Towards Fair Knowledge Distillation using Student Feedback

Abhinav Java [*][1] **Surgan Jandial** [*][1] **Chirag Agarwal** [2]

Abstract

With the advent of large-scale models and their success in diverse fields, Knowledge Distillation (KD) techniques are increasingly used to deploy them to edge devices with limited memory and computation constraints. However, most distillation works focus on improving the prediction performance of the student model with little to no work in studying the effect of distillation on key fairness properties, ensuring trustworthy distillation. In this work, we propose a fairness-driven distillation framework, BIRD (BIas-awaRe Distillation), which introduces a FAIRDISTILL operator to collect feedback from the student through a meta-learning-based approach and selectively distill teacher knowledge. We demonstrate that BIRD can be augmented with different KD methods to increase the performance of foundation models and convolutional neural networks. Extensive experiments across three fairness datasets show the efficacy of our framework over existing state-of-the-art KD methods, opening up new directions to develop trustworthy distillation techniques.

1. Introduction

Recent years have witnessed an alarming trend toward developing large-scale vision and language foundation models (Radford et al., 2021; Singh et al., 2022) trained on large datasets from unvetted data sources, leading to several deployment and societal issues (Agarwal; Birhane et al., 2021; Mehrabi et al., 2021; Naik & Nushi, 2023; Seth et al., 2023). To address these constraints, model compression techniques have been recently used to preserve their predictive performance while significantly reducing their parameter size (Hsieh et al., 2023; Wang et al., 2022). Amongst these techniques, Knowledge Distillation (KD) has been shown to work for a wide range of models (Wang et al., 2022; Sanh et al., 2019), where it distills the internal feature representations and/or the outputs logits of a larger teacher model into a smaller student model (Bucilua et al., 2006; Hinton et al., 2015; Romero et al., 2015; Zagoruyko & Komodakis, 2017; Yim et al., 2017; Jandial et al., 2023). As the distillation of large models becomes increasingly prevalent in real-world applications, it becomes crucial to ensure that the resulting student models and their output representations are safe and reliable. In particular, they should not learn discriminatory features, and their bias should not be exacerbated due to distillation from the teacher model.

Previous fairness works have argued that bias in machine learning is primarily due to the dataset and/or the training process (Agarwal et al., 2021; Hooker, 2021; Yucer et al., 2022). However, they cannot be directly extended to KD frameworks as there is an inherent trade-off between *"what"* and *"how much"* a student distills knowledge from a teacher. While a student model can learn useful teacher predictive properties, it is also prone to inheriting the bias of an unfair teacher. These contrasting aspects of the distillation framework make the problem of fair KD non-trivial. While several existing works focus on either KD or fairness in ML, there is little to no research on addressing them simultaneously. Despite independent efforts in these two fields, there remain open questions about whether a student inherits biases from a teacher and *"what"* teacher features distill more bias during KD. To this end, Jung et al. (2021) proposes a method to perform fair KD which enforces fairness constraints by matching the student's group conditioned features with the teacher's group indistinguishable features (Gretton et al., 2012), where a group is a protected attribute like gender, race, etc. However, there is no guarantee that the group averaging will wash away the bias from the teacher features and scale to large-scale foundation models. In contrast to existing works, we propose to transform the biased knowledge from the teacher as well as debias the student model with fairness objectives. More specifically, our framework learns to exclude the biased features from the teacher by incorporating *student feedback* via meta-learning.

Present Work. In this work, we address fairness in KD and ensure that the student learns fair and accurate representations. In doing so, we first identify the key connection between the bias induced by the teacher and the bias inherited

[*]Equal contribution [1]Media and Data Science Research, Adobe [2]Harvard University. Correspondence to: Chirag Agarwal <chiragagarwall12@gmail.com>.

Published at the ICML 2023 Workshop on Efficient Systems for Foundation Models. Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

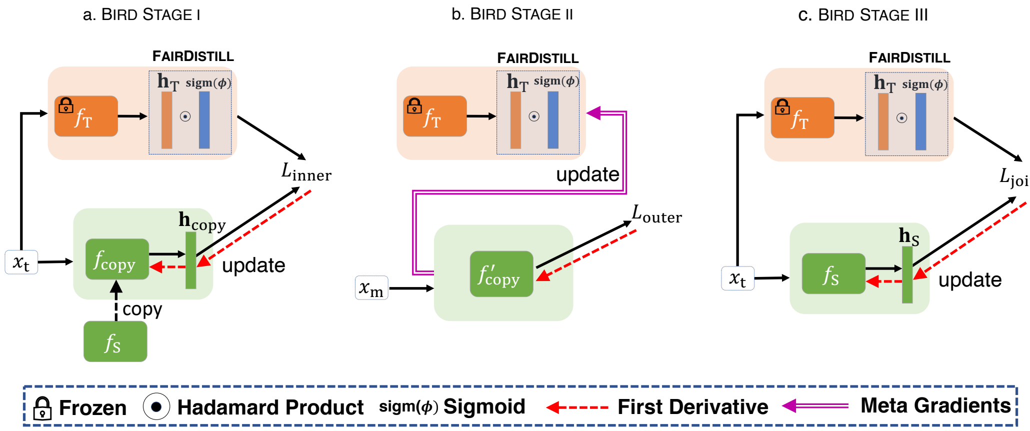

Figure 1: Overview of our BIRD framework – BIRD learns bias-aware representations from the teacher f_T by training the FAIRDISTILL operator using a meta-learning framework: **a)** In Stage I, BIRD updates a copy of the student model with $\mathcal{L}_\mathrm{in}$, **b)** in Stage II, the updated model f'_c is used to train ϕ with bias-feedback information in the form of meta-gradients from $\mathcal{L}_\mathrm{o}$, and **c)** in Stage III, the student model f_S is distills unbiased representations using FAIRDISTILL (from Stage II).

by the student (refer to Sec. 4.1 Q1). We show that there is a trade-off between the predictive and biased knowledge a student distills from a teacher. While the teacher does improve the predictive performance of the student, we find that existing KD techniques lead a student model to mimic the bias in teacher's predictions. We leverage this connection to propose a novel framework, BIRD (BIas-awaRe Distillation), which can be used with any existing KD methods and architectures to learn fair student representations. We introduce a fair-distillation operator that selects and filters a subset of uncorrelated features from the teacher during distillation. The learnable operator is updated using meta-gradients from the student objective functions. To the best of our knowledge, this work is the first to tackle the problem of fairness in KD using student feedback in a meta-learning pipeline.

