ensemble models that can improve the WGA (Ko et al., 2023), ensemble knowledge distillation tends to amplify bias in the distilled model.

To address this problem, we propose AGRE-KD, an adaptive ensembling knowledge distillation strategy that ensures the student model captures robust knowledge from the teachers. AGRE-KD relies on a model pretrained with ERM that captured the dataset's spurious correlation (biased model). Intuitively, suppose a student model takes gradient steps toward the direction that minimizes its KD loss with the *biased* model. In that case, the resulting student model will likely capture and even amplify the reliance on the spurious correlation, i.e., the local/global minimum in that direction likely provides the worst performance for the underrepresented subgroups. Following this intuition, we derive a weighting scheme that prioritizes teacher models whose gradient direction disagrees with the biased ERM model. Figure 2 provides an overview of our proposed method.

Gradient-based weighting scheme. For a given minibatch, we compute the student KD loss regarding the biased model b and each teacher t. Before aggregating the teachers' outputs, we compute sample weights based on the similarity between the gradient direction of each teacher and the biased model. Consider $\ell_i^t(\theta)$ the KD loss on sample i regarding the t-teacher, and $\ell_i^b(\theta)$ KD loss regarding the biased model, with θ the parameters of the student model. The dot product ($\langle\cdot,\cdot\rangle$) between the *normalized*[2] gradients of the student KD loss with the teacher and the biased model indicates which teachers align with the biased model in gradient space. In particular, when $\langle\nabla\ell_i^t(\theta), \nabla\ell_i^b(\theta)\rangle > 0$ both models have the same gradient directions and their gradients are in the opposite direction when $\langle\nabla\ell_i^t(\theta), \nabla\ell_i^b(\theta)\rangle < 0$. In extreme cases, when the dot product of the gradients is closer to 1, both models have precisely the same directions; if the dot product is close to -1, the models are in opposite directions. Thus, to penalize teachers with a gradient direction closer to the biased model, we downweight teachers' output for samples having dot products get closer to -1 and upweight samples as their dot products get closer to 1. More formally, we compute the sample-wise teacher's weight as follows:

$$W_t(x_i) = 1 - \langle\nabla\ell_i^t(\theta), \nabla\ell_i^b(\theta)\rangle \tag{4}$$

Adaptive knowledge distillation. The weighting scheme in Equation 4 suggests that teacher models who behave similarly to the biased model in the gradient space will have less influence on the aggregated outputs. Therefore, we aggregate the teacher's outputs using the sample-wise weighted average KD (wKD) loss defined as follows:

$$\mathcal{L}_{\text{wKD}} = \frac{W_t(x_i) \cdot \mathcal{L}_{\text{KD}}}{\sum_t W_t} \tag{5}$$

And we train the student model with the following final loss $\mathcal{L} = \alpha\mathcal{L}_{\text{wKD}} + (1-\alpha)\mathcal{L}_{\text{cls}}$

As aforementioned, we focus on unsupervised knowledge distillation, i.e., we set $\alpha = 1$ to investigate the robustness of the "dark knowledge" provided by an ensemble of eventually debiased teachers. Supplementary A provides a detailed algorithm of the training process of AGRE-KD. Note that the use of an additional biased model to orchestrate the distillation process does not add extra complexity compared to related work using a reference classifier to identify the worst-performing subgroups (Nam et al., 2020; Liu et al., 2021; Kenfack et al., 2024a). Furthermore, as we will see in the experiments, our proposed method is robust to the choice of biased model architecture trained using ERM without.

5 Experimental Results

In this section, we present the experimental setup and the empirical results. We experiment with several ensemble KD techniques and analyze their impact on the word-group accuracy of the student model. We compare our adaptive ensemble knowledge distillation to these methods and demonstrate its effectiveness in improving the student model's worst-group performance.

[2] Table 5 in the Supplementary shows the importance of ignoring the magnitude of the gradients.

5.1 Setup

We evaluate the worst-case performance of the proposed method on three classification tasks: two from the vision domain (Waterbirds and CelebA datasets) and one from the language domain (CivilComments)

- **Waterbirds** (Sagawa et al., 2019; Liu et al., 2021) is a dataset of birds derived from Caltech-UCSD Birds (CUB) (Wah et al., 2011) by synthetically creating a spurious correlation between bird species and the background. In particular, the class label is the type of bird appearing in the image (waterbirds, landbirds), and the background landscape (water, land) spuriously correlates with the bird type. Here, the minority subgroups represent images with the background landscape not aligned with the bird type, i.e., {`waterbird`, `land background`} and {`landbird`, `water background`}.

- **CelebA** (Liu et al., 2015) dataset contains images of celebrities with 40 facial attributes. In this dataset, the attribute `hair color` is spuriously correlated `gender`. We consider hair color {`blond`, `non-blond`} as the class label and gender {`male`, `female`} as group information.

- **CivilComments** (Koh et al., 2021) is a textual dataset collected from online comments. The task is to predict whether a comment is `toxic` or `non-toxic`. The label is spuriously correlated with comments related to some demographic subgroups such as gender (male, female), race (white, black), and sexual orientation (LGBT). We consider a binary indicator of comments related to these demographic subgroups as spurious group information.

Network Architecture and Training. Following prior work (Tiwari et al., 2024; Lee & Lee, 2023), we use the Resnet-18 (He et al., 2016) architecture for the student model and the Resnet-50 (He et al., 2016) architecture for the teacher models for the forvision tasks. Both networks are pretrained on the ImageNet-1K (Russakovsky et al., 2015) dataset. For the language task, we used the BERT (Devlin et al., 2019) model for teachers and the DistilBERT (Sanh et al., 2019) for the student model; and language models are pretrained on Book Corpus and English Wikipedia. Following related works on KD (Du et al., 2020; Fukuda et al., 2017; Chen et al., 2022), we set the temperature hyperparameter $\tau = 4$ (Eq. 1) and show in an ablation study in Supplementary D (Figure 4) that increasing τ can exert positive effect on WGA up to certain values. We provide further details about hyperparameters in the Supplementary B.

