NEURAL INFORMATION PROCESSING SYSTEMS

Fair Deepfake Detectors Can Generalize

Harry Cheng[1], Ming-Hui Liu[2], Yangyang Guo[1], Tianyi Wang[2], Liqiang Nie[3], Mohan Kankanhalli[1]

[1]National University of Singapore

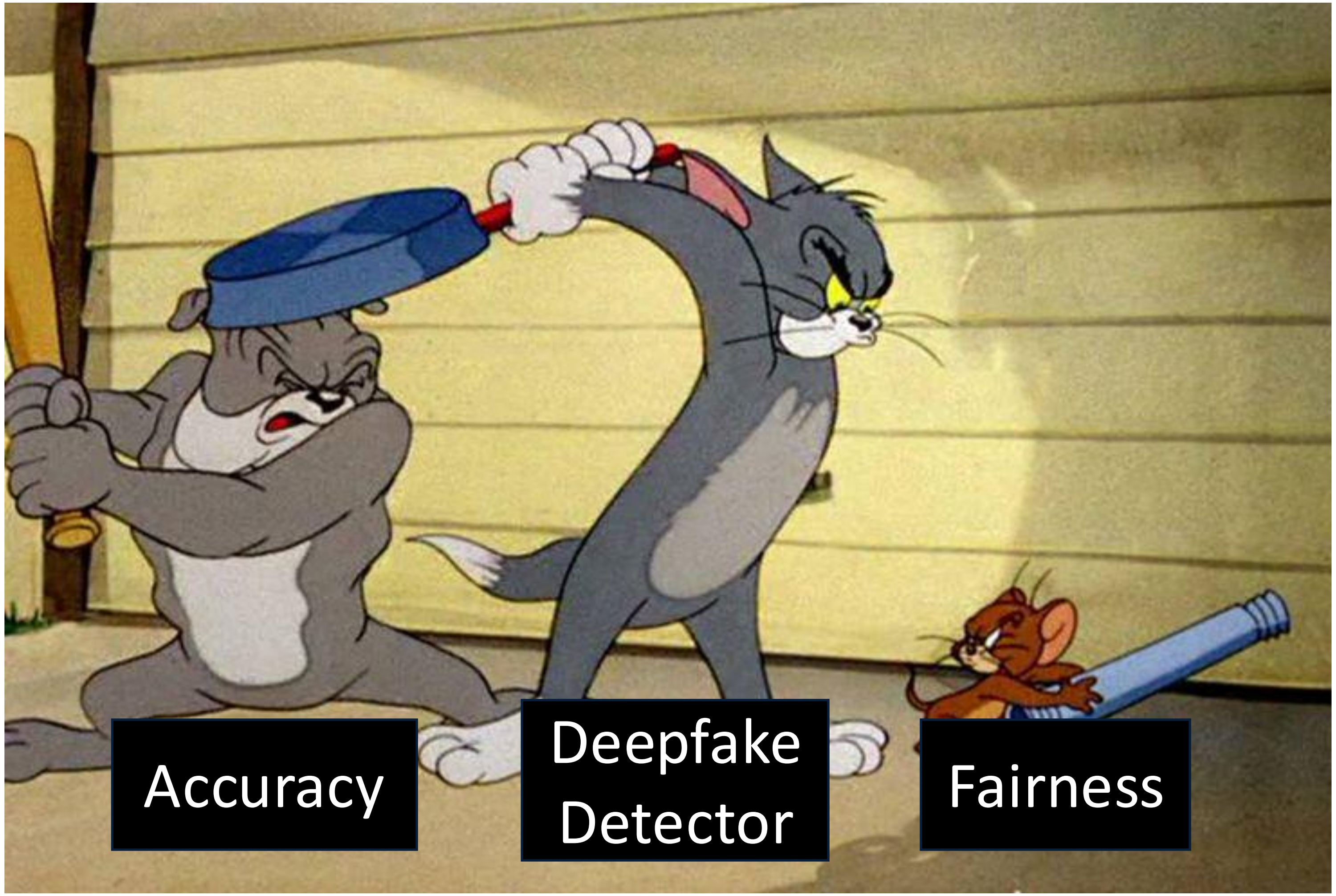
[2]Shandong University

[3]Harbin Institute of Technology (Shenzhen)

