Generalizable Deepfake Detection via Simplicity-Bias-aware CLIP Adaptation

ICPR 2026

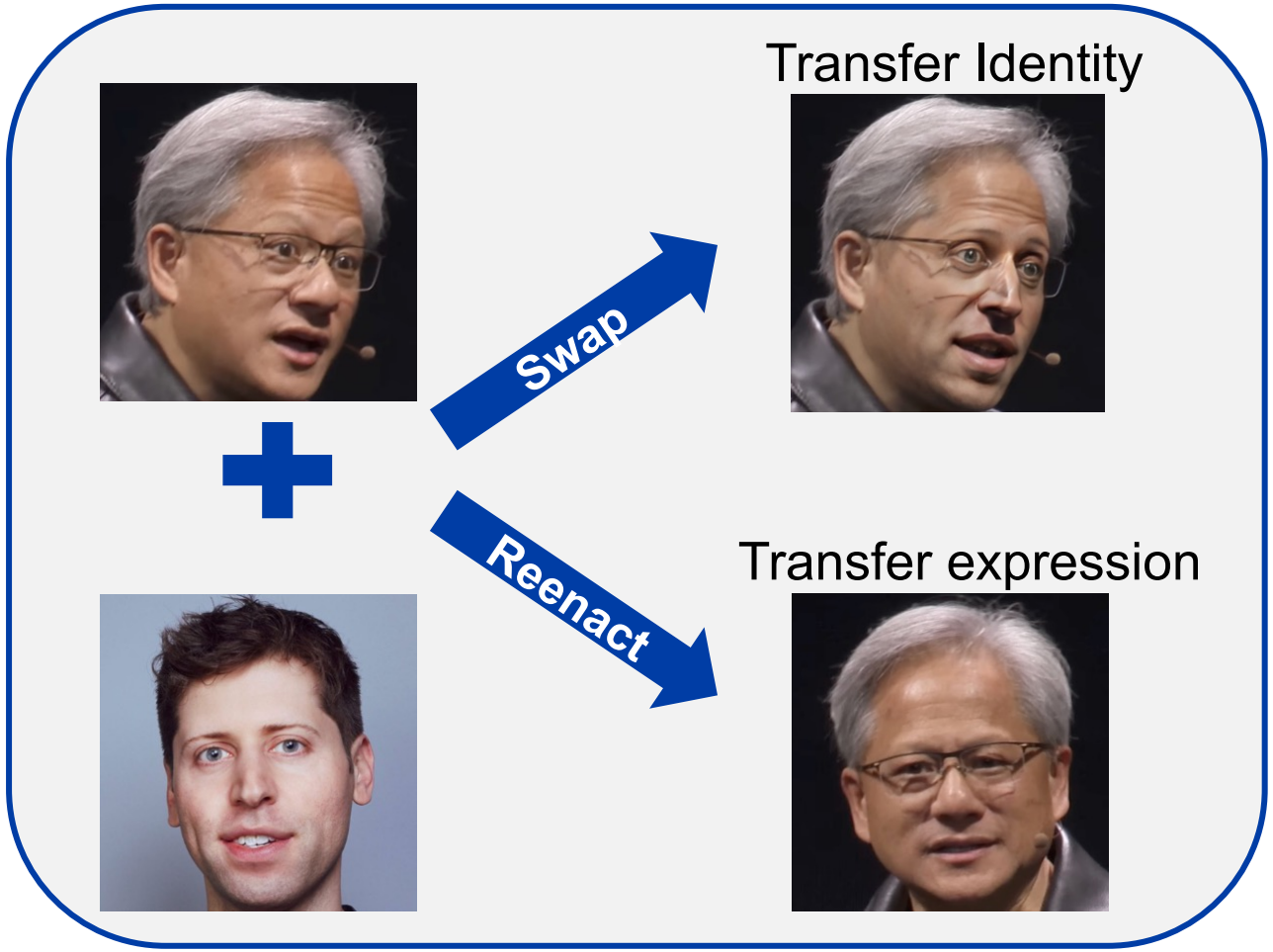
Charbel Yahchouchi[1,2], Noemi Roggero[2], Laurent Saroul[2], Antitza Dantcheva[1]

[1]Inria Center at Université Côte d'Azur, [2]Probayes

Outline

1. Introduction

Deepfakes [1] are videos that have been edited or fully generated using Deep Generative models



[1] Pei et al., "Deepfake generation and detection: A benchmark and survey". ACM Computing Surveys'26

Outline

2. Related Work

Approaches	Cues
Space domain	Frame level artifacts [2,3] : Texture, Noise and Boundaries
Time domain	Motion artifacts [5,6]: Gaze, Lips motion
Frequency domain	Frequency artifacts [4], Fourier patterns [7]
Data driven	Learn features and consistency from data [8,9]