Teacher training. We train each teacher model in the ensemble independently using standard ERM and cross-entropy loss with the ground truth class labels. Teacher models have the same architecture and hyperparameters and only differ in random seeds used for weights initialization; prior studies have shown that independent training with different random weight initializations can provide models sufficient diversity to improve the performance of the model ensemble (Allen-Zhu & Li, 2020; Ganaie et al., 2022). As aforementioned, we obtain debiased teachers using the approach introduced by Kirichenko et al. (2022) to retrain the last layer of the ERM model on the small proportion of the held-out group-balanced dataset. We perform the retraining step of the last layer using group-balanced batch sampling instead of the averaging models trained over group-balanced subsets of the data (LaBonte et al., 2024). This debiasing process is very simple and computationally inexpensive since it involves training a logistic regression model on a smaller dataset.

For the *biased* model used by our method to compute teachers's weights, we randomly select one teacher model trained with ERM and without DFR. We provide in Supplementary D (Table 4) experiment with different choices of biased model showing the robustness of the proposed gradient-based weighting to the biased model choice. Following Kirichenko et al. (2022); LaBonte et al. (2024), we use half of the validation set of each benchmark to perform last-layer retraining with DFR; as we do not use the group and class labels during the KD training, we do not perform further hyperparameter tuning or model selection.

Baselines. In addition to the standard training process using the one-hot class label for training each teacher model (Section 2), we consider other ensemble knowledge distillation methods aiming to improve the student's performance. In particular, we consider the following baseline:

Models	Debiased	KD	Waterbirds		CelebA		CivilComment	
			Average	WGA	Average	WGA	Average	WGA
Teacher	✗	—	$85.7_{\pm 1.80}$	$65.6_{\pm 5.75}$	$95.4_{\pm 0.07}$	$37.5_{\pm 2.27}$	$90.0_{\pm 0.25}$	$75.9_{\pm 1.15}$
	✓	—	$94.2_{\pm 0.53}$	$90.9_{\pm 1.00}$	$93.7_{\pm 2.44}$	$90.1_{\pm 1.68}$	$85.9_{\pm 0.49}$	$77.9_{\pm 0.49}$
Deep Ensemble	✗	—	$84.4_{\pm 0.00}$	$59.3_{\pm 0.44}$	$95.6_{\pm 0.02}$	$37.2_{\pm 0.00}$	$90.6_{\pm 0.02}$	$76.1_{\pm 0.10}$
	✓	—	$93.5_{\pm 0.17}$	$90.0_{\pm 0.56}$	$92.4_{\pm 0.18}$	$90.3_{\pm 0.55}$	$85.9_{\pm 0.17}$	$76.6_{\pm 0.09}$
One-hot	✗	✗	$83.8_{\pm 0.97}$	$54.1_{\pm 2.21}$	$95.5_{\pm 0.07}$	$32.3_{\pm 2.66}$	$90.1_{\pm 0.21}$	$75.5_{\pm 0.84}$
	✓	✗	$91.4_{\pm 1.31}$	$86.7_{\pm 1.85}$	$92.3_{\pm 0.39}$	$88.9_{\pm 2.90}$	$85.7_{\pm 0.47}$	$76.5_{\pm 1.09}$
Random	✗	✓	$80.0_{\pm 0.37}$	$39.6_{\pm 3.63}$	$95.2_{\pm 0.03}$	$27.2_{\pm 0.00}$	$90.8_{\pm 0.21}$	$\underline{75.2}_{\pm 0.95}$
	✓	✓	$89.6_{\pm 0.71}$	$77.3_{\pm 2.23}$	$\mathbf{92.5}_{\pm 0.29}$	$85.1_{\pm 0.84}$	$\mathbf{91.0}_{\pm 0.18}$	$74.2_{\pm 1.04}$
AVER	✗	✓	$79.2_{\pm 0.80}$	$46.4_{\pm 2.39}$	$\underline{95.5}_{\pm 0.05}$	$28.8_{\pm 0.78}$	$\underline{90.9}_{\pm 0.03}$	$74.7_{\pm 1.16}$
	✓	✓	$90.8_{\pm 2.44}$	$82.9_{\pm 1.23}$	$92.4_{\pm 0.25}$	$83.4_{\pm 0.45}$	$90.8_{\pm 0.07}$	$75.0_{\pm 0.73}$
AE-KD	✗	✓	$81.7_{\pm 0.80}$	$46.9_{\pm 7.57}$	$95.3_{\pm 0.06}$	$30.5_{\pm 1.88}$	$90.8_{\pm 0.54}$	$73.1_{\pm 3.05}$
	✓	✓	$90.9_{\pm 1.72}$	$85.0_{\pm 1.23}$	$92.3_{\pm 0.26}$	$87.5_{\pm 1.17}$	$90.7_{\pm 0.23}$	$74.8_{\pm 1.11}$
AGRE-KD(Ours)	✗	✓	$\underline{82.2}_{\pm 1.37}$	$\underline{55.0}_{\pm 5.47}$	$95.4_{\pm 0.04}$	$\underline{37.6}_{\pm 0.78}$	$89.3_{\pm 3.65}$	$74.7_{\pm 3.00}$
	✓	✓	$\mathbf{91.3}_{\pm 0.49}$	$\mathbf{87.9}_{\pm 1.23}$	$91.7_{\pm 0.20}$	$\mathbf{91.9}_{\pm 0.71}$	$90.2_{\pm 0.49}$	$\mathbf{75.9}_{\pm 1.75}$

Table 1: **Comparison of ensemble KD methods**. We report the average and worst-group test accuracy (WGA) on each dataset. The *Debiased* column indicates whether the model involves debiasing with DFR or whether teacher models are debiased when the *KD* column is checked (✓). Bolded represents the best-performing student with the debiased teacher ensemble and underlined represents the best-performing with the biased teacher ensemble.