Our contributions. We present our contributions as follows: i) we propose BIRD, a fair knowledge distillation framework that achieves more effective debiasing for KD compared to existing techniques. ii) BIRD is model-agnostic and can be integrated with diverse foundational models and CNNs across knowledge distillation methods. iii) BIRD introduces a simple, flexible, and computationally inexpensive fair-distillation operator trained using meta-learning for selectively distilling fair and accurate teacher features. iv) We conduct experiments on multiple fairness datasets and demonstrate the effectiveness of BIRD through extensive empirical analysis. Results show that BIRD improves the fairness of knowledge distillation approaches by 32.53% (on average) without sacrificing predictive performance.

2. Preliminaries

Notation. Let $\mathbf{D}$=$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ be the labelled dataset, where each input image sample $\mathbf{x}_i \in \mathbb{R}^{3\times\mathrm{h}\times\mathrm{w}}$ has height h and width w, and each label $y_i \in \{1, 2, \ldots, C\}$ represents one of the C classes in $\mathbf{D}$. In addition to the ground-truth prediction label, each sample comprises a protected attribute label y_p which may be used unfairly against the subject in the image. Following previous meta-learning works (Finn et al., 2017), we split the dataset $\mathbf{D}$ into three mutually exclusive sets $\mathbf{D}^\mathrm{train}$, $\mathbf{D}^\mathrm{test}$, and $\mathbf{D}^\mathrm{meta}$.

Knowledge Distillation. Let f_T and f_S denote the teacher and student model parameterized by θ_T and θ_S. We denote the output representations of the f_T and f_S as $\mathbf{h}_\mathrm{T}$ and $\mathbf{h}_\mathrm{S}$ respectively. Methods such as Zagoruyko & Komodakis (2017) and Romero et al. (2015) perform Feature Knowledge Distillation (FKD) by optimizing $\mathcal{D}(\mathbf{h}_\mathrm{S}, \mathbf{h}_\mathrm{T})$, where $\mathcal{D}$ is any distance metric (*e.g.,* Euclidean). For the remainder of this work, we define FKD as:

$$\mathcal{L}_\mathrm{FKD} = \alpha\mathcal{D}\left(\mathbf{h}_\mathrm{S}, \mathbf{h}_\mathrm{T}\right) + (1-\alpha)\mathcal{L}_\mathrm{CE}(\hat{\mathbf{y}}, y) \tag{1}$$

where α is the weight coefficient, $\mathcal{L}_\mathrm{CE}$ denotes the cross entropy loss function, and $\hat{\mathbf{y}}$ represents the softmax outputs of f_S and y are the ground truth labels respectively. Formally, the definition of KD (Eqn. 1) suggests that the student models are solely optimized for their predicted performance. Since teacher models may be biased, the KD frameworks may distill spurious correlations in the student, motivating the need for a bias-aware distillation framework.

3. Our method: BIRD

Our framework demonstrates that we can obtain a fair student model by i) eliminating the biased features in the teacher representation and ii) using fairness objectives for optimizing the student during distillation. BIRD achieves this by introducing a fairness-aware distillation operator (Sec. 3.1), as well as the student copy update (Sec. 3.2) and meta-update (Sec. 3.3) objectives in the distillation pipeline.

Problem Formulation (Bias-Aware Distillation). *Given a dataset $\mathbf{D}^{train}$ and a biased teacher model f_T optimized for predictive performance on $\mathbf{D}^{train}$, we aim to learn a student model f_S whose representations do not reflect any undesirable discriminatory biases (i.e., they are fair) and achieve high predictive performance (i.e., they are accurate).*

3.1. Fairness-Aware Distillation Operator

We propose a fairness-aware distillation operator, denoted as FAIRDISTILL, which aims to identify and distill a subset of features from the teacher, or in other words *selectively distill fair teacher representations*. Next, we present the formal definition of FAIRDISTILL, which is generic in its formulation and can be augmented with any existing KD method.

FAIRDISTILL operator. We define a computationally inexpensive operator FAIRDISTILL : $\mathbb{R}^d \rightarrow \mathbb{R}^d$, consisting of a d-dimensional learnable weight vector ϕ. For a given teacher representation $\mathbf{h}_\mathrm{T}$, we define $\textsc{FairDistill}(\mathbf{h}_\mathrm{T}) = \mathrm{sigmoid}(\phi) \odot \mathbf{h}_\mathrm{T}$, where $\phi \in [0,1]^\mathrm{d}$ are the learnable parameters, $\odot$ is the Hadamard product, and $\mathrm{sigmoid}$ is the non-linear activation function. We apply sigmoid to ϕ before performing element-wise multiplication to re-weight $\mathbf{h}_\mathrm{T}$ based on their correlation with the protected attributes.

