CERIAS

The Center for Education and Research in Information Assurance and Security

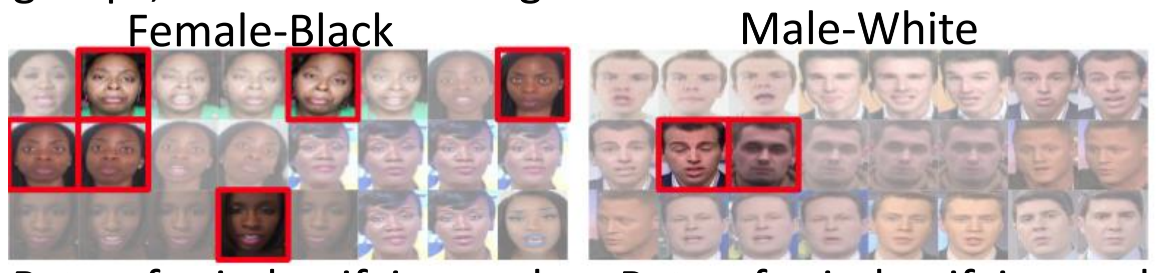
Preserving Fairness Generalization in Deepfake Detection (CVPR'24)

Li Lin[1], Xinan He[2], Yan Ju[3], Xin Wang[4], Feng Ding[2], Shu Hu[1]*
[1]Purdue University {lin1785, hu968}@purdue.edu
[2]Nanchang University {shahur, fengding}@ncu.edu.cn
[3]University at Buffalo, State University of New York yanju@buffalo.edu
[4]University at Albany, State University of New York xwang@albany.edu

Motivation and Introduction

1.Deepfakes, created through advanced AI techniques, are highly realistic media that can pose serious threats like misinformation and political manipulations.

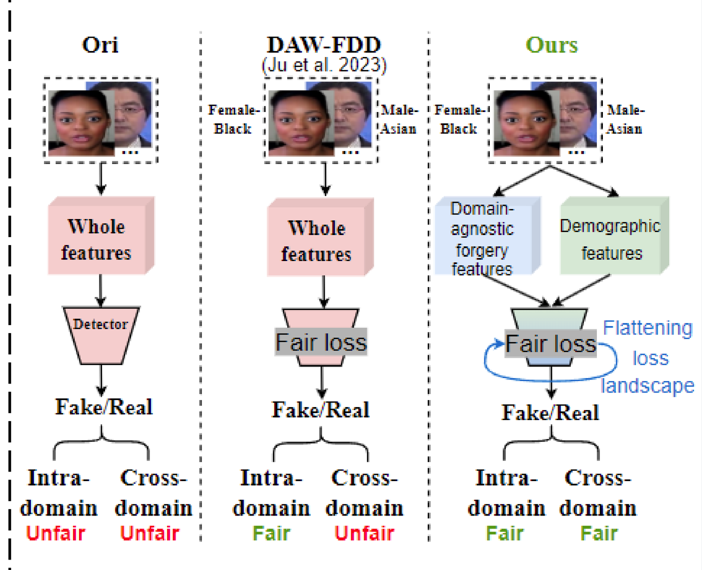
2.Unfairness in Deepfake: detection models have unfair performance disparities among demographic groups, such as race and gender.



Rate of misclassifying real as fake = 26.27% » Rate of misclassifying real as fake = 8.33%