- **Deep Ensemble**: This baseline corresponds to deep ensembling using a majority voting scheme of models with the same capacity as the student model. In particular, given a set of models, the predicted class label represents the class that received the most votes for models in the ensemble.

- **KD with averaged teachers outputs** (AVER) (You et al., 2017): Here, we perform standard knowledge distillation following equation 3 by minimizing the KL loss between the student's softmax output and the averaged softened outputs (dark knowledge) from teachers.

- **Random** (Fukuda et al., 2017): During each mini-batch training, this method randomly selects a teacher model from the ensemble to train the student model. Fukuda et al. (2017) referred to this technique as *switched-training* as the weights of the student model are updated by switching across teacher labels at the minibatch level.

- **AE-KD** (Du et al., 2020): This method is an adaptive ensembling distillation technique closest to ours. However, the method postulates that when teachers have conflicting gradient directions, a multi-objective optimization problem is solved to select the gradient direction (teachers) that satisfies most of the teachers in the ensemble.

5.2 Results

We train each model using three random seeds and report the means and standard deviations. We consider ensemble distillation with ten teacher models randomly sampled from a poll of pretrained (*biased*) teachers across independent random seeds. In addition to the teacher's performance and deep ensemble approach, we report the performance of the "student" trained only using the ground truth class labels (One-hot), i.e., without knowledge distillation. We report the average accuracy and WGA. We consider models trained using ERM or with last-layer retraining for debiasing. Specifically, for ensemble knowledge distillation methods, we train the student model with *biased teachers* (ERM model) and *debiased teachers* (last layer retrained with

DFR (Kirichenko et al., 2022)). Table 1 summarizes the main results of the paper from which we draw the following observations:

- **Ensemble knowledge distillation amplifies bias.** From the results in Table 1, we observe that deep ensemble using models with the same architecture as the student model improves the worst-case group performance, while for all ensemble knowledge distillation, the worst-case group performance can drop up to 10%. The drop is more pronounced for the random ensemble distillation but improved average accuracy, which suggests switching between teachers during the training harms worst-case groups. On the other hand, AE-KD (Du et al., 2020) performs better than other standard ensemble KD methods, showing the effectiveness of the proposed gradient-based adaptive distillation for the worst group samples.
- **DFR teachers lead to robust student models.** When we train the teacher models using ERM and then retrain the classification layer with the group-balanced set, the resulting student models achieve significantly better worst-case group accuracy. We only train the student model using the aggregated teachers' outputs without any feature distillation. These results demonstrate that the retrained classifications can provide pseudo-label distribution (dark knowledge) that reduces students' reliance on spurious features. This shows that by mimicking the teacher's outputs, the classifier layer of the student model also downweights the spurious features in its last layers (Izmailov et al., 2022). As the temperature parameter (τ) also alters the teacher's output by softening its predicted probability distribution, we provide an ablation study (Appendix C) showing that increasing τ positively impacts the worst-group accuracy. On the other hand, the improved students' WGA across ensemble distillation methods still does not match the WGA of the debiased teacher models. We attribute this to the smaller capacity of the student model, which we discuss in the next experiment.
- **Our weighting scheme can consistently improves the WGA compared to other ensembling methods.** On the CelebA dataset, we achieve better worst-case group accuracy compared to the deep ensembling of models with the same capacity as the student. This observation is consistent with settings where teacher models are biased or debiased. On the other hand, we can see that our method does not provide significant improvement for the worst-case group when all teacher models in the ensemble are biased; we attribute this to the fact all these teachers likely have a similar gradient direction with the biased model for many samples. Nevertheless, our weighting scheme still identifies and upweights the least biased teachers. In the Supplementary D (Table 6), we provide the group-wise accuracy of each baseline method on the Waterbirds and CelebA datasets.

5.2.1 Impact of the model capacity.

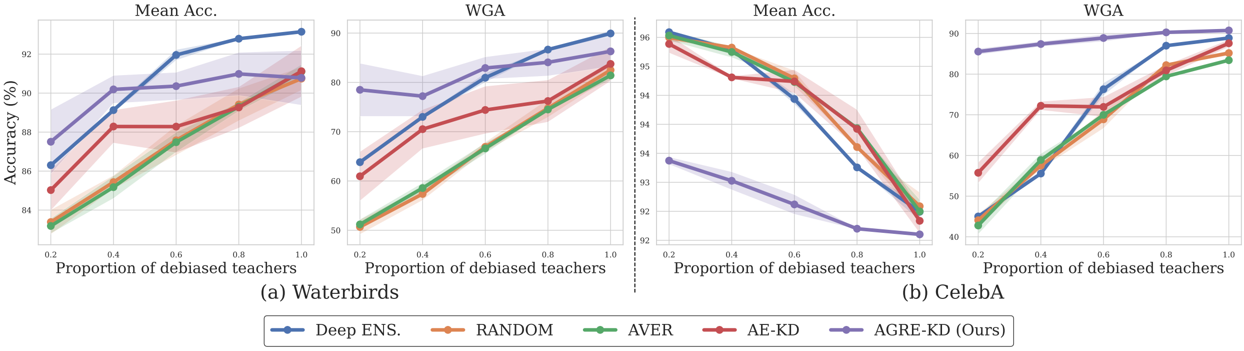
In this experiment, we study whether the student's network capacity is a source of the implication of spurious feature learning in ensemble KD. In particular, we perform the same experiments as previously but using self-distillation, i.e., we use the same architecture for the student and teacher models (e.g., resnet50 on image tasks). Results in Table 2 show that ensemble KD methods with self-distillation significantly improve the worst-group test accuracy. The gap between the teacher models and students is reduced compared to KD settings in Table 1, where the students' models have smaller capacity. The results indicate that the students' higher capacity can help the network learn more core features and reduce the influence of spurious features in the last layer. On the other hand, AGRE-KD outperforms other baselines, showing that our adaptive weighting scheme effectively guides the student models to focus on minimizing worst-case group error during training. We further illustrate the effectiveness of our weighting scheme in the next experiment by adjusting the number of debiased teachers in the ensemble.