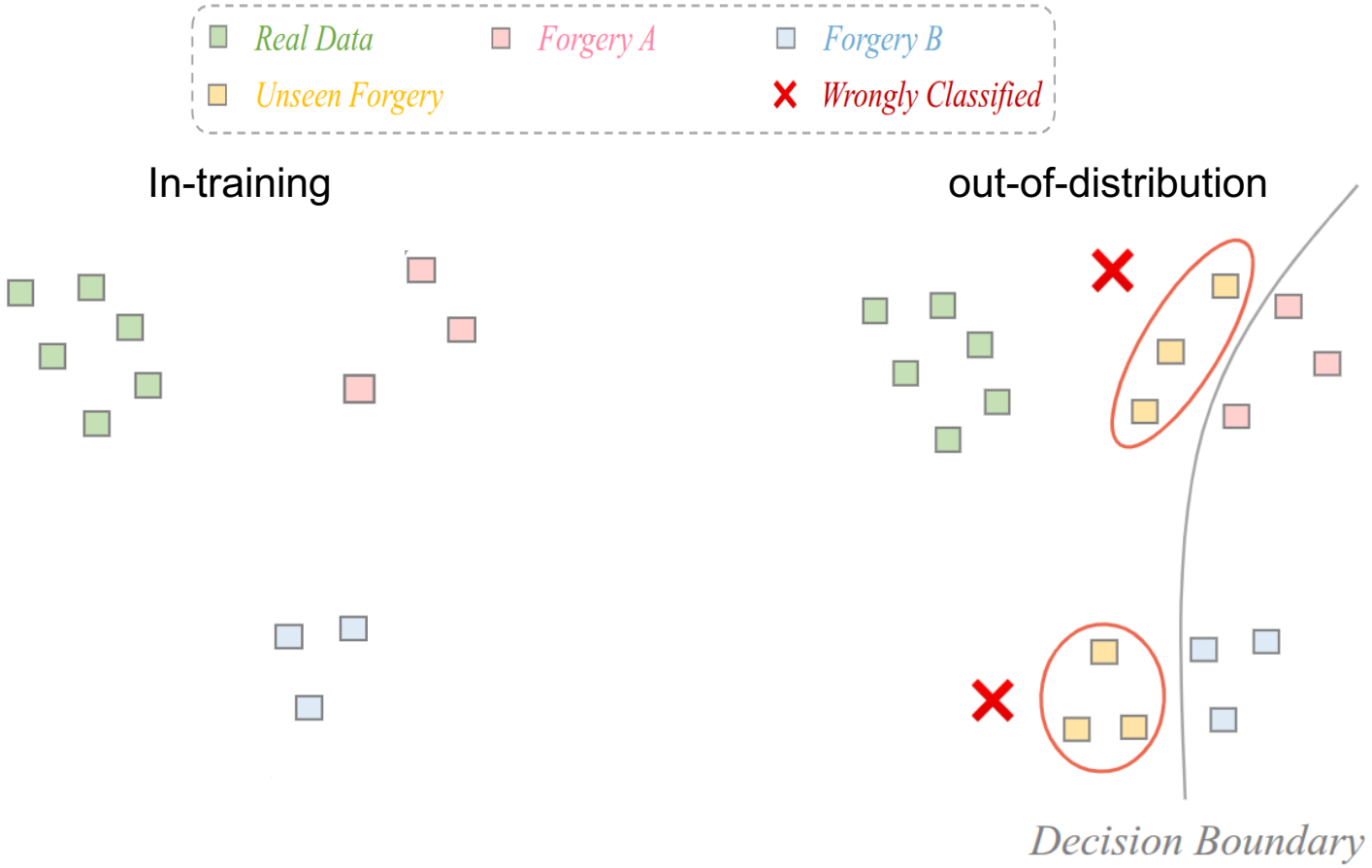
[2] Zhao et al., "Multi-attentional deepfake detection." CVPR'21.
[3]Tianyi et al., "Noise based deepfake detection via multi-head relative-interaction" AAAI'23.
[4] Guo et al.,"Constructing new backbone networks via space-frequency interactive convolution for deepfake detection." TIFS'23
[5] Haliassos et al., "Lips don't lie: A generalisable and robust approach to face forgery detection. CVPR'21
[6] Chunlei et al.,"Where deepfakes gaze at? Spatialtemporal gaze inconsistency analysis for video face forgery detection." TIFS'24
[7] Dutta et al,."WaveDIF: Wavelet sub-band based Deepfake Identification in Frequency Domain." CVPR'25
[8] Zhai et al.,"Towards generic image manipulation detection with weakly-supervised self-consistency learning" ICCV'23
[9] Zhap et al.,"Learning self-consistency for deepfake detection." ICCV'21

