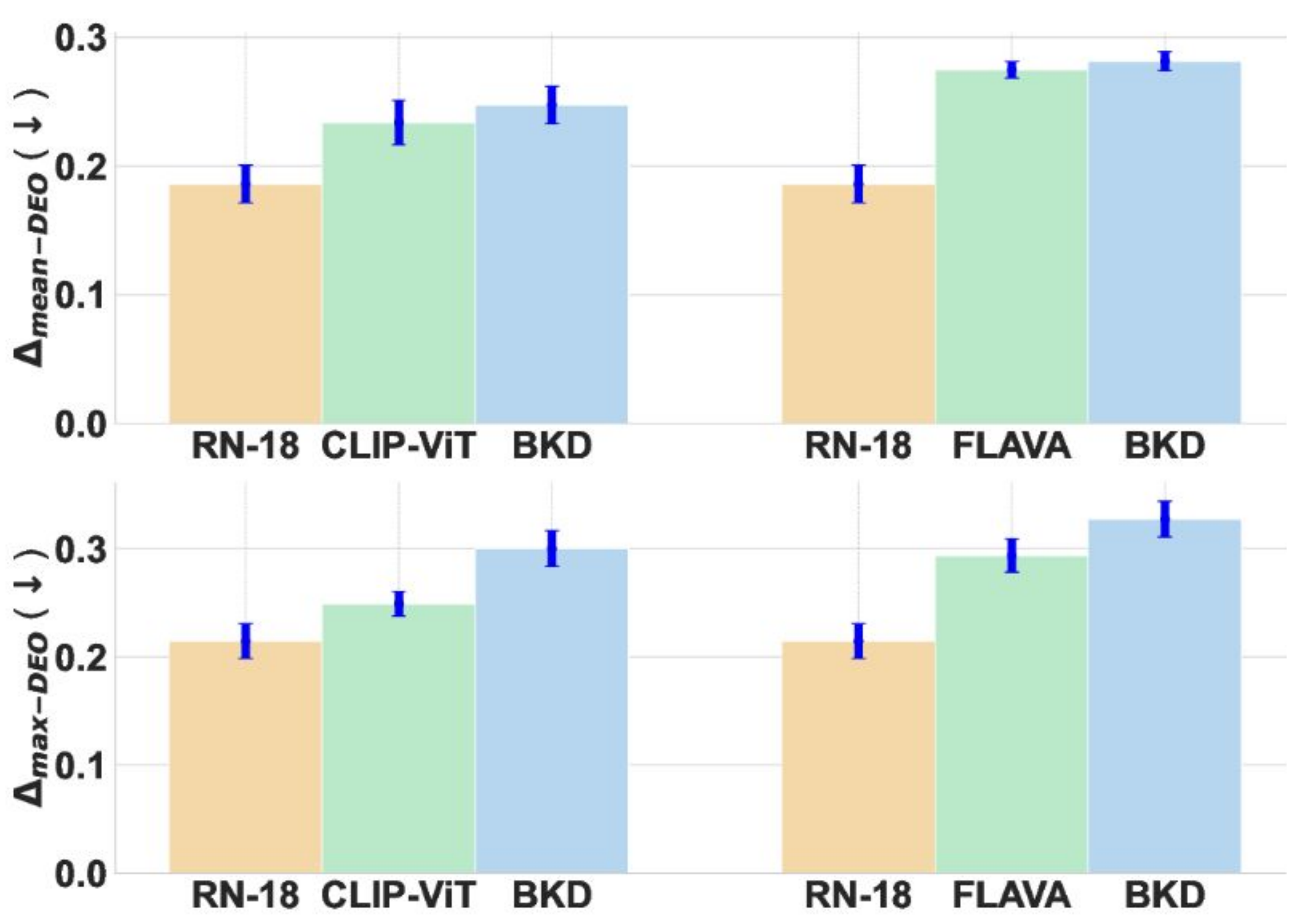
Towards Fair Knowledge Distillation using Student Feedback

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

[1]MDSR Labs, Adobe, India [2]Harvard University

Fairness & Knowledge Distillation: How can we incorporate student feedback to perform Bias Aware Distillation?