3.2. Student copy update

At the beginning of the inner optimization loop, we create a copy of the original student model f_c. Next, using $(\mathbf{x}_\mathrm{t}, y_\mathrm{t}) \sim \mathbf{D}^\mathrm{train}$, we obtain the penultimate layer representations $\mathbf{h}_\mathrm{T}, \mathbf{h}_\mathrm{c}$ from $f_\mathrm{T}, f_\mathrm{c}$, respectively. We then leverage the FAIRDISTILL at the current step to transform $\mathbf{h}_\mathrm{T}$ and update f_c with the following objective:

$$\mathcal{L}_\mathrm{in} = \alpha\mathcal{D}\left(\mathbf{h}_\mathrm{c}, \textsc{FairDistill}(\mathbf{h}_\mathrm{T})\right) + (1-\alpha)\mathcal{L}_\mathrm{CE}(\hat{\mathbf{y}}_\mathrm{t}, y_\mathrm{t}) \tag{2}$$

where $\mathcal{D}$ can be any distance-based metric and $\hat{\mathbf{y}}_\mathrm{t}$ is the output predicted by $f_\mathrm{c}(\mathbf{x}_\mathrm{t})$. We update FAIRDISTILL in the meta-step based on the resulting fairness properties of f'_c. Intuitively, we update FAIRDISTILL such that f'_c is fair, which in turn teaches FAIRDISTILL to distill fairly.

3.3. Meta Update Phase

We first sample data from the meta-subset $(\mathbf{x}_\mathrm{m}, y_\mathrm{m}) \sim \mathbf{D}^\mathrm{meta}$ and pass it through f'_c (Sec. 3.2). Also, θ_c is a function of ϕ which implies that $\nabla\theta'_\mathrm{c}$ is a function of the *gradients of* ϕ. Consequently, we use θ'_c to perform meta-updates on ϕ using a bias-aware objective function:

$$\mathcal{L}_\mathrm{o} = \sum_{i=1}^{C} \max \sum_{j=1}^{M} \mathrm{abs}\Big(\mathcal{L}_\mathrm{CE}(\hat{\mathbf{y}}_i|y_p{=}j, y_i|y_p{=}j) - \mathcal{L}_\mathrm{CE}(\hat{\mathbf{y}}_i, y_i)\Big) \tag{3}$$

where M is the number of unique values in the given protected attribute p, $\hat{\mathbf{y}}_i|y_p = j$ denotes the output of the network $f'_\mathrm{c}(\mathbf{x}_\mathrm{m})$ such that the unprotected class label is i and the label of the protected attribute is j. Intuitively, $\nabla_\phi\mathcal{L}_\mathrm{o}$ implies that ϕ is updated such that f'_c exhibits equal predictive performance across all protected demographic groups for every task category (i.e., it is fair), following the fairness definition proposed in Hardt et al. (2016). In Fig. 1 (Stage II), we show the information flow using the meta-gradients obtained from $\mathcal{L}_\mathrm{o}$. In particular, the gradients of $\mathcal{L}_\mathrm{o}$ with respect to ϕ are backpropagated via θ'_c and ϕ is subsequently updated by computing the *gradients of gradients* or *meta-gradients*.

Algorithm 1 BIRD: BIas-awaRe Distillation.

```
BIRD ($\theta_\mathrm{S}, \theta_\mathrm{T}, \phi$)                                    ▷ Input Parameters
Hyperparameters: $\mu_1, \mu_2, \mu_3$                                  ▷ Learning Rates
Dataset: $\mathbf{D}^\mathrm{train}, \mathbf{D}^\mathrm{meta}$                                    ▷ Data Splits
while not done do
    $\theta_\mathrm{c} \leftarrow \theta_\mathrm{S}$                                    ▷ Save current student state
    $\theta'_\mathrm{c} \leftarrow \theta_\mathrm{c} - \mu_1\nabla_{\theta_\mathrm{c}}\mathcal{L}_\mathrm{in}(f_\mathrm{c})$ ▷ Update with $(\mathbf{x}_\mathrm{t}, y_\mathrm{t}) \sim \mathbf{D}^\mathrm{train}$ using Eqn. 2
    $\phi \leftarrow \phi - \mu_2\nabla_\phi\mathcal{L}_\mathrm{o}(f'_\mathrm{c})$ ▷ Update with $(\mathbf{x}_\mathrm{m}, y_\mathrm{m}) \sim \mathbf{D}^\mathrm{meta}$ using Eqn. 3
    $\theta_\mathrm{S} \leftarrow \theta_\mathrm{S} - \mu_3\nabla_{\theta_\mathrm{S}}\mathcal{L}_\mathrm{j}(f_\mathrm{S})$ ▷ Update with $(\mathbf{x}_\mathrm{t}, y_\mathrm{t}) \sim \mathbf{D}^\mathrm{train}$ using Eqn. 4
end while
```

3.4. Overall Optimization

To learn $\mathbf{h}_\mathrm{S}$ that is invariant to y_p, we train BIRD using the following objectives: i) gather feedback from f_S in the form of meta-gradients to learn an optimal FAIRDISTILL (see Sec. 3.3) and ii) apply the learned FAIRDISTILL on f_T to selectively perform KD. In addition, we use a model-agnostic regularization on f_S that further penalizes student bias. Finally, the joint objective which updates the original student model f_S using the *updated* FAIRDISTILL is given as:

$$\mathcal{L}_\mathrm{j} = \alpha\mathcal{D}(\mathbf{h}_\mathrm{S}, \textsc{FairDistill}(\mathbf{h}_\mathrm{T})) + (1-\alpha)\mathcal{L}_\mathrm{CE}(\hat{\mathbf{y}}_\mathrm{t}, y_\mathrm{t}) + \lambda\mathcal{L}_\mathrm{reg} \tag{4}$$

where $\mathcal{L}_\mathrm{reg}$ is the regularization on f_S that penalizes student bias, λ is a regularization coefficient, and $\hat{\mathbf{y}}_t$ is the softmax output of $f_\mathrm{S}(\mathbf{x}_\mathrm{t})$. Without any loss of generality, we use Eqn. 3 as the regularization term in our BIRD framework.