Outline

3. Problem Setting

Challenge Detectors suffer from lack of out-of-distribution generalization

Why? Diversity of artifacts, dataset shortcut

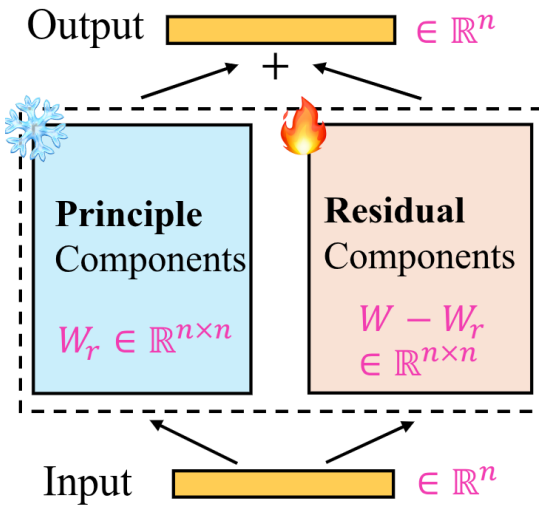


[10] Yan et al.,Transcending forgery specificity with latent space augmentation for generalizable deepfake detection." CVPR'24

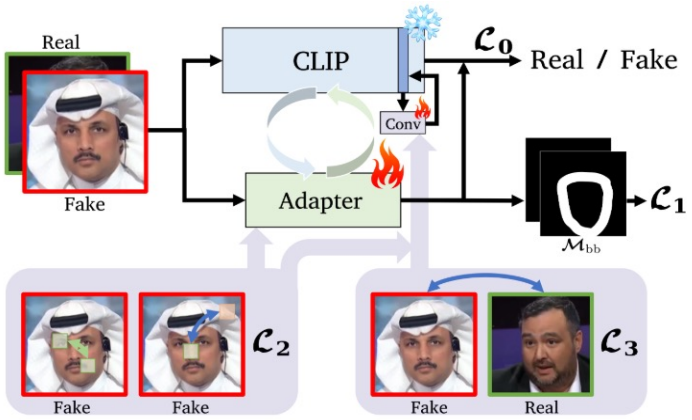
3. Prior work towards this challenge

Foundation models have strong priors and zero-shot capabilities

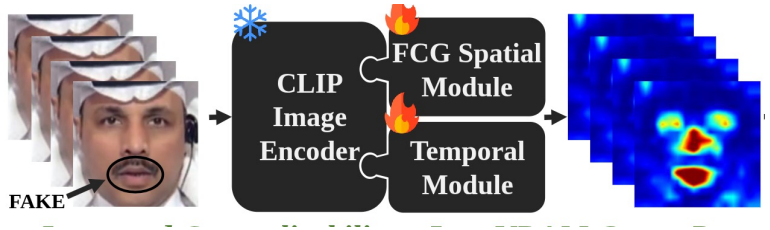
PEFT (Parameter-Efficient Fine-Tuning) preserve priors and adapt to deepfake domain



Lora based [11]



Adapters [12]



Side module [13]

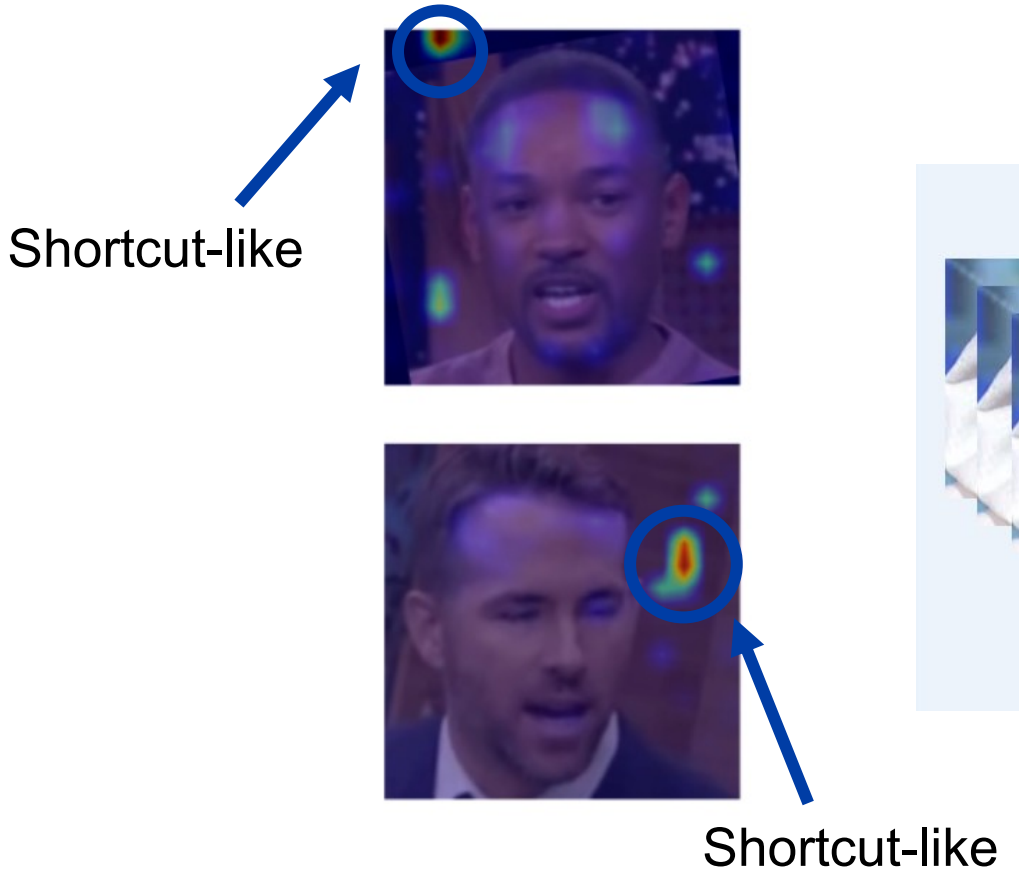
[11] Yan et al,."Orthogonal subspace decomposition for generalizable ai-generated image detection". ICML'25
[12] Cui et al.," Forensics Adapter: Adapting CLIP for Generalizable Face Forgery Detection". CVPR'25
[13] Han et al,."Towards more general video-based deepfake detection through facial component guided adaptation for foundation model". CVPR'25