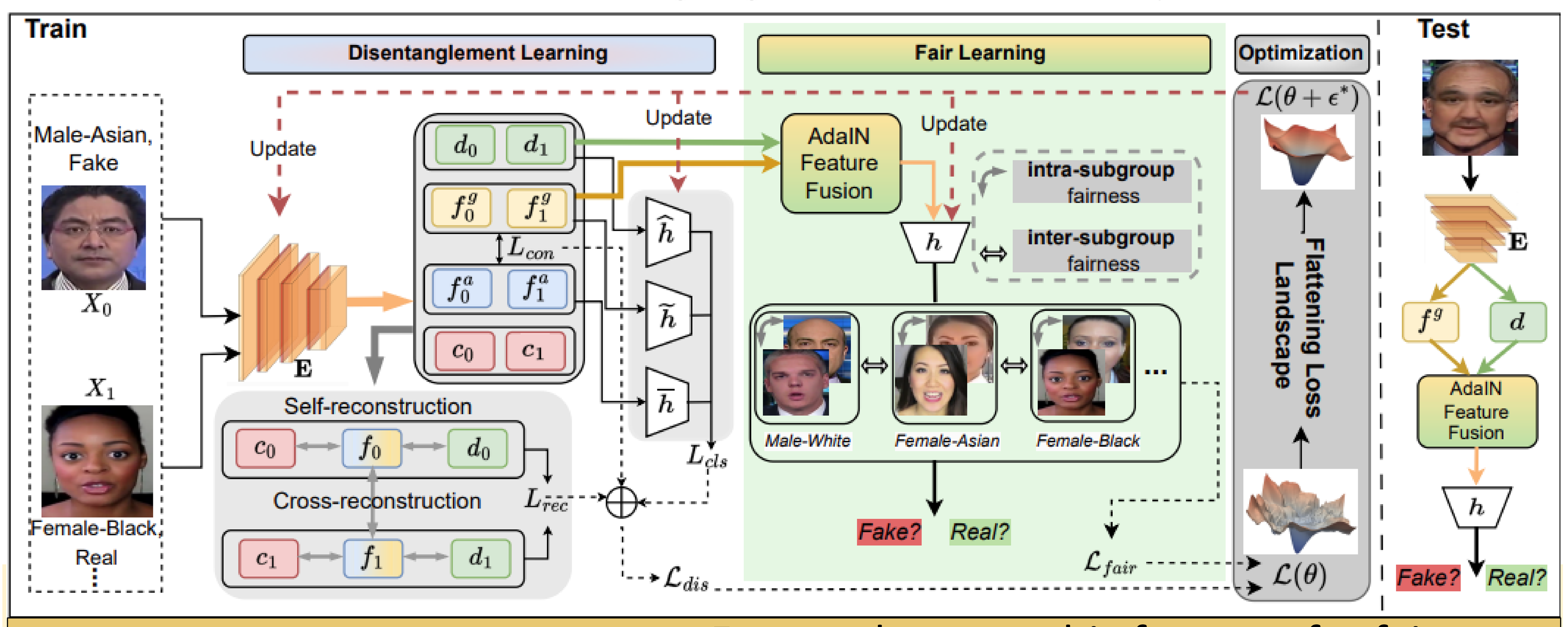
Such misclassification in AI detection models can erode public trust, highlighting the necessity of preserving fairness for secure AI system.

3.Fairness Generalization can guarantee fair detection under intra-domain and cross-domain scenarios (detect deepfakes generated by unknown forgeries). Enhance trustworthiness of AI systems and strengthen AI security.



Methodology

Objective: Train a fair deepfake detector using S can then generalize to an unseen dataset while maintaining high detection accuracy.



Disentanglement Learning	Expose demographic features for fairness and forgery features for generalization.

$\mathcal{L}_{dis} = \frac{1}{n}\sum_i[L_{cls} + \rho_1 L_{con} + \rho_2 L_{rec}]$, ρ_1, ρ_2 are trade-off hyperparameters.

Classification Loss L_{cls}: disentangle demographic, forgery features.

Contrastive Loss L_{con}: enhance the encoder's representation capabilities.

Reconstruction Loss L_{rec}: ensure the reconstructed image and the original images are consistent at the pixel level.

Fair Learning	Fuse demographic and domain-agnostic forgery features using AdaIN with fair loss.

$\mathcal{L}_{fair} = \min_{\eta\in\mathbb{R}} \eta + \frac{1}{\alpha|J|}\sum_{j=1}^{|J|}[L_j - \eta]_+$, J represents user-defined subgroups e.g., Male-Asian or Female-Black. $|J|$ represents the size of set J, L_j is the subgroup loss. $\alpha \in (0,1)$ is user-predefined hyperparameter.

Optimization	Flatten loss landscape with sharpness-aware minimization method.

$\min_\theta(\mathcal{L}_{dis} + \lambda\mathcal{L}_{fair})(\theta + \epsilon^*)$, θ is model weights, ϵ^* is obtained through perturbing θ and calculating the gradient of $\mathcal{L}$. The intuition is that the perturbation along the gradient norm direction increases the loss value significantly and then makes the model more generalizable in terms of fairness.