5.2.2 Effect of the number of debiased teachers.

In this experiment, we study how an ensemble with different proportions of biased and debiased models impacts the WGA of the student model. We use the same training process as previously and consider an ensemble of five teachers with different proportions of debiased teachers (i.e., $\{0.2, 0.4, 0.6, 0.8, 1\}$); we report the average and WGA in Figure 3 for the Waterbirds and CelebA datasets. The results show that

Models	Debiased	KD	Waterbirds		CelebA	
			Average	WGA	Average	WGA
Deep Ensemble	✗	—	$84.4_{\pm 0.00}$	$59.3_{\pm 0.44}$	$95.6_{\pm 0.02}$	$37.7_{\pm 0.00}$
	✓	—	$93.5_{\pm 0.17}$	$90.0_{\pm 0.56}$	$92.4_{\pm 0.18}$	$88.8_{\pm 0.55}$
One-hot	✗	✗	$85.7_{\pm 1.80}$	$65.6_{\pm 5.75}$	$95.4_{\pm 0.07}$	$37.5_{\pm 2.27}$
	✓	✗	$92.2_{\pm 0.53}$	$90.9_{\pm 1.00}$	$92.5_{\pm 0.34}$	$88.2_{\pm 1.68}$
Random	✗	✓	$83.5_{\pm 0.68}$	$64.0_{\pm 2.88}$	$95.5_{\pm 0.00}$	$35.8_{\pm 1.17}$
	✓	✓	$92.8_{\pm 0.49}$	$90.3_{\pm 0.23}$	$\mathbf{92.6}_{\pm 0.26}$	$87.9_{\pm 0.64}$
AVER	✗	✓	$83.5_{\pm 0.75}$	$63.7_{\pm 2.72}$	$95.6_{\pm 0.02}$	$35.3_{\pm 1.95}$
	✓	✓	$\mathbf{92.8}_{\pm 0.64}$	$90.2_{\pm 0.23}$	$\mathbf{92.6}_{\pm 0.20}$	$88.3_{\pm 1.66}$
AE-KD	✗	✓	$84.2_{\pm 1.52}$	$61.0_{\pm 4.45}$	$95.6_{\pm 0.02}$	$36.9_{\pm 0.39}$
	✓	✓	$91.9_{\pm 1.89}$	$89.0_{\pm 3.59}$	$92.3_{\pm 0.04}$	$89.4_{\pm 1.11}$
AGRE-KD	✗	✓	$84.9_{\pm 1.40}$	$66.3_{\pm 4.76}$	$94.5_{\pm 3.22}$	$39.2_{\pm 0.78}$
	✓	✓	$91.4_{\pm 1.99}$	$\mathbf{91.1}_{\pm 2.56}$	$91.1_{\pm 0.09}$	$\mathbf{91.9}_{\pm 1.17}$

Table 2: **Results on self-distillation**. The student and the teacher models have the same network architecture (ResNet50). The student's worst group test accuracy increases when we distill to a network with higher capacity.



(a) Waterbirds

(b) CelebA

Figure 3: **Results on the proportion of debiased teachers in the ensemble**. We trained the student model using an ensemble of 5 teachers with different ratios of debiased teachers within the ensemble ({0.2, 0.4, 0.6, 0.8, 1}). AGRE-KD effectively upweights and favors the least biased teachers in the ensemble, while other ensemble methods rely more on biased teachers' output and decrease the WGA. AGRE-KD maintains significantly higher WGA, despite having only a single debiased model in the ensemble..

students' WGA increases as the proportion of debiased models in the ensemble increases. When the ensemble contains a single debiased model, the aggregation process of other ensemble KD methods relies more on the majority of biased teachers, leading the student model to capture spurious correlation. On the other hand, our method can implicitly identify and adaptively rely on the knowledge of debiased teachers while reducing reliance on biased teachers. This is reflected by the significantly higher WGA in settings where the ensemble contains a single debiased teacher. Additionally, AGRE-KD outperforms or matches the performance of the deep ensembling model, where the last layer of each model in the ensemble is directly retrained.

5.2.3 Ablation on teacher architecture heterogeneity

Experiments in Section 5.2.1 revealed that the capacity of the student model plays an important role in bias amplification of ensemble KD. Prior work has demonstrated that increasing the diversity of the ensembles

Models	Architecture	KD	Debiased Teachers		ERM Teachers	
			Avg	WGA	Avg	WGA
Teacher	ViT	✗	$96.1_{\pm 0.3}$	$90.1_{\pm 1.5}$	$92.7_{\pm 0.0}$	$73.8_{\pm 1.5}$
	ResNet50	✗	$94.3_{\pm 0.8}$	$89.5_{\pm 2.5}$	$88.7_{\pm 0.09}$	$68.0_{\pm 2.20}$
	ConvNeXtV2	✗	$92.4_{\pm 0.6}$	$85.7_{\pm 1.5}$	$86.4_{\pm 0.41}$	$62.0_{\pm 0.55}$
RANDOM	ResNet18	✓	$\mathbf{94.3}_{\pm 0.18}$	$80.6_{\pm 2.30}$	$88.9_{\pm 0.1}$	$57.7_{\pm 3.21}$
AVER		✓	$94.2_{\pm 0.21}$	$80.7_{\pm 1.10}$	$84.2_{\pm 0.4}$	$41.7_{\pm 2.51}$
AE-KD		✓	$93.5_{\pm 0.65}$	$84.9_{\pm 2.81}$	$89.1_{\pm 1.4}$	$57.9_{\pm 1.90}$
AGRE-KD		✓	$93.5_{\pm 0.63}$	$\mathbf{87.9}_{\pm 1.25}$	$\mathbf{89.5}_{\pm 0.7}$	$\mathbf{61.6}_{\pm 2.10}$