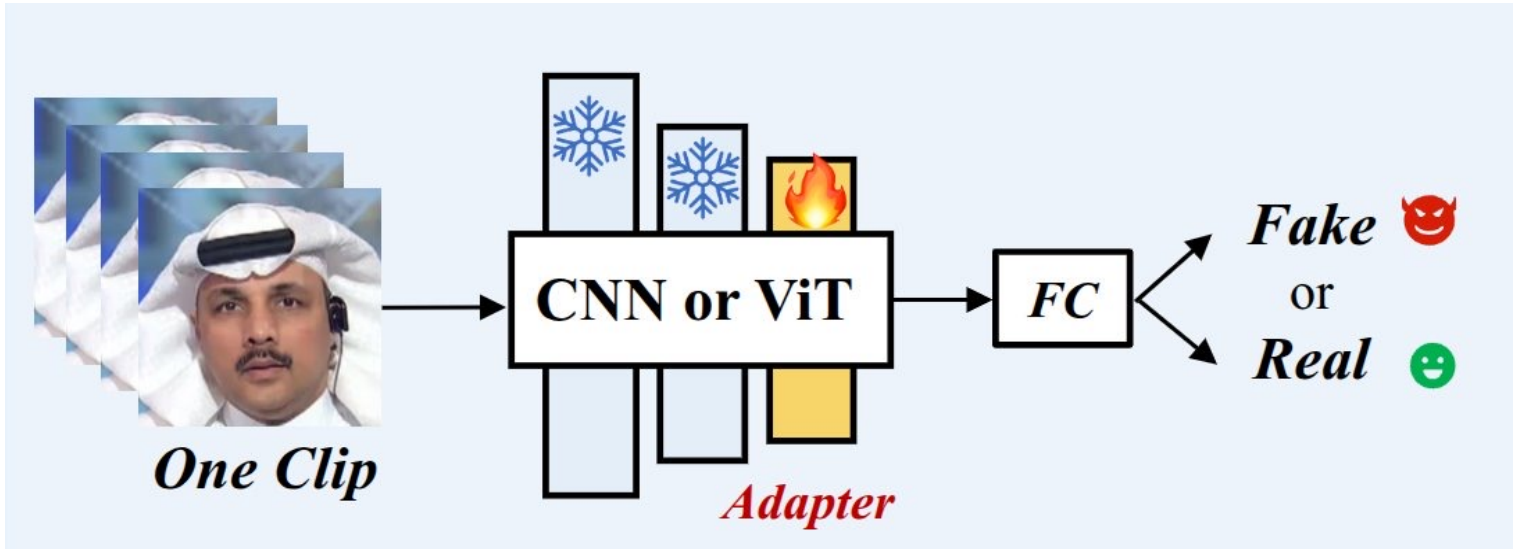
Outline

4. Method

Under binary objective, PEFT may rely on easily exploitable cues

=> can we control what the adapter is allowed to rely on and avoid shortcuts?