4. Experiments

We present the experimental results for BIRD and address the following questions: Q1) Does KD worsen/improve student's unfairness? Q2) Can we selectively distill from the teacher to ensure bias-free distillation? Q3) How do meta gradients from student models improve debiasing? Q4) Can BIRD be augmented with existing KD baselines? Refer to Appendix for details on experimental setup.

Table 1: Results of KD on three FMs using CelebA dataset. Shown is the avg. performance across five independent runs. Arrows (↑, ↓) indicate the direction of better performance. BIRD retains the predictive power (AUROC) of the baseline model while improving their fairness (shaded area).

Model	Method	AUROC (↑)	$\Delta_{\text{mean-DEO}}(\downarrow)$	$\Delta_{\text{max-DEO}}(\downarrow)$
Flava	Baseline	84.43±0.12	27.48±0.64	29.37±1.53
	BKD	84.42±0.11	27.39±0.58	29.36±1.41
	FitNet	84.47±0.10	26.59±0.62	28.56±0.68
	AD	84.35±0.05	10.54±0.80	12.93±0.79
	MFD	84.45±0.11	26.64±0.62	28.63±0.68
	BIRD	85.48±0.02	**2.53**±0.17	**4.12**±0.59
CLIP-ViT/32	Baseline	87.01±0.26	23.38±1.72	24.91±1.15
	BKD	87.07±0.26	23.26±1.67	24.62±1.14
	FitNet	87.17±0.13	22.84 ±1.03	24.17±1.22
	AD	88.20 ±0.17	17.02 ±1.03	17.82 ±0.97
	MFD	87.22±0.11	21.99±0.70	23.70±1.58
	BIRD	88.55±0.03	**3.44**±0.92	**5.19**±1.06
CLIP-R50	Baseline	87.72±0.06	21.11±0.30	21.97 ±0.41
	BKD	87.72±0.06	21.10±0.40	22.07 ±0.41
	FitNet	87.54±0.14	22.01 ±1.05	23.30 ±1.15
	AD	88.51 ±0.02	5.33 ±0.19	7.93 ±0.22
	MFD	87.49±0.12	22.56 ±0.56	23.52±0.33
	BIRD	87.93±0.01	**2.65**±0.29	**4.49**±0.48

4.1. Results

Q1) Student inherits the fairness properties of the teacher. We use BKD (Hinton et al., 2015) with CLIP-ViT-B/32 and FLAVA models as teachers and the ResNet-{18,34} as students. Across different teacher-student combinations, results in Fig. 2 show that the student, which originally had better fairness performance, becomes unfair (higher metric values) after inheriting the teacher's biased features. On average, we find an increase of 38.54% in $\Delta_{\text{mean-DEO}}$ and 37.26% in $\Delta_{\text{max-DEO}}$. These results support our hypothesis that KD introduces bias in the student model. See the Appendix (Table 6) for similar insights on additional teacher-student architectures.

Q2) BIRD improves the fairness of knowledge distillation. Across state-of-the-art foundational models (see Table 1), we show that BIRD consistently learns fairer student representations while preserving the predictive performance of the original model. On average, BIRD improves the fairness of the underlying model by **87.60%** (in $\Delta_{\text{mean-DEO}}$) and **81.67%** (in $\Delta_{\text{max-DEO}}$), respectively. Further, we consider a CLIP-ViT-B/32 → ResNet-18 teacher-student distillation setting, where we find an average improvement of about 64% in the fairness performance of the ResNet-18 student model compared to its baselines (Fig. 3). Please refer to Table 2-4 for results on CNNs, CIFAR-10S and UTKFace.

Q3) Meta-updates improve fairness. We conduct an ablation on the importance of the meta-step in our BIRD framework. In doing so, we do not update the parameters of the FAIRDISTILL operator using the gradients of gradients obtained from the student model. Results show that the meta-learning component is necessary to learn fair representations (Table 7). In particular, we observe an 18.27% improvement in the fairness of BIRD, as compared to BIRD w/o Meta, providing empirical evidence that the meta-gradients improve the fairness of the student model.

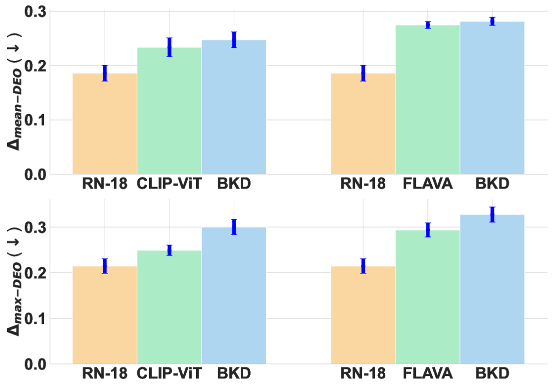

Figure 2: Δ_{DEO} scores for baseline teacher (CLIP-ViT-B/32, FLAVA), baseline student (ResNet-18), and distilled student models using base KD. We find that KD results in unfairer student predictions as compared to baseline students across both fairness metrics.

Q4) BIRD-augmented methods are fairer than their vanilla counterpart. We augment BIRD with two widely used KD methods (FitNet and AT) and MFD, a baseline to achieve fairness in KD. Our results in Table 5 demonstrate that BIRD-augmented KD techniques learn fairer representations than their vanilla counterparts. On average, BIRD improves the fairness of three existing KD methods by 41.86% (in $\Delta_{\text{mean-DEO}}$) and 41.80% (in $\Delta_{\text{max-DEO}}$), respectively. A key takeaway from our experiments is that BIRD learns a small distillation operator using meta-learning that can be easily integrated with any existing KD frameworks.