Motivation

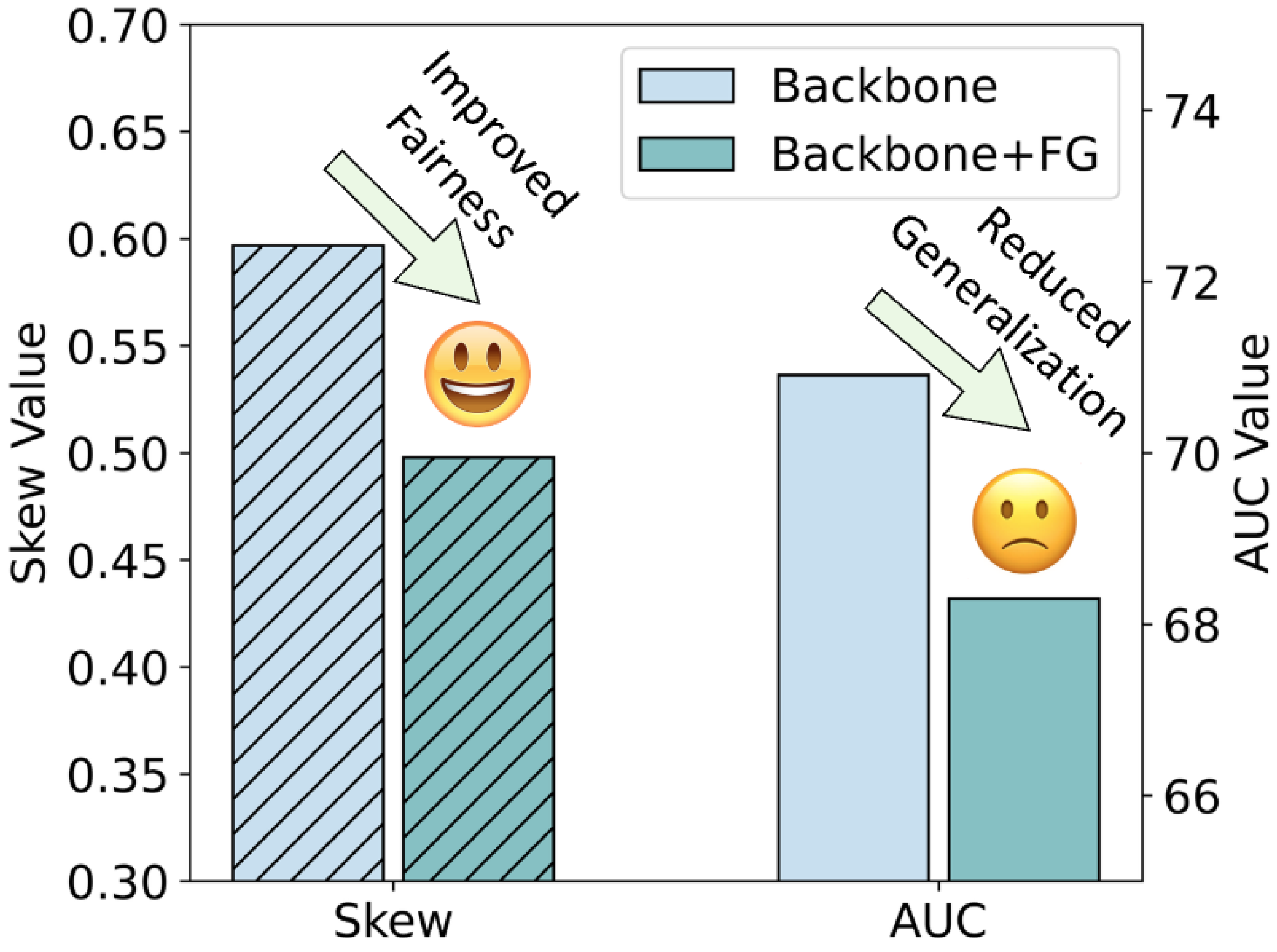
There is often a trade-off between fairness and generalization of a deepfake detector.