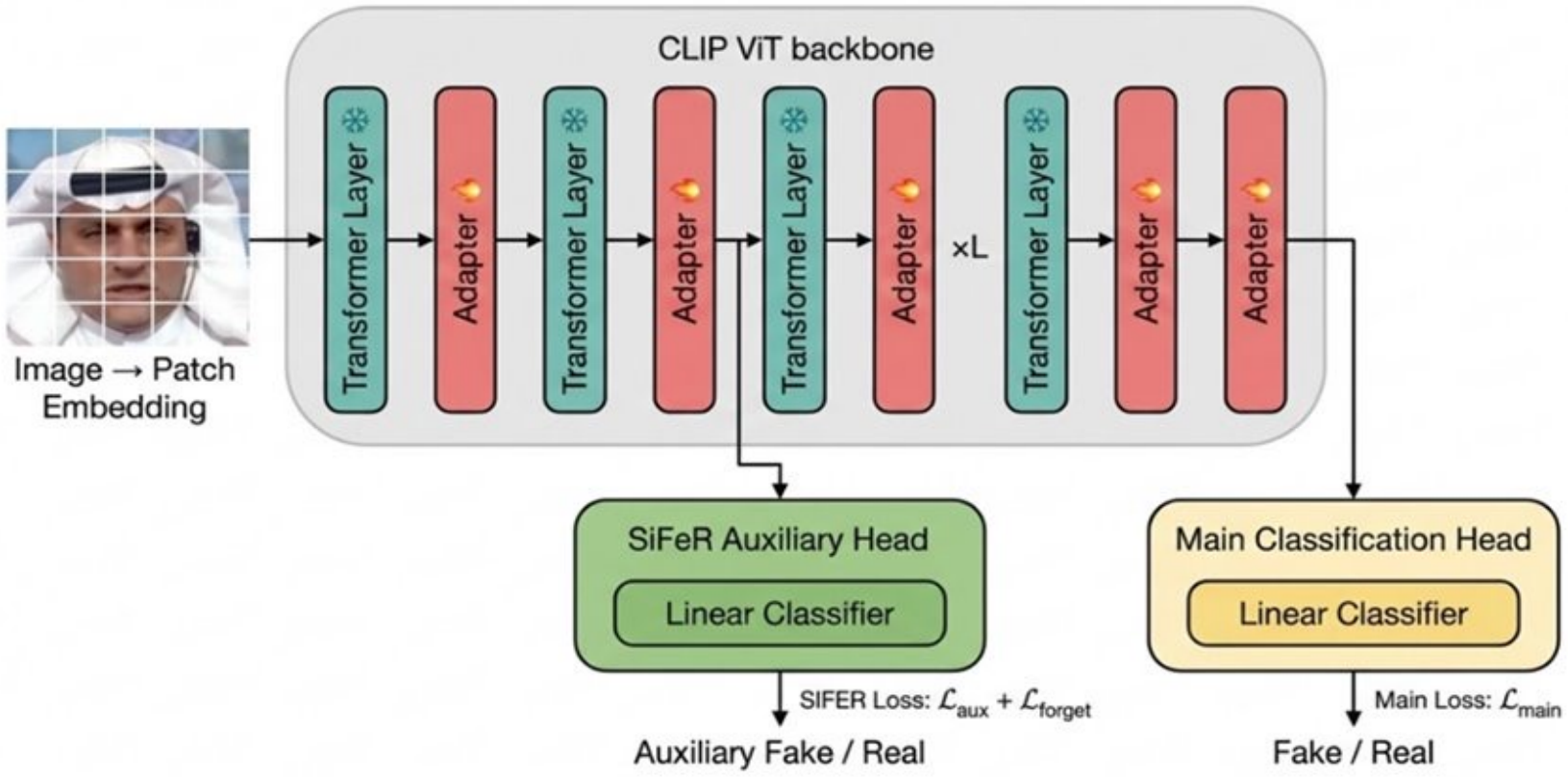


4. Method

1 Freeze backbone

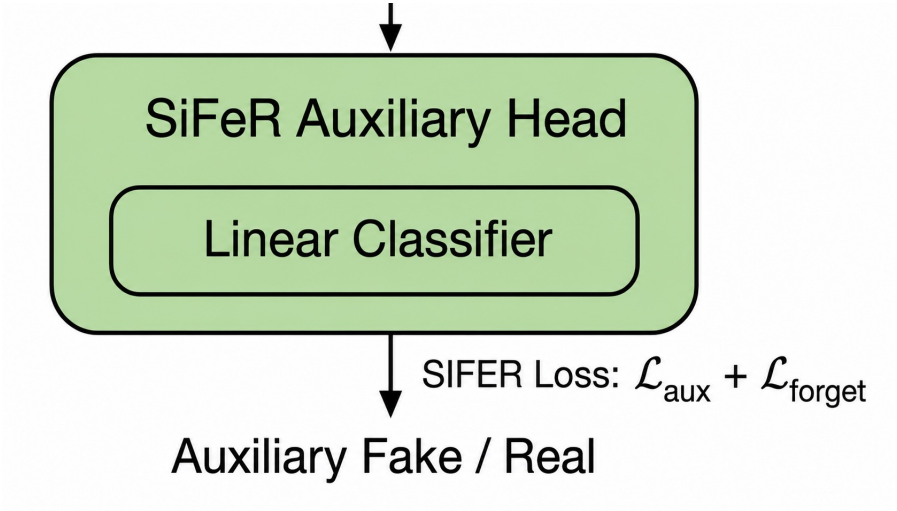
2 Train adapters + heads

3 Remove aux head at inference



4. Method

We employ SIFER [15] mechanism



$$\mathcal{L}_{\text{aux}} = \text{CE}(\hat{y}_{\text{aux}}, y), \qquad \mathcal{L}_{\text{forget}} = \text{CE}\big(U, \hat{y}_{\text{aux}}\big), \quad U = \left[\tfrac{1}{2}, \tfrac{1}{2}\right].$$

$$\mathcal{L}(k) = \alpha_1 \mathcal{L}_{\text{main}} + \alpha_2 \mathcal{L}_{\text{aux}} + \mathbf{1}\big[(k \bmod F) = 0\big]\, \alpha_3 \mathcal{L}_{\text{forget}},$$

