Appendix for "Open-Unfairness Adversarial Mitigation for Generalized Deepfake Detection"

1. Complexity analysis

Table 1. Comparison of computation costs.

Method	Eff-b3 [2]	UCF [3]	LSDA [4]	FG-DD [1]	AdvOU Ours
Num. Param.(M)	10.70	26.09$^{+15.39}$	96.71$^{+86.01}$	38.77$^{+28.07}$	**13.03**$^{+2.33}$
Train time(s/epoch)	942.7 ± 5.1	3568.1 ± 124.7	3510.9 ± 4023.8	7072.6 ± 18.4	**1677.6** $\pm$ **101.4**

We compare the parameter count and per-epoch training time of our AdvOU framework with state-of-the-art methods. As shown in Table 1, AdvOU introduces only 2.33M additional parameters to the EfficientNet-b3 (Eff-b3) backbone, attributable to its lightweight unfairness discoverer module. With a training cost of 1678 seconds/epoch caculated over 5 times of training epoch, AdvOU remains competitive against UCF [3], LSDA [4], and FG-DD [1] . This efficiency stems from structural advantages: UCF [3] and FG-DD [1] require computationally reconstruction operations, while LSDA [4] relies on multi-teacher distillation that scales parameter overhead. In contrast, AdvOU achieves fairness integration solely through alternating optimization between the unfairness discoverer and deepfake detector.

2. Trade-offs between fairness and accuracy

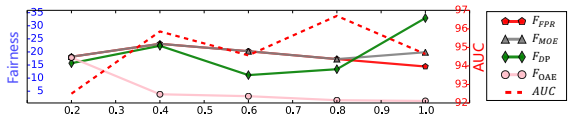

Figure 1. Trade-offs between fairness and accuracy metrics.

We analyze the trade-off between fairness and accuracy on FF++ by varying the weight of the fairness mitigation loss in Equation 9. As shown in Figure 1, increasing the mitigation strength (x-axis) steadily improves F_{OAE} and F_{FPR}, while other fairness metrics and AUC follow a rise-then-drop trend. These results underscore the importance of balancing fairness and accuracy when addressing open unfairness in deepfake detection.

3. Hyperparameter analysis

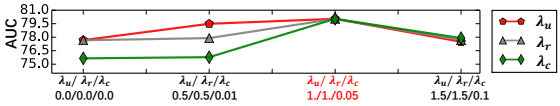

Figure 2. Sensitivity analysis for hyperparameters.

We conduct sensitivity analysis on three key hyperparameters (λ_u in Equation 4 as well as λ_r and λ_c in Equation 9) using FF++ for training and evaluating generalization on unseen datasets(CDF, DFDC and DFD). Figure 2 illustrates the trade-off between hyperparameter values and cross-dataset average AUC. Our analysis reveals that performance is most sensitive to λ_c: AUC rises from 75.65 ($\lambda_c = 0.0$) to 80.04 ($\lambda_c = 0.1$) before declining to 77.90 at $\lambda_c = 0.15$. This non-monotonic relationship suggests λ_c critically balances robustness against unfairness-induced feature perturbations, with optimal regularization occurring at $\lambda_c = 0.1$.

References

[1] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2024. 1

[2] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[3] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. 1

[4] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024. 1