Table 3: Results with heterogeneous teacher ensemble on the Waterbirds dataset. We train teachers with different architectures (ViT, ResNet50, and ConvNeXtV2) and perform knowledge distillation to a ResNet18 student model.

results in better generalization performance. We enforced diversity by initializing the weights with independent random seeds (Allen-Zhu & Li, 2020). We now explore the heterogeneity of teacher models by building ensembles with diverse network architectures and analyzing the impact on the WGA accuracy of the student models. We train teacher models using three different network architectures, i.e., ViT (Dosovitskiy et al., 2020), ConvNeXtV2 (Woo et al., 2023), and ResNet50 (He et al., 2016). Using the same experimental setup as previously, we built an ensemble of nine models with an equal proportion of each model architecture, i.e., three models with each architecture, all trained using different random initialization. This ensures diversity between and within teacher architectures.

The results in Table 3 show, for the Waterbirds dataset, the average and WGA accuracy when the heterogeneous ensemble contains ERM-trained or debiased teacher models. We observe that even under a more heterogeneous teacher ensemble, student models trained with classical ensemble KD methods can match the teacher's average accuracy but incur a significant drop in WGA. Compared to student models trained with homogeneous ensembles of ResNet50 models (Table 1), we observe that using a more diverse teacher's backbone marginally improves performance across baselines, yet the drop in WGA remains important across ensemble KD methods. This highlights that the bias amplification in ensemble KD does not necessarily depend on ensemble diversity, further supporting our hypothesis of student capacity as a source of the disparity. On the other hand, the orchestration process in AGRE-KD shows better performance as its distillation process encourages better WGA. Despite the teachers' architecture being heterogeneous, we use a ResNet50 biased model to compute teachers' weights in AGRE-KD during the distillation process, demonstrating the robustness of the proposed approach to the choice of teacher model architecture. The results on the CelebA dataset show the same trends, which we provide in Table 7 in the Appendix.

6 Limitation and conclusion

In this paper, we studied bias in ensemble knowledge distillation (KD) and demonstrated that, unlike deep ensemble models that reduce bias, traditional ensemble KD methods can amplify it. We proposed AGRE-KD, an adaptive gradient-based weighting method that improves group robustness in ensemble KD by guiding the student model to learn core features and boosting worst-group accuracy. Our experiments across several benchmarks demonstrated the effectiveness of our approach in distilling knowledge with reduced spurious correlations. While our results highlight AGRE-KD's advantages over existing methods, several limitations and open questions remain. Specifically, the study focused on unsupervised KD using the teachers' logits alone, without access to group or class labels. Although our primary interest was in examining the ability of an ensemble of teachers to distill more robust knowledge, exploring scenarios where class labels are available could enhance the performance of the student further. Moreover, we observed that the WGA improvement in the proposed method is less pronounced when all teachers in the ensemble are biased. This indicates an opportunity to leverage class labels in this context to further enhance group robustness, e.g., by considering the teachers' misclassifications. Finally, additional evaluation using more complex datasets, such as health datasets, is necessary to validate the approach across a wider range of applications.

References

Sumyeong Ahn, Seongyoon Kim, and Se-Young Yun. Mitigating dataset bias by using per-sample gradient. *arXiv preprint arXiv:2205.15704*, 2022.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Pedro RAS Bassi, Andrea Cavalli, and Sergio Decherchi. Explanation is all you need in distillation: Mitigating bias and shortcut learning. *arXiv preprint arXiv:2407.09788*, 2024.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11933–11942, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL `https://api.semanticscholar.org/CorpusID:52967399`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *advances in neural information processing systems*, 33:12345–12355, 2020.

William Falcon and The PyTorch Lightning team. Pytorch lightning, 2019. URL `https://www.pytorchlightning.ai`. Available at: `https://github.com/Lightning-AI/lightning`.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pp. 3697–3701, 2017.

Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

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.

Patrik Joslin Kenfack, Adil Mehmood Khan, SM Ahsan Kazmi, Rasheed Hussain, Alma Oracevic, and Asad Masood Khattak. Impact of model ensemble on the fairness of classifiers in machine learning. In *2021 International conference on applied artificial intelligence (ICAPAI)*, pp. 1–6. IEEE, 2021.

Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. A survey on fairness without demographics. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL `https://openreview.net/forum?id=3HE4vPNIfX`.

Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. Fairness under demographic scarce regime. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL `https://openreview.net/forum?id=TB18G0w6Ld`.

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.

Wei-Yin Ko, Daniel D'souza, Karina Nguyen, Randall Balestriero, and Sara Hooker. Fair-ensemble: When fairness naturally emerges from deep ensembling. *arXiv preprint arXiv:2303.00586*, 2023.

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 *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. Adaptive knowledge distillation based on entropy. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7409–7413. IEEE, 2020.

Tyler LaBonte. Milkshake: Quick and extendable experimentation with classification models. `http://github.com/tmlabonte/milkshake`, 2023.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiwoon Lee and Jaeho Lee. Debiased distillation by transplanting the last layer. *arXiv preprint arXiv:2302.11187*, 2023.

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.

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.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *International Conference on Learning Representations (ICLR)*, 2019.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *arXiv preprint arXiv:2106.10494*, 2021.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*, 2023.

Aida Mohammadshahi and Yani Ioannou. What is left after distillation? how knowledge transfer impacts fairness and bias. *arXiv preprint arXiv:2410.08407*, 2024.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS Workshop Autodiff*, 2017.

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. *Advances in neural information processing systems*, 32, 2019.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, pp. 28448–28467. PMLR, 2023.