Results

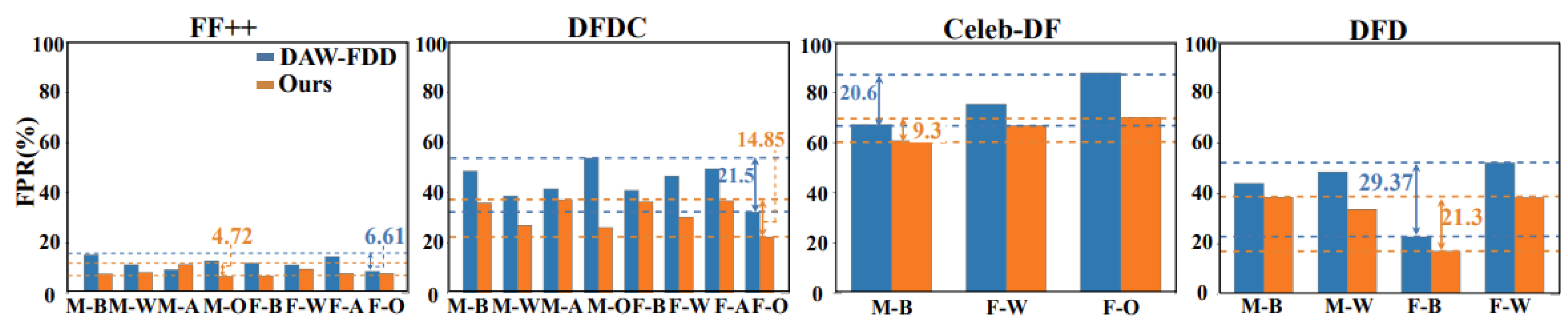
Performance of Fairness Generalization

Dataset	Method	Xception [63]					ResNet-50 [64]					EfficientNet-B3 [65]				
		Fairness Metrics(%)↓				Detection Metric(%)↑	Fairness Metrics(%)↓				Detection Metric(%)↑	Fairness Metrics(%)↓				Detection Metric(%)↑
		F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC	F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC	F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC
FF++	Ori [34]	31.31	17.69	11.12	10.08	92.77	34.69	17.29	9.83	8.85	94.83	18.78	33.21	31.36	26.01	93.55
	DAW-FDD [6]	14.06	10.55	10.97	8.72	97.46	30.36	9.74	8.89	7.42	93.23	23.33	26.15	24.74	21.23	94.92
	UCF [17]	21.52	13.06	15.06	10.58	97.10	35.13	10.87	10.81	8.05	95.92	20.92	33.08	30.01	24.56	94.21
	Ours	**10.63**	**8.15**	**10.41**	**7.60**	**98.28**	**22.70**	**9.28**	**8.72**	**5.74**	**97.72**	**11.19**	**20.61**	**18.40**	**16.18**	**95.39**
DFDC	Ori [34]	52.77	37.78	13.87	30.30	56.72	45.84	28.89	16.67	26.25	58.08	62.38	37.56	22.44	25.93	57.81
	DAW-FDD [6]	45.14	35.77	18.59	14.07	59.96	44.07	34.14	18.72	24.58	60.11	50.73	43.79	18.31	29.57	58.29
	UCF [17]	53.07	44.44	15.70	23.22	60.03	43.39	35.62	15.86	19.15	**61.06**	42.79	40.54	19.35	21.13	58.85
	Ours	**40.73**	**34.48**	**9.69**	**13.71**	**61.47**	**37.17**	**27.78**	**10.94**	**18.52**	59.76	**22.89**	**33.78**	**12.35**	**16.73**	**60.67**
Celeb-DF	Ori [34]	27.55	25.65	17.74	58.44	62.66	24.94	22.32	19.47	48.62	70.64	30.86	27.47	19.15	59.32	62.36
	DAW-FDD [6]	22.31	20.60	**11.65**	49.71	69.55	26.82	21.93	20.80	47.14	75.70	31.36	21.79	6.91	**50.86**	70.14
	UCF [17]	27.81	25.96	16.51	48.63	71.73	32.17	28.28	19.38	45.15	76.44	24.95	22.41	15.14	58.48	72.65
	Ours	**10.62**	**12.77**	15.04	**36.01**	**74.42**	**11.55**	**17.01**	**17.21**	**29.58**	**78.55**	**13.00**	**9.73**	**5.21**	55.74	**75.32**
DFD	Ori [34]	35.14	28.52	15.31	12.95	74.34	31.76	26.91	5.90	28.48	76.02	39.37	38.57	20.01	17.00	75.87
	DAW-FDD [6]	34.02	29.37	15.75	11.31	71.42	33.05	24.24	7.12	27.08	77.05	32.72	28.74	17.12	24.70	74.76
	UCF [17]	42.66	33.41	20.24	19.84	81.88	42.54	33.17	5.24	30.98	78.97	36.59	27.32	25.83	9.36	76.76
	Ours	**26.08**	**21.37**	**11.65**	**8.37**	**84.82**	**25.71**	**20.02**	**2.34**	**25.60**	**79.67**	**29.34**	**24.52**	**11.46**	**5.11**	**77.28**

Table 1. Comparison with different methods in improving fairness and detection generalization under both intra-domain (FF++) and cross-domain (DFDC, Celeb-DF, and DFD) scenarios.