5. Conclusion

In this work, we address the problem of learning fair distilled students. To this end, we introduce BIRD, a meta-learning framework that exploits a critical connection between *"what"* and *"how much"* knowledge to distill from a given teacher. We demonstrate that BIRD leverages important student feedback to identify and transfer teacher features uncorrelated to a given protective attribute, resulting in fairer and more accurate student representation. Our results on three benchmark fairness datasets show that BIRD consistently improves the fairness (in terms of difference of equalized odds metric) compared to state-of-the-art knowledge distillation and debiasing techniques. This work paves the way for an exciting direction to develop trustworthy knowledge distillation techniques, where student feedback can guide the distillation process to distill trustworthy features from the teacher.

References

Agarwal, C. Intriguing properties of visual-language model explanations. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Agarwal, C., Lakkaraju, H., and Zitnik, M. Towards a unified framework for fair and stable graph representation learning. In *UAI*, 2021.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv*, 2021.

Bucilua, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *KDD*, 2006.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 2012.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Hooker, S. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2021.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv*, 2023.

Jandial, S., Khasbage, Y., Pal, A., Krishnamurthy, B., and Balasubramanian, V. N. Retrokd: Leveraging past states for regularizing targets in teacher-student learning. CODS-COMAD '23. ACM, 2023.

Jung, S., Lee, D., Park, T., and Moon, T. Fair feature distillation for visual recognition. In *CVPR*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *ICLR*, 2015.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *ICCV*, 2015.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Naik, R. and Nushi, B. Social biases through the text-to-image generation lens. *arXiv*, 2023.

PyTorch. 10. model zoo — pytorch/serve master documentation. https://pytorch.org/serve/model_zoo.html. (Accessed on 05/11/2023).

Quadrianto, N., Sharmanska, V., and Thomas, O. Discovering fair representations in the data domain. In *CVPR*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*, 2019.

Seth, A., Hemani, M., and Agarwal, C. Dear: Debiasing vision-language models with additive residuals. *arXiv*, 2023.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *CVPR*, 2022.

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., and Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

Wang, Z., Codella, N., Chen, Y.-C., Zhou, L., Yang, J., Dai, X., Xiao, B., You, H., Chang, S.-F., and Yuan, L. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv*, 2022.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, Online, October 2020.

Yim, J., Joo, D., Bae, J., and Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017.

Yucer, S., Tektas, F., Al Moubayed, N., and Breckon, T. P. Measuring hidden bias within face recognition via racial phenotypes. In *WACV*, 2022.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017.

A. Additional Experimental Details

In this section, we first discuss the experimental setup and datasets used in detail (Sec A.1). We then present our experimental results of BIRD on additional architectures (ShuffleNet-v2, ResNet-18, and ResNet-34) and datasets (UTKFace (Zhang et al., 2017), and CIFAR-10S (Wang et al., 2020)) in Sec A.2. Finally, we present additional experiments and the full set of numbers described in Sec 4.1.

A.1. Datasets and Experimental Setup

Datasets. We evaluate BIRD on three widely-used fairness datasets. 1) CelebA (Liu et al., 2015) dataset comprises more than 200,000 images with 40 binary attribute annotations. Following Quadrianto et al. (2019) and Jung et al. (2021), we only consider the binary protected group and binary task class in our experiment, namely, we set Gender (male/female) as the protected attribute and Attractive (yes/no) as the target variable. 2) UTKFace (Zhang et al., 2017) dataset consists of approximately 20,000 face images with annotations of age (from 0 to 116), gender (male/female), and ethnicity (White, Black, Asian, and Indian). Images in the dataset are diverse and encompass different variations in pose, facial expression, illumination, occlusion, resolution, etc. We follow the setup described by Jung et al. (2021) and use ethnicity as the protected attribute with four classes and age as the task attribute bucketed into three classes. 3) CIFAR-10 Skewed (CIFAR-10S) (Wang et al., 2020) dataset is a modified version of CIFAR-10 to study bias mitigation in image classification and consists of 32×32 images categorized into one of 10 classes.

Evaluation metrics. We report AUROC and F1-score on the test set to evaluate the predictive performance of the student. While for fairness, we use two types of difference of equalized odds (DEO) metrics as proposed by Jung et al. (2021), defined upon taking the *maximum* and the *average* over the given prediction label $\hat{y}$. Further, $\Delta_{\text{max-DEO}}$ denotes the worst-case unfairness performance and $\Delta_{\text{mean-DEO}}$ represents the overall fairness across all classes.

Baseline methods. We consider the standard knowledge distillation baselines: Base KD (BKD) (Hinton et al., 2015), Attention Transfer (Zagoruyko & Komodakis, 2017), and FitNet (Romero et al., 2015) – they entirely focus on improving student's prediction accuracy. In addition, we include recent methods proposed to tackle fairness in knowledge distillation: Adversarial debiasing (Zhang et al., 2018) and MFD (Jung et al., 2021). All hyperparameters of the chosen baselines were set following the author's guidelines.

Model architectures. We investigate the flexibility of BIRD using three established foundation models (FMs): CLIP-ResNet-50, CLIP-ViT-B/32 (Radford et al., 2021), and FLAVA (Singh et al., 2022). In addition, we consider three widely used Convolutional Neural Network (CNN) architectures in knowledge distillation to show the generalizability of BIRD in performing bias-aware distillation: ShufflenetV2 (Ma et al., 2018), ResNet-18, and ResNet-34 (He et al., 2016). We use the public implementations and pre-trained weights for FMs and CNNs models from HuggingFace (Wolf et al., 2020) and PyTorch model-zoo (PyTorch), respectively. Note that while we initialize the FMs using their pre-trained weights, the CNN models were trained from scratch.