Ahmed Radwan, Layan Zaafarani, Jetana Abudawood, Faisal AlZahrani, and Fares Fourati. Addressing bias through ensemble learning and regularized fine-tuning. *arXiv preprint arXiv:2402.00910*, 2024.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

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.

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. *arXiv preprint arXiv:1911.08731*, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. URL `https://api.semanticscholar.org/CorpusID:203626972`.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In *International Conference on Machine Learning*, pp. 34330–34343. PMLR, 2023.

Rishabh Tiwari, Durga Sivasubramanian, Anmol Mekala, Ganesh Ramakrishnan, and Pradeep Shenoy. Using early readouts to mediate featural bias in distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2638–2647, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

Serena Wang, Harikrishna Narasimhan, Yichen Zhou, Sara Hooker, Michal Lukasik, and Aditya Krishna Menon. Robust distillation for worst-class performance: on the interplay between teacher and student objectives. In *Uncertainty in Artificial Intelligence*, pp. 2237–2247. PMLR, 2023.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16133–16142, 2023.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1285–1294, 2017.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4498–4502. IEEE, 2022.

Wenjie Zhou, Zhenxin Ding, Xiaodong Zhang, Haibo Shi, Junfeng Wang, and Dawei Yin. Govern: Gradient orientation vote ensemble for multi-teacher reinforced distillation. *arXiv preprint arXiv:2405.03764*, 2024.

A AGRE-KD: Algorithm and method overview

In Algorithm 1, we provide a high-level description of the AGRE-KD methodology for training and validation of group-robust distilled models. We also provide an overview of our method in Figure 2.

Algorithm 1 Adaptive Group Robust Ensemble Knowledge Distillation (AGRE-KD)

```
Input: Ensemble of pretrained teachers $T = \{T_1, T_2, \dots, T_M\}$, biased model $T_b$, student model $S$ with
  parameters $\theta$, dataset $\mathcal{D}$, distillation coefficient $\alpha$, temperature parameter $\tau$
for each minibatch $\{(x_i, y_i)\}_{i=1}^{B}$ from $\mathcal{D}$ do
    for each teacher $T_t \in T$ do
        Compute knowledge distillation loss $\ell_i^t(\theta)$ for sample $i$ between $S$ and $T_t$    ▷ Equation 1.
        Compute biased model distillation loss $\ell_i^b(\theta)$ for sample $i$ between $S$ and $T_b$    ▷ Equation 1.
        Compute gradient alignment $G_i^t = \langle \nabla\ell_i^t(\theta), \nabla\ell_i^b(\theta) \rangle$ ▷ Dot product of normalized gradient vectors.
  Table 5 shows the importance of ignoring the magnitude of the gradients.
        Compute the adaptive sample weight for teacher $t$ on sample $i$: $W_t(x_i) = 1 - G_i^t$;
    end for
    Compute weighted knowledge distillation loss:
        $\mathcal{L}_{\text{wKD}} = \frac{\sum_t W_t(x_i) \cdot \ell_i^t(\theta)}{\sum_t W_t(x_i)}$
    Compute classification loss (if labeled data available): $\mathcal{L}_{\text{cls}}$
    Compute final loss:
        $\mathcal{L} = \alpha\mathcal{L}_{\text{wKD}} + (1-\alpha)\mathcal{L}_{\text{cls}}$
    Update student model parameters $\theta$ using gradient descent on $\mathcal{L}$
end for
```

B Hyperparameters

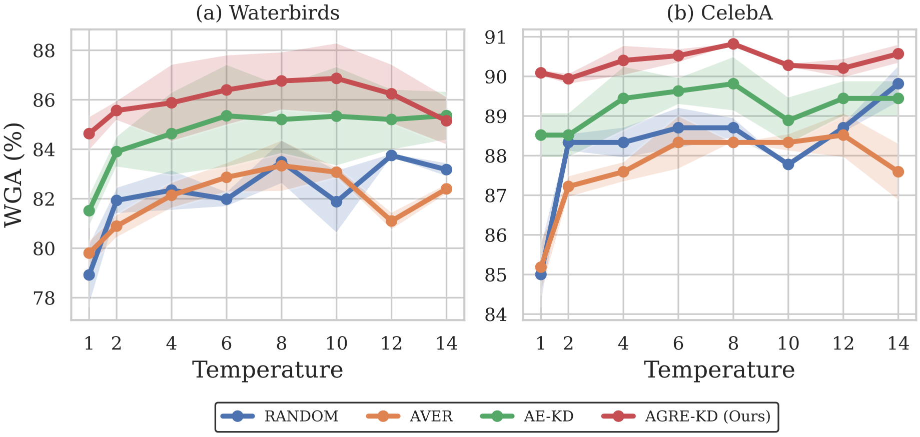
We train all models using standard hyperparameters from previous work (LaBonte et al., 2024; Kirichenko et al., 2022; Izmailov et al., 2022) and keep their value fixed across experiments. For the vision tasks, we used an initial learning rate of 1×10^{-3} with a cosine learning rate scheduler; we used a batch size of 32 and 100 for the Waterbirds and the CelebA datasets, respectively. For the CivilComments dataset, we use an initial learning rate of 1×10^{-5} with a linear learning rate scheduler, a batch size 16, and train for ten epochs. We keep all hyperparameters fixed to train the teacher and student models. For the optimizer, we used AdamW (Loshchilov et al., 2019) and SGD for the language and vision datasets, respectively, with a weight decay of 1×10^{-4}. Our implementation uses PyTorch (Paszke et al., 2017; 2019), Torch Lightning (Falcon & team, 2019), and Milkshake (LaBonte, 2023).