- **Improving generalization does not necessarily enhance fairness.**
- **Bolstering fairness can inadvertently undermine a model's ability to generalize.**



Motivation

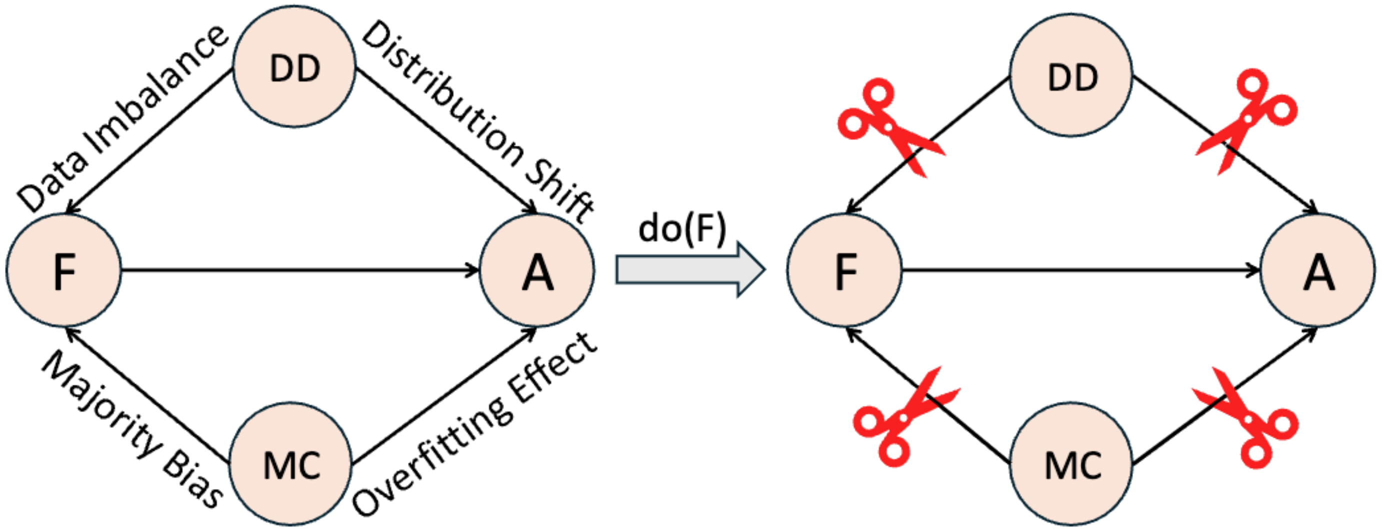
Skew (fairness metric, the lower the better)



AUC (generalization metric, the higher the better).

Comparison of model performance on Celeb-DF

Our Solution

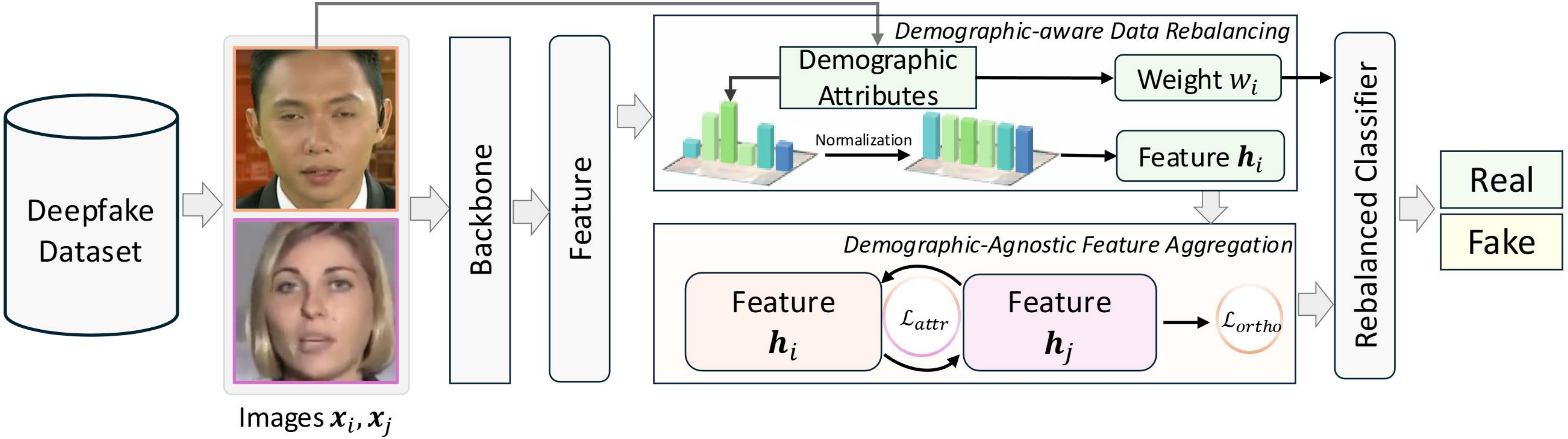


In our work, we demonstrate that improving fairness can, in some cases, enhance generalization—when confounding factors are rigorously controlled