Baseline Implementation. We use Adam optimizer (Kingma & Ba, 2015) with its default parameters and a learning rate of $1e^{-3}$ to train all our baseline models. For the CelebA dataset, all models are trained for 10 epochs with a constant learning rate. However, for the UTK dataset, we train the CNNs for 50 epochs with a decay factor of $1e^{-1}$ every 10 epochs and FMs for 10 epochs with a constant learning rate. We follow previous works and use $\alpha = 0.90$ and $\tau = 4$ for the knowledge distillation parameters for all experiments. We implement FitNet (Romero et al., 2015) following Jung et al. (2021), AT loss using the official repository [1] with $\beta = 1e^{6}$, and Adversarial Debiasing from the repository[2] provided by Wang et al. (2020). Refer to the Appendix for details on baseline hyperparameters.

Compute details. For all our experiments we use a single A100 GPU with 80GB GPU memory and CUDA version 11.2.

A.2. Additional Results

CNNs, UTKFace, and CIFAR-10S Results. We observe that BIRD consistently improves the fairness performance for all architectures in the CNN experiments for two real-world, widely used visual fairness datasets while maintaining their predictive performance (See Table 2). Interestingly, while AD does improve fairness in some cases (Table 2 CelebA) it fails to do so consistently while maintaining the F1 and AUROC scores. In Table 4, we present the results for foundation models

[1] https://github.com/szagoruyko/attention-transfer

[2] https://github.com/princetonvisualai/DomainBiasMitigation

Table 5: Results of BIRD for three different KD methods. Shown is the average performance across five independent runs on the Celeb-A dataset with ResNet18→ResNet18. BIRD consistently improves the fairness performance (shaded area) of all existing KD methods.

Method	AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}$(↓)	$\Delta_{\text{max-DEO}}$(↓)
FitNet	85.22±0.25	75.19±0.76	20.62±0.94	24.23±1.94
BIRD + FitNet	84.49±0.18	77.04±0.35	**11.32**±1.29	**16.15**±1.38
AT	86.03±0.20	75.07±1.50	18.00±1.19	22.16±1.54
BIRD + AT	86.92±0.16	78.56±0.33	**3.55**±0.37	**5.24**±0.56
MFD	86.24±0.09	77.32±0.26	19.34±0.47	21.46±0.64
BIRD + MFD	82.61±0.34	75.20±0.63	**15.25**±0.45	**18.09**±0.45

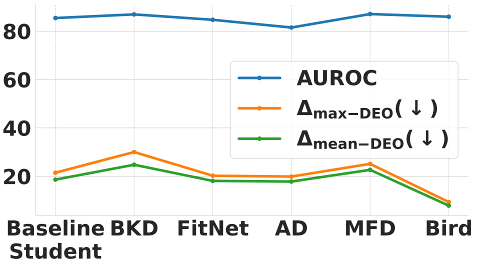

Figure 3: The effects of distillation on the fairness performance of a ResNet-18 student when distilled using CLIP-ViT-B/32 teacher for Celeb-A dataset. BIRD outperforms all methods on both fairness metrics without sacrificing ResNet-18's predictive ability.

on UTKFace dataset, showing that BIRD achieves the best fairness amongst all the predominant baselines

We reproduce the experimental setup by Wang et al. (2020) for CIFAR-10S, and show the results for the same in Table 3. We show that BIRD is able to improve both $\Delta_{\text{max-DEO}}$ by 20.19% (47.94→38.26) and $\Delta_{\text{mean-DEO}}$ by 24.71% (26.26→19.77). It is noteworthy that even AD significantly improves the fairness metrics, however, this results in a substantial drop in F1-Scores. On the other hand, BIRD obtains the best predictive performance metrics amongst all baselines.

Ablation Studies (Cont. from Sec 4). Here, we discuss our experimental results that address the questions presented in Sec 4 (Q1-Q4).

Q1). In Table 6, we present the exact metrics for several teacher-student pairs. Specifically, we observe that while the KD process improves predictive performance, it comes with a tradeoff of fairness. On average, we observe an increase of 38.54% in $\Delta_{\text{mean-DEO}}$ and an increase of 37.26% in $\Delta_{\text{max-DEO}}$ for the selected models in Table 6.

Q2). Described in Additional Results (see Sec A.2).

Q3). In Table 7, we show the efficacy of meta-gradients in our pipeline for the CelebA dataset with ResNet18→ResNet18. While a simple regularization term does produce improvements in fairness performance, it is alone not sufficient, and incorporating student feedback through meta-gradients further improves fairness.

Q4) Detailed in Sec 4.1..

A.3. Additional Hyperparameter Details

Here, we discuss the hyperparameter choices for BIRD and predominant baselines. For the CIFAR-10S dataset, we find that the widely accepted temperature τ=4 and α=0.90 do not give optimal BKD performance. Thereby, we use τ=10 and α=0.50 instead. We observe that setting the training ratio parameter (base model updates compared to domain classifier) to 3 if the total number of epochs is greater than 20, and 1 otherwise helps retain the predictive performance the best when using AD. We observe that this largely retains the predictive performance of AD. The feature distillation strength for FitNet loss is kept constant to 0.1 as in repository [3]. For the experiments conducted on foundation models in our paper, we operate under the assumption that only the penultimate representation layer is accessible, following a black-box setting. As a result, employing AT is not feasible, and that is why the corresponding results are excluded from Table 1. Lastly, since Jung et al. (2021) does not provide a public codebase, we try to implement MFD as faithfully as possible keeping testing conditions consistent across all our experiments. Please refer to Table 8 for the list of hyperparameters for different BIRD experiments.

[3] https://github.com/AberHu/Knowledge-Distillation-Zoo/

Table 2: Results of knowledge distillation on three CNN models using two fairness datasets. Shown is the average performance across five independent runs. Arrows (↑, ↓) indicate the direction of better performance. BIRD retains the predictive power (AUROC and F1-score) of the baseline model while improving their fairness (shaded area).