C Ablation on temperature parameter

Our experiments using the DFR teacher models revealed that reweighting the feature by retraining the last classification layer provides a "dark knowledge" (soft target) to the student model that is more robust to spurious correlation. Since DFR teachers share the same feature representation as their ERM-trained counterparts but differ only in their final classification layer, the enhanced robustness of the student when using DFR teachers indicates that the probability distribution over all classes predicted by the teacher significantly influences the student's dependence on spurious correlations. Since the temperature hyperparameter (Equation 1) softens the probability distribution of the teacher to better expose dark knowledge. We perform an ablation study by varying the temperature parameter ($\tau \in \{1, 2, 4, ..., 10\}$) and analyze its influence on the spurious correlation captured by the student model.

Figure 4 shows the WGA averaged across three independent random seeds in WaterBirds and CelebA datasets. We considered ensembles of five ResNet50 DFR teachers and a ResNet18 student. The results show that

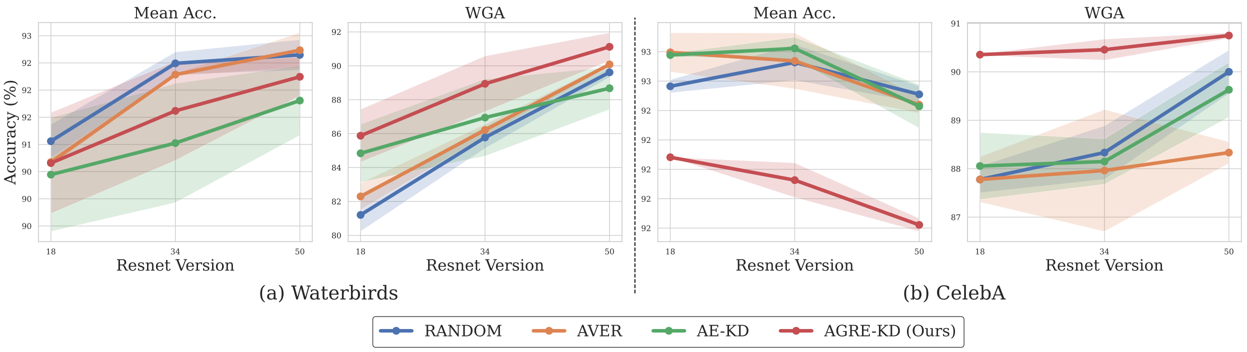
WGA increases with temperature and then remains steady or decreases after higher values of temperature ($\tau \geq 10$ and $\tau \geq 8$ for the Waterbirds and CelebA datasets, respectively). The smallest WGA is achieved when $\tau = 1$, which means the training uses a regular cross-entropy loss; it increases when we soften the probability distribution ($\tau > 1$) and eventually decreases when the distribution gets almost uniform ($\tau \gg 1$). These findings suggest that the dark knowledge provided by the softened probability distribution of the teacher models reduces students' reliance on spurious correlation during training. These results align with recent work by Mohammadshahi & Ioannou (2024) studying the benefit of increased temperature for fairness in knowledge distillation.


(a) Waterbirds

(b) CelebA

Figure 4: Effect of the temperature parameter on the worst-group accuracy.

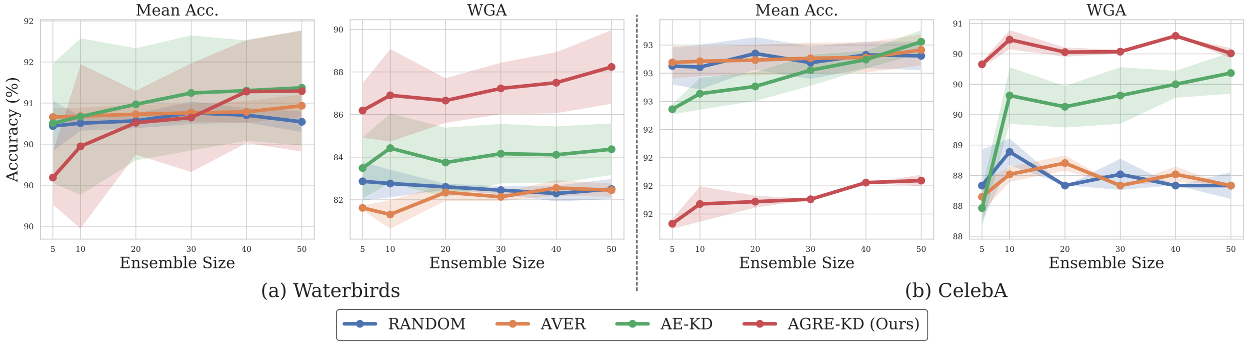
D Supplemental results



(a) Waterbirds

(b) CelebA

Figure 5: **Results on student model capacity**. We perform experiments on WaterBirds and CelebA using ResNet50 teachers and varying the capacity of the student network (ResNet18, ResNet34, and ResNet50). We plot the average and the worst-case accuracy of different ensemble distillation methods across three random seeds. The WGA tends to increase as the student model has more capacity for learning the core features. Most baselines match or outperform the teachers when the student model has the same capacity as the teachers (i.e., ResNet50), and our method remains superior in terms of WGA. These results suggest that the reduced capacity of the student model is a source of the disparity observed.



(a) Waterbirds

(b) CelebA

Figure 6: **Effect of teacher size**. We train the resnet18 student model with different ensemble sizes of resnet50 teachers (5, 10, 20, . . . , 50). Increasing ensemble size exerts a positive effect on both the average and the worst-group performance. However, the Random KD method tends to get worse as we increase the number of teachers.