$\mathcal{L}_{\text{aux}}$ exposes easily exploitable intermediate directions
$\mathcal{L}_{\text{forget}}$ forget reduces their discriminative utility

[15] Tiwari et al.,"Overcoming Simplicity Bias in Deep Networks using Feature Sieve". ICML'23

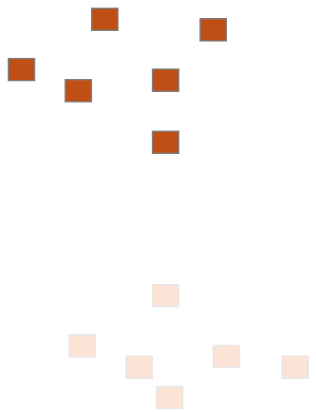
Outline

5. Experimental setup

Training set: FaceForensics++ [16]
Evaluation metric: Video-AUC
Two different evaluation protocols

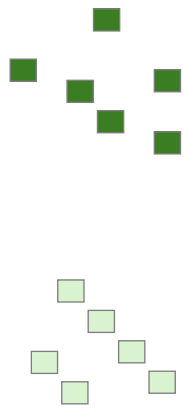
Protocol-1: Cross-dataset [17]

- Unseen Real
- Fake from unseen Real



Protocol-2: Cross-manipulation [17]

- Seen Real (FF)
- Fake from seen Real



[16] Rossler et al. ,"Faceforensics++: Learning to detect manipulated facial images." ICCV'19.
[17] Yan et al., "DF40: Toward Next-Generation Deepfake Detection." NeurIPS'24

5. Results: Protocol-1

Improved generalization on cross-dataset (CDF-v2, CDF-v3, and DFD)

Detector	Backbone	CDF-v2	CDF-v3	DFD	DFDC	DFDCP
SBI[18]	EFNB4	93.2	–	82.7	72.4	–
UCF[19]	Xception	83.7	72†	86.7	74.2	77
LSDA[10]	EFNB4	87.5	–	88.1	70.1	81.2
ProDet[20]	EFNB4	92.6	–	90.1	70.7	82.8
CDFA[21]	SwinTransformerV2	93.8	–	95.4	83.0	88.1
UDD[22]	CLIP ViT-B/16	93.1	–	95.5	81.2	88.1
P&P[23]	CLIP ViT-L/14	94.7	–	96.5	84.3	
ForADA[12]	CLIP ViT-L/14	95.7	75.6†	97.2	**87.2**	**92.9**
DFD-FCG[13]	CLIP ViT-L/14	95	84†	–	81.8	–
GenD[24]	CLIP ViT-L/14	96.0	85.9	97.0	87.1	–
EFFORT[11]	CLIP ViT-L/14	95.6	78.7	96.5	84.3	90.9
Ours	CLIP ViT-L/14	**96.3**	**87.7**	**97.6**	84.8	88.4

[18] Shiohara et al,. "Detecting deepfakes with self-blended images." CVPR'22.
[19]Yan et al,."UCF: Uncovering Common Features for Generalizable Deepfake Detection". CVPR'23
[20] Cheng et al., "Can We Leave Deepfake Data Behind in Training Deepfake Detector?" NeurIPS'24
[21] Lin et al. "Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection." ECCV'24.
[22] Fu et al.,"Exploring unbiased deepfake detection via token-level shuffling and mixing." AAAI'25
[23] Yan et al.,"Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning."CVPR'25
[24] Yermakov et al.,"Deepfake detection that generalizes across benchmarks." WACV'26

5. Results: Protocol-2

Improved generalization on cross manipulation

Method	UniFace	BleFace	MobSwap	FaceDan	FSGAN	InSwap	SimSwap	Avg.
F3Net[25]	80.9	80.8	86.7	71.7	84.5	75.7	67.4	78.2
SPSL[26]	74.7	74.8	88.5	66.6	81.2	64.3	66.5	73.8
RECCE[27]	89.8	83.2	92.5	84.8	94.9	84.8	76.8	86.7
SLADD[28]	87.8	88.2	95.4	82.5	94.3	87.9	79.4	87.9
SBI[18]	72.4	89.1	95.2	59.4	80.3	71.2	70.1	76.8
UCF[19]	83.7	83.1	82.7	73.1	86.2	93.7	80.9	83.3
IID[29]	83.9	78.9	88.8	84.4	92.7	78.9	64.4	81.7
LSDA[10]	87.2	87.5	93.0	72.1	93.9	85.5	79.3	85.5
ProDet[20]	90.8	92.9	97.5	74.7	92.8	83.7	84.4	88.1
CFDA[21]	76.2	75.6	82.3	80.3	94.2	77.2	75.7	80.2
EFFORT[11]	96.2	87.3	95.3	92.6	95.7	93.6	92.6	93.3
ForADA[12]	95.1	87.9	97.9	94.6	97.5	94.8	91.3	94.2
Ours	**98.0**	**94.4**	**98.3**	**97.4**	**98.8**	**97.1**	**95.0**	**97.0**