Model	Dataset	Method	AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}$(↓)	$\Delta_{\text{max-DEO}}$(↓)
ShuffleNetV2	UTKFace	Baseline	89.15±0.37	74.05±0.56	20.33±1.17	42.29±2.42
		BKD	90.43±0.13	74.60±0.23	20.00±0.68	38.31±2.13
		FitNet	90.02±0.29	74.55±0.42	19.67±1.02	40.40±1.68
		AT	90.70 ±0.29	76.12 ±0.44	19.34 ±1.74	40.10 ±2.58
		AD	90.16 ±0.15	75.60 ±0.29	19.60 ±0.63	40.30 ±2.19
		MFD	90.11±0.27	74.75±0.85	19.77±0.67	37.91 ±0.81
		BIRD	90.53±0.37	74.88±0.61	**16.92**±1.15	**36.12**±1.39
	CelebA	Baseline	86.01±0.04	76.44 ±1.21	23.11±0.20	28.38±0.60
		BKD	86.20±0.11	76.81 ±0.86	23.26±0.59	26.72±1.78
		FitNet	85.84±0.20	76.93±0.40	22.98±0.89	25.17±1.76
		AT	86.27 ±0.11	75.51 ±1.50	25.17 ±1.76	27.54 ±1.97
		AD	86.51 ±0.18	77.64 ±0.79	8.04 ±1.96	11.18 ±2.33
		MFD	85.88±0.08	76.72±0.52	21.59±0.39	23.81±1.06
		BIRD	88.01±0.27	79.82±0.29	**5.01**±1.04	**8.16**±2.17
ResNet18	UTKFace	Baseline	92.25±0.14	78.73±0.27	17.21±0.40	36.92±1.13
		BKD	93.06±0.17	80.22 ±0.49	18.54±0.81	39.00±2.13
		FitNet	92.75±0.11	79.35±0.20	17.41±0.97	38.31±1.58
		AT	92.92 ±0.12	80.30 ±0.24	17.88 ±0.71	36.22 ±1.52
		AD	90.93 ±0.46	78.61 ±0.47	17.18 ±0.73	36.32 ±2.32
		MFD	93.03±0.11	80.10±0.19	16.62±0.70	36.22±0917
		BIRD	91.67±0.29	77.71±0.42	**15.49**±0.77	**30.65**±3.42
	CelebA	Baseline	85.44±0.29	74.26±1.59	18.60±1.46	21.46±1.61
		BKD	86.17±0.18	74.52±1.53	17.98±1.81	21.12±1.99
		FitNet	85.22±0.25	75.19±0.76	20.62±0.94	24.23±1.94
		AT	86.03 ±0.20	75.07 ±1.50	18.00 ±1.19	22.16 ±1.54
		AD	60.19 ±2.88	55.63 ±1.69	13.78 ±4.54	16.89 ±4.42
		MFD	86.24±0.09	77.32 ±0.26	19.34±0.47	21.46±0.64
		BIRD	84.49±0.18	77.04±0.35	**11.32**±1.29	**5.31**±1.38
ResNet34	UTKFace	Baseline	92.18±0.35	78.96±0.33	17.61 ±0.90	36.52 ±1.16
		BKD	92.36±0.43	79.95 ±0.30	17.41±0.63	37.91 ±1.45
		FitNet	92.67±0.32	80.00±0.30	17.18±0.43	38.11±1.67
		AT	92.06 ±0.27	79.50 ±0.27	16.72 ±1.05	34.03 ±2.26
		AD	92.08 ±0.40	79.05 ±0.70	18.04 ±1.45	36.52 ±3.33
		MFD	92.48±0.16	78.76±0.31	16.98 ±0.78	35.42 ±1.08
		BIRD	90.90±0.16	77.74±0.43	**15.92**±0.55	**33.13**±1.31
	CelebA	Baseline	85.93±0.31	75.73±0.53	20.43±1.11	24.71 ±1.71
		BKD	86.32±0.36	77.04±0.26	20.88±1.43	23.29 ±1.83
		FitNet	85.93±0.19	75.30±0.56	19.27±1.74	23.30±2.10
		AT	85.95 ±0.38	74.89 ±0.80	21.05 ±0.79	25.35 ±1.59
		AD	69.32 ±4.14	61.83 ±3.10	28.52 ±7.25	34.12 ±8.88
		MFD	87.10±0.10	78.04 ±0.06	17.48 ±0.80	17.91 ±0.90
		BIRD	84.31±0.37	73.90±0.55	**10.31**±2.88	**13.59**±3.92

Table 3: Results on CIFAR-10S dataset across 5 independent runs for ResNet18→ResNet18. Arrows (↑, ↓) indicate the direction of better performance. BIRD retains the predictive power (AUROC and F1-Score) of the baseline model while improving the fairness criterion ($\Delta_{\text{mean-DEO}}$ and $\Delta_{\text{max-DEO}}$)

Method	AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}(\downarrow)$	$\Delta_{\text{max-DEO}}(\downarrow)$
Baseline	98.91±0.02	88.34±0.17	26.26±0.70	47.94±1.94
BKD	98.95±0.02	88.90±0.13	25.30±0.63	46.92±2.16
FitNet	98.89±0.01	88.15±0.08	26.55±0.66	48.86±1.85
AT	98.99±0.02	88.95±0.12	25.16±0.33	46.08±2.27
AD	98.44±0.11	85.98±0.43	**16.20**±1.18	**31.94**±3.89
MFD	98.93±0.03	88.32±0.10	27.27±0.34	49.16±1.62
BIRD	**99.12**±0.02	**89.45**±0.14	**19.77**±0.37	**38.26**±1.73

Table 4: Results of knowledge distillation on three foundation models using UTK dataset. Shown is the average performance across five independent runs. Arrows (↑, ↓) indicate the direction of better performance. BIRD retains the predictive power (AUROC and F1-score) of the baseline model while improving their fairness (shaded area). Here, R50 is ResNet-50, and ViT-32 is ViT-B/32.