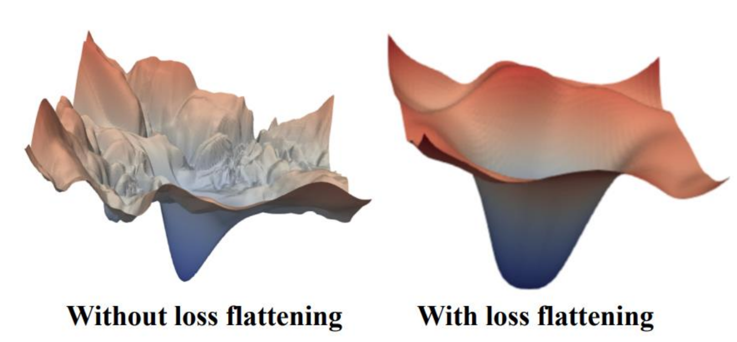
- Our method achieves the **best performance on all datasets** with different settings of backbones.

Comparison on Intersectional Subgroups



- Ours leads to a **marked reduction** in the FPR across all test datasets.

Visualization of Loss Landscape



- With flattening, the landscape becomes smoother, suggesting an **easier optimization path**, potentially leading to better training and generalization.

Visualization of the Saliency Map

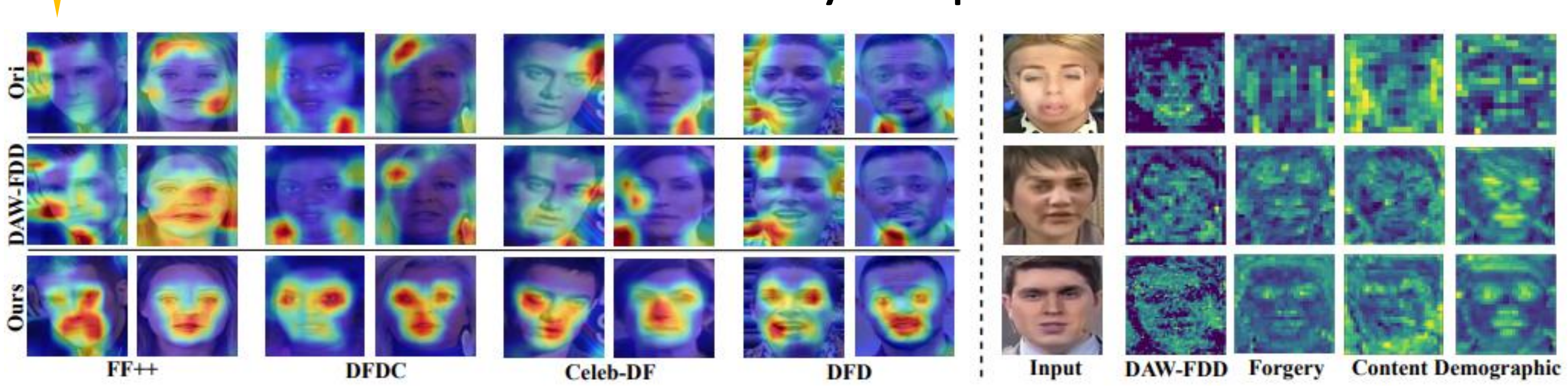


Figure 1. (**Left**) Grad-CAM visualization of Ori's, DAW-FDD, and ours on the intra-domain dataset (FF++), and cross-domain datasets. (**Right**) Visualization of the image, and features.

- Our method's activation region demonstrates a consistent model **focus on facial** salient features, irrespective of the dataset.

Conclusion & Future Work

- We propose the **first** method to improve fairness generalization in deepfake detection by addressing features, loss, and optimization.
- Our method outperforms state-of-the-art approaches in preserving fairness generalization.

In the Future, we aim to preserve fairness generalization while detecting images generated by diffusion models or GANs.

Acknowledgements. This work is accepted by CVPR 2024.
Shu Hu is supported by the Purdue University start-up grant.

Lab GitHub

PURDUE UNIVERSITY

CVPR JUNE 17-21, 2024 SEATTLE, WA

CERIAS