]25] Qian et al., "Thinking in frequency: Face forgery detection by mining frequency-aware clues." ECCV'20
[26] Liu et al., "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain."CVPR'21
[27] Caeo et al., "End-to-end reconstruction-classification learning for face forgery detection." CVPR'22
[28] Chen et al., "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection." CVPR'22
[29] Huang et al.,"Implicit identity driven deepfake face swapping detection." CVPR'23

5. Results

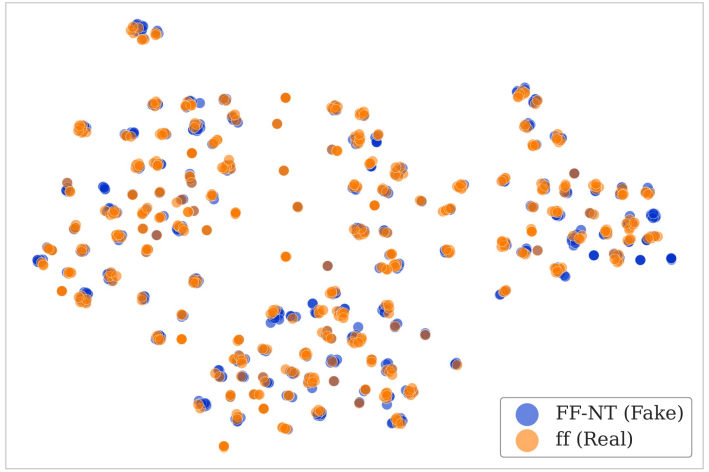
Attention shifts toward face-relevant regions

W\o SIFER

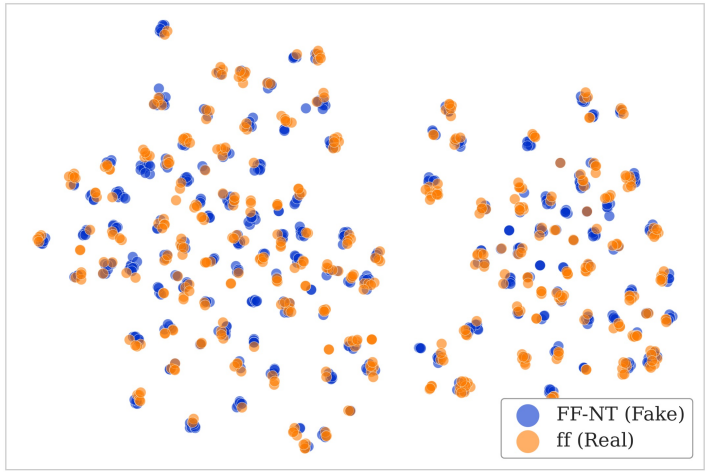
Ours

5. Results

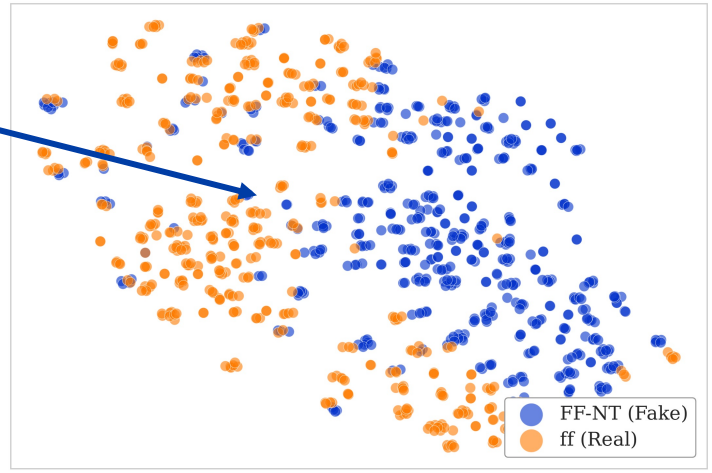
Improved separation between real and face reenactment