Model	Method	AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}(\downarrow)$	$\Delta_{\text{max-DEO}}(\downarrow)$
Flava	Baseline	94.43±0.04	81.12±0.28	14.49±0.49	32.54±1.62
	BKD	94.43±0.04	81.02±0.23	15.02±0.51	32.84±1.51
	FitNet	94.42±0.10	81.54±0.48	15.32±1.32	32.74±3.55
	AD	92.81 ±0.09	77.21 ±0.44	17.08 ±0.12	37.11 ±0.21
	MFD	94.42±0.10	81.54±0.48	15.42±1.38	33.03 ±3.73
	BIRD	94.00±0.02	80.74±0.54	**14.23**±0.55	**28.16**±2.89
CLIP-ViT/32	Baseline	95.96±0.03	86.22±0.16	13.47±0.20	25.07±0.96
	BKD	95.95±0.05	86.02 ±0.15	13.73±0.15	25.27±0.99
	FitNet	96.00±0.03	86.27 ±0.31	14.16±0.37	25.27±1.74
	AD	96.05 ±0.06	86.34 ±0.32	11.84 ±0.73	22.49±0.31
	MFD	96.05±0.04	86.64±0.15	12.11±0.16	22.79±0.37
	BIRD	95.50±0.04	85.67±0.08	**12.07**±0.27	**16.92**±0.82
CLIP-R50	Baseline	95.67±0.04	84.90±0.20	13.70 ±0.58	23.08 ±1.06
	BKD	95.67±0.03	84.85 ±0.20	13.57±0.50	23.28 ±0.99
	FitNet	95.73±0.04	85.02 ±0.44	13.53 ±0.63	22.89 ±1.40
	AD	95.67 ±0.05	83.86 ±0.16	14.83 ±1.58	26.27 ±0.70
	MFD	95.69±0.03	84.90±0.52	14.16 ±0.60	22.99 ±1.60
	BIRD	95.43±0.02	84.05±0.13	**12.43**±0.14	**23.28**±0.43

Table 6: Results of Base Knowledge Distillation on different teacher-student pairs. Shown is the average performance across five independent runs. We establish that across different architectures, knowledge distillation results in unfair student models by following the fairness properties ($\Delta_{\text{mean-DEO}}$, $\Delta_{\text{max-DEO}}$) of the teacher.

Baselines		AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}(\downarrow)$	$\Delta_{\text{max-DEO}}(\downarrow)$
FLAVA		84.43±0.12	74.87±0.63	27.48±0.64	29.37±1.53
CLIP-ViT-32		87.01±0.26	78.15±0.52	23.38±1.72	24.91±1.15
CLIP-R50		87.72±0.06	78.71±0.21	21.11±0.30	21.97±0.40
ResNet18		85.44±0.29	74.26±1.59	18.60±1.46	21.46±1.61
ResNet34		85.93±0.31	75.73±1.25	20.43±1.11	24.71±1.71
Teacher	Student	AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}(\downarrow)$	$\Delta_{\text{max-DEO}}(\downarrow)$
FLAVA	ResNet18	85.24±0.09	76.30±0.81	28.15±0.75	32.74±1.65
FLAVA	ResNet34	84.81±0.32	75.46±0.89	29.50±1.29	33.34±2.29
CLIP-ViT-32	ResNet18	86.94±0.37	77.14±1.55	24.75±1.46	30.01±1.65
CLIP-ViT-32	ResNet34	86.71±0.15	78.11±0.39	23.68±0.24	25.38±1.00
CLIP-R50	ResNet18	86.71±0.37	77.96±0.63	25.61±1.86	30.08±1.49
CLIP-R50	ResNet34	87.31±0.15	78.11±0.39	23.68±0.24	25.38±1.0
ResNet34	ResNet18	86.25±0.76	73.45±1.30	19.80±1.62	23.86±1.67

Table 7: Ablation study to understand the impact of meta-gradients in BIRD. Shown is the average performance across five independent runs on the Celeb-A dataset with ResNet18→ResNet18, evidencing that the student feedback through the meta-step update improves fairness.

Method	AUROC (↑)	F1-score (↑)	$\Delta_{\text{mean-DEO}}(\downarrow)$	$\Delta_{\text{max-DEO}}(\downarrow)$
Baseline	85.44±0.29	74.26±1.59	18.60±1.46	21.46±1.61
BIRD w/o Meta	88.33±0.22	77.25±1.56	13.85±3.20	17.90±4.04
BIRD	84.49±0.18	77.04±0.35	**11.32**±1.29	**16.15**±1.38

Table 8: Hyperparameters for BIRD for different datasets and architectures. We perform minimal linear probing to find the optimal λ (See 3.4) for each setting.

Architecture	Dataset	λ	Warmup	Total Epochs
FLAVA	CelebA	0.1	5	10
CLIP-ViT-32	CelebA	0.2	5	10
CLIP-R50	CelebA	0.1	5	10
ShuffleNetV2	CelebA	0.05	5	10
ResNet18	CelebA	0.2	5	10
ResNet34	CelebA	0.1	5	10
FLAVA	UTK	0.1	20	50
CLIP-ViT-32	UTK	0.2	20	50
CLIP-R50	UTK	0.1	20	50
ShuffleNetV2	UTK	0.05	20	50
ResNet18	UTK	0.2	20	50
ResNet34	UTK	0.1	20	50
ResNet18	CIFAR-10S	0.2	70	100