Biased model	Waterbirds		CelebA	
	Average	WGA	Average	WGA
Resnet50	$90.6_{\pm 0.49}$	$86.7_{\pm 2.82}$	$91.8_{\pm 0.20}$	$90.5_{\pm 0.24}$
Resnet34	$90.0_{\pm 1.63}$	$85.2_{\pm 3.48}$	$91.8_{\pm 0.14}$	$90.5_{\pm 0.24}$
Resnet18	$90.2_{\pm 1.10}$	$86.8_{\pm 2.88}$	$91.5_{\pm 0.09}$	$89.8_{\pm 0.16}$

Table 4: **Sensitivity of AGRE-KD to the biased model architecture**. In the main paper, we used a biased model with the same architecture as the teacher models (i.e., resnet50). We experiment with different network backbones for the biased models in AGRE-KD. The results below do not show significant differences across biased model choices, demonstrating the robustness of the proposed method to the choice of biased model. These results suggest that the gradient direction of any biased pretrained model can provide sufficient guidance for debiased distillation.

Model	Waterbirds		CelebA	
	Average	WGA	Average	WGA
AGRE-KD w/ grad norm	$90.6_{\pm 0.49}$	$\mathbf{86.8}_{\pm 1.86}$	$91.7_{\pm 0.20}$	$\mathbf{90.9}_{\pm 0.71}$
AGRE-KD w/o grad norm	$91.0_{\pm 1.32}$	$81.5_{\pm 4.07}$	$93.1_{\pm 0.21}$	$87.9_{\pm 0.52}$

Table 5: **Training AGRE-KD using gradient direction with (w/) and without (w/o) gradient normalization**. On the Waterbirds and CelebA datasets, we study how using the gradient magnitude in the weighting scheme impact the results. We trained our AGRE-KD method without normalizing the gradient vectors in the dots product, i.e., accounting for the gradient magnitude of the losses. The results below show that accounting for the gradient magnitude in the weighting scheme reduces the WGA performance. This shows the importance of using normalized dot products to compare directional changes in gradients, making the computed weights independent of the gradient scales themselves, which are generally very noisy.

Datasets	Sub Groups	#Samples	ERM	Teacher	Random	AVER	AEKD	AGRE-KD
Waterbirds	(landbirds,land)	3498	99.3$_{\pm 0.13}$	95.5$_{\pm 0.97}$	96.7$_{\pm 0.57}$	96.8$_{\pm 0.62}$	96.1$_{\pm 1.29}$	94.3$_{\pm 1.46}$
	(**landbirds,water**)	184	74.0$_{\pm 2.54}$	94.3$_{\pm 0.96}$	86.2$_{\pm 2.20}$	87.4$_{\pm 0.91}$	86.4$_{\pm 3.88}$	87.5$_{\pm 3.14}$
	(waterbirds,land)	56	54.1$_{\pm 1.80}$	92.1$_{\pm 1.39}$	82.3$_{\pm 1.65}$	82.1$_{\pm 1.06}$	84.6$_{\pm 3.37}$	86.7$_{\pm 2.82}$
	(waterbirds,water)	1057	93.4$_{\pm 0.72}$	90.9$_{\pm 0.81}$	91.5$_{\pm 1.08}$	90.9$_{\pm 0.67}$	91.4$_{\pm 1.97}$	90.6$_{\pm 1.85}$
CelebA	(nonblond,female)	71629	96.0$_{\pm 0.39}$	91.1$_{\pm 0.37}$	91.8$_{\pm 0.52}$	91.8$_{\pm 0.52}$	91.3$_{\pm 0.36}$	90.4$_{\pm 0.77}$
	(nonblond,male)	66874	99.5$_{\pm 0.06}$	92.9$_{\pm 0.20}$	94.3$_{\pm 0.37}$	94.2$_{\pm 0.20}$	93.7$_{\pm 0.40}$	92.4$_{\pm 0.53}$
	(blond,female)	22880	85.2$_{\pm 1.21}$	94.4$_{\pm 0.30}$	93.8$_{\pm 0.35}$	93.9$_{\pm 1.13}$	94.6$_{\pm 0.34}$	95.0$_{\pm 0.83}$
	(blond,male)	1387	32.3$_{\pm 2.17}$	90.1$_{\pm 0.86}$	88.3$_{\pm 0.78}$	87.5$_{\pm 0.52}$	89.4$_{\pm 1.63}$	91.6$_{\pm 0.78}$

Table 6: **Group-wise accuracy comparison**. We report the group-wise accuracy of different ensemble KD methods on the Waterbirds and CelebA datasets. As in previous experiments, we average the performances across three different random seeds and considered ensembles of five teachers.

Models	Architecture	KD	Debiased Teacher		ERM Teacher	
			Avg	WGA	Avg	WGA
Teacher	ViT	✗	93.9$_{\pm 0.48}$	85.0$_{\pm 1.46}$	95.8$_{\pm 0.00}$	38.8$_{\pm 0.78}$
	ResNet50	✗	93.2$_{\pm 0.92}$	90.1$_{\pm 0.86}$	95.8$_{\pm 0.17}$	42.2$_{\pm 3.14}$
	ConvNeXtV2	✗	92.9$_{\pm 0.07}$	86.9$_{\pm 0.39}$	95.9$_{\pm 0.05}$	44.4$_{\pm 0.78}$
RANDOM	ResNet18	✓	94.1$_{\pm 0.48}$	79.4$_{\pm 3.14}$	95.9$_{\pm 0.07}$	36.3$_{\pm 0.39}$
AVER		✓	94.1$_{\pm 0.38}$	80.0$_{\pm 3.10}$	95.9$_{\pm 0.09}$	36.1$_{\pm 0.00}$
AE-KD		✓	93.9$_{\pm 0.19}$	82.2$_{\pm 1.57}$	95.9$_{\pm 0.10}$	40.2$_{\pm 1.17}$
AGRE-KD		✓	90.4$_{\pm 0.69}$	87.6$_{\pm 0.94}$	95.9$_{\pm 0.03}$	41.2$_{\pm 1.57}$

Table 7: Results with heterogeneous teacher ensemble on the CelebA dataset. We train teachers with different architectures (ViT, ResNet50, and ConvNeXtV2) and perform knowledge distillation to a ResNet18 student model.