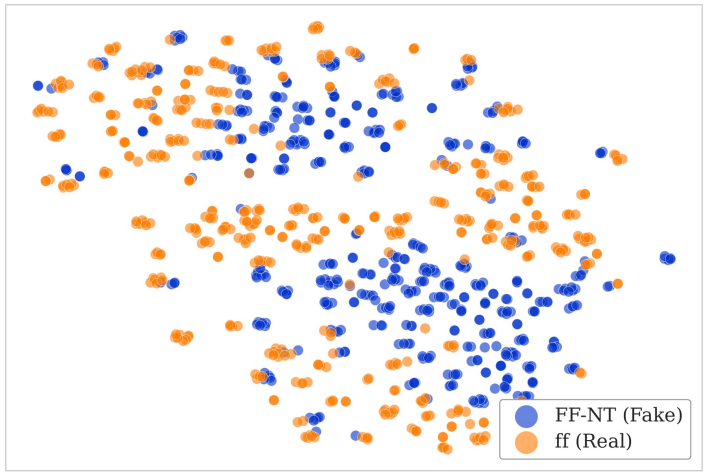
CLIP + LP



FOR-ADA



OURS

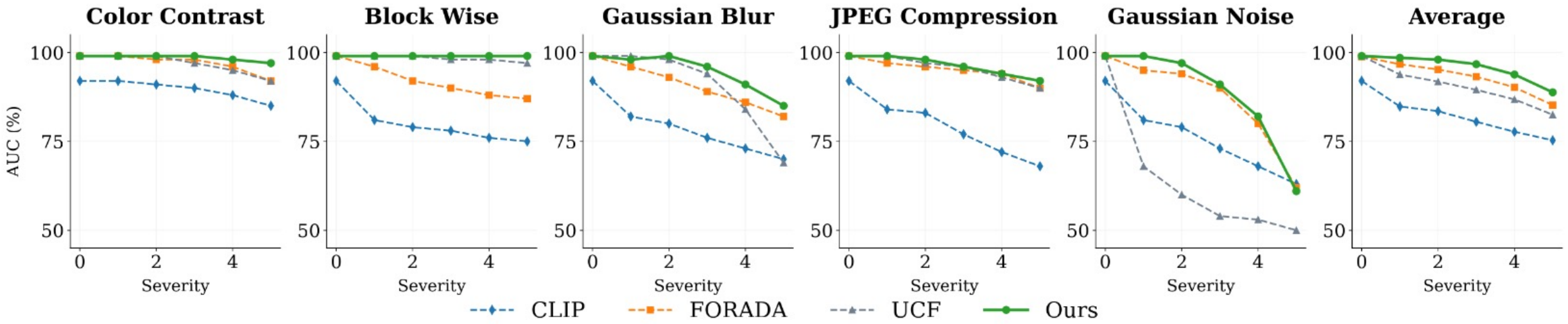


CLIP + Adapater (w/o SIFER)

Method	Wav2Lip	LIA	MCNet	FaceVid2Vid	DaNet	Avg.
UCF [44]	71.0	87.2	67.9	75.2	59.7	72.2
EFFORT [42]	64.6	79.6	64.0	71.5	62.0	68.3
ForADA [7]	57.8	79.5	54.2	62.8	50.1	60.88
Ours	**81.3**	**93.4**	**78.9**	**87.6**	**73.3**	**82.9**

5. Robustness

Improved robustness compared to prior works under common perturbations [30]



[30] Jiang et al., "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection." CVPR'20

Outline

6. Future Work

1. Extend simplicity-bias-aware adaptation to video architectures: examining whether temporal adaptation exploit non-generalizable cues

2. Characterize the features suppressed by SIFER: using interpretability analyses to determine what shortcut is being suppressed

6. Conclusions

- We frame generalizable deepfake detection as a simplicity-bias problem
- Training strategy discourages easily predictive intermediate features during CLIP adaptation through periodic feature forgetting
- Stronger cross-dataset and cross-manipulation generalization, e.g., unseen manipulations with subtle artifacts
- The exact features being suppressed remain to be characterized

Thank you for your attention

Questions & Discussion

Charbel Yahchouchi
Inria Center at Université Côte d'Azur, Probayes

charbelyah/Simplicity-Bias-CLIP

Supplementary Material

SIFER Ablations

Layer	CDF-v2	mcnet	DFDC	fc-v2v	AVG
6	96.1	74.9	85.8	83.6	85.1
8	96.9	78.5	85.6	85.1	86.5
10	96.3	78.9	84.8	87.6	**86.9**
12	96.1	73.6	85.9	82.8	84.6
14	96.7	79.8	85.3	85.5	<u>86.8</u>

Forgetting Events	CDFv3
No SIFER	84.3
2	86.9
10	**87.7**
16	87.04

Ablations

Adapter	SIFER	CDF-v2	CDF-v3	DFDC	DFD
×	×	74.6	75.6	75.3	89.2
✓	×	94.4	84.7	84.3	94.0
✓	✓	**96.3**	**87.7**	**84.8**	**97.6**

Method	SIFER	CDF-v2	CDF-v3
Bit-Fit	×	93.8	76.0
	✓	94.4	85.2
LORA	×	94.4	77.0
	✓	94.08	81.0
Adapter	×	94.4	84.7
	✓	**96.3**	**87.7**