This phenomenon is captured by the causal graph we constructed, in which the data distribution (DD) and the model capacity (MC) are defined as two confounding variables.

$$\mathbb{P}(A \mid \mathrm{do}(F=f)) = \sum_{dd,mc} \underbrace{\mathbb{P}(A \mid F=f, DD=dd, MC=mc)\,\mathbb{P}(DD=dd, MC=mc)}_{\text{Back-door Adjustment}},$$

Our Solution


Deepfake Dataset
Images $\boldsymbol{x}_i, \boldsymbol{x}_j$
Backbone
Feature
Demographic-aware Data Rebalancing
Demographic Attributes
Weight w_i
Normalization
Feature $\boldsymbol{h}_i$
Demographic-Agnostic Feature Aggregation
Feature $\boldsymbol{h}_i$
$\mathcal{L}_{attr}$
Feature $\boldsymbol{h}_j$
$\mathcal{L}_{ortho}$
Rebalanced Classifier
Real
Fake

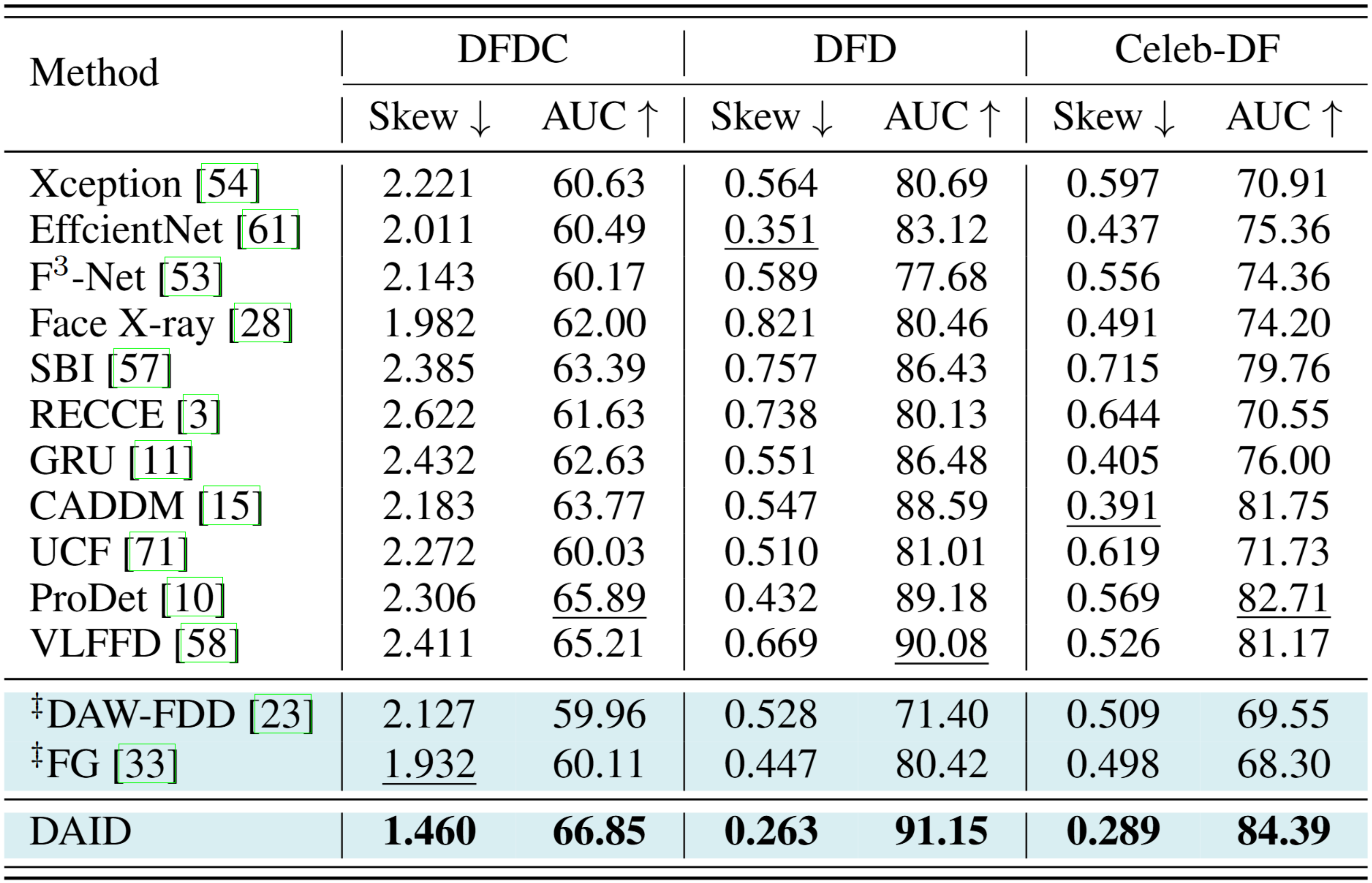
$$w_i = \left(\prod_{k=1}^{K} \widehat{\mathbb{P}}(\mathbf{s}_i^{(k)}) \right)^{-1},$$