Student mimics the fairness properties of the teacher

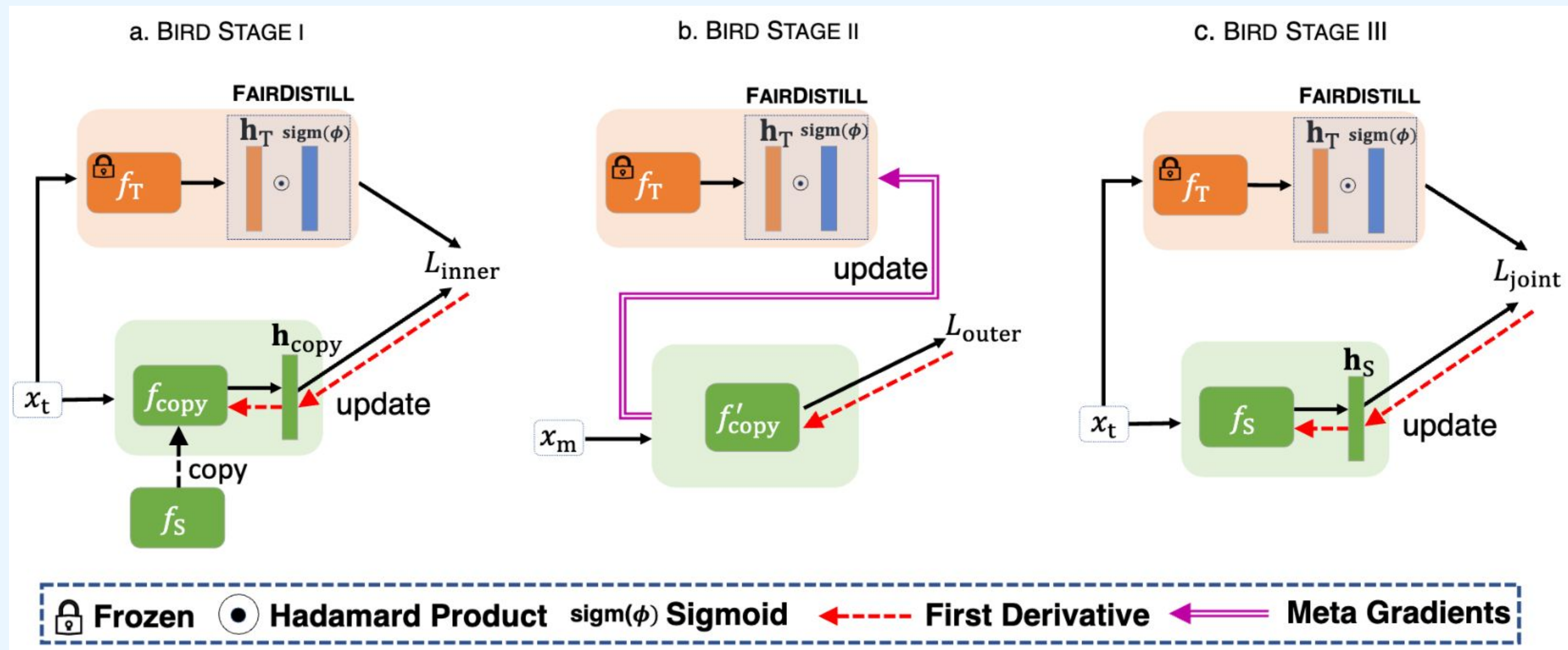


Fairness scores for baseline teacher (CLIP-ViT/B32, FLAVA), baseline student (ResNet-18), and distilled student models using base KD (BKD)

Problem Formulation

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).

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 L_{inner},
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 L_{outer}, and
c. in Stage III, the student model f_S is distills unbiased representations using ***FAIRDISTILL*** (from Stage II).

Model	Method	AUROC (↑)	$\Delta_{mean-DEO}$ (↓)	$\Delta_{max-DEO}$ (↓)
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

BIRD improves fairness of knowledge distillation

Shown is the comparative performance of **BIRD** on CelebA Dataset (Left) for three foundation models and on CIFAR10-S dataset (Bottom) for ResNet18→ResNet18. Note that all results indicate avg. performance across five independent runs. Arrows (↑↓) indicate the direction of desired performance. *BIRD* retains the predictive power (AUROC) of the baseline while improving fairness criterion (shaded).

Method	AUROC (↑)	F1-score (↑)	$\Delta_{mean-DEO}$ (↓)	$\Delta_{max-DEO}$ (↓)
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

java.abhinav99@gmail.com