$$\mathcal{L}_{\text{attr}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathcal{L}_{\cos}(\hat{\mathbf{h}}_i, \hat{\mathbf{h}}_j)$$

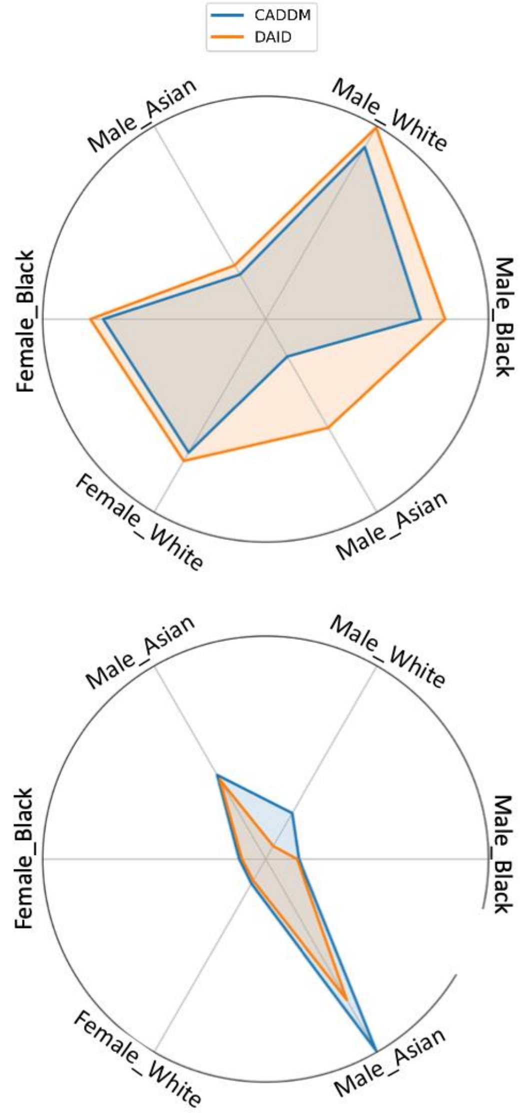
Penalizing groups with too many samples

Features from different groups but sharing the same label should be mapped to similar features

Experiments



Method	DFDC		DFD		Celeb-DF	
	Skew ↓	AUC ↑	Skew ↓	AUC ↑	Skew ↓	AUC ↑
Xception [54]	2.221	60.63	0.564	80.69	0.597	70.91
EffcientNet [61]	2.011	60.49	0.351	83.12	0.437	75.36
F^3-Net [53]	2.143	60.17	0.589	77.68	0.556	74.36
Face X-ray [28]	1.982	62.00	0.821	80.46	0.491	74.20
SBI [57]	2.385	63.39	0.757	86.43	0.715	79.76
RECCE [3]	2.622	61.63	0.738	80.13	0.644	70.55
GRU [11]	2.432	62.63	0.551	86.48	0.405	76.00
CADDM [15]	2.183	63.77	0.547	88.59	0.391	81.75
UCF [71]	2.272	60.03	0.510	81.01	0.619	71.73
ProDet [10]	2.306	65.89	0.432	89.18	0.569	82.71
VLFFD [58]	2.411	65.21	0.669	90.08	0.526	81.17
‡DAW-FDD [23]	2.127	59.96	0.528	71.40	0.509	69.55
‡FG [33]	1.932	60.11	0.447	80.42	0.498	68.30
DAID	**1.460**	**66.85**	**0.263**	**91.15**	**0.289**	**84.39**



Thanks!

Project

xaCheng1996 AT gmail DOT com

Contract