Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization

Shichao Dong[1*], Jin Wang[1*], Renhe Ji[1†], Jiajun Liang[1], Haoqiang Fan[1], Zheng Ge[1]

Paper tag: TUE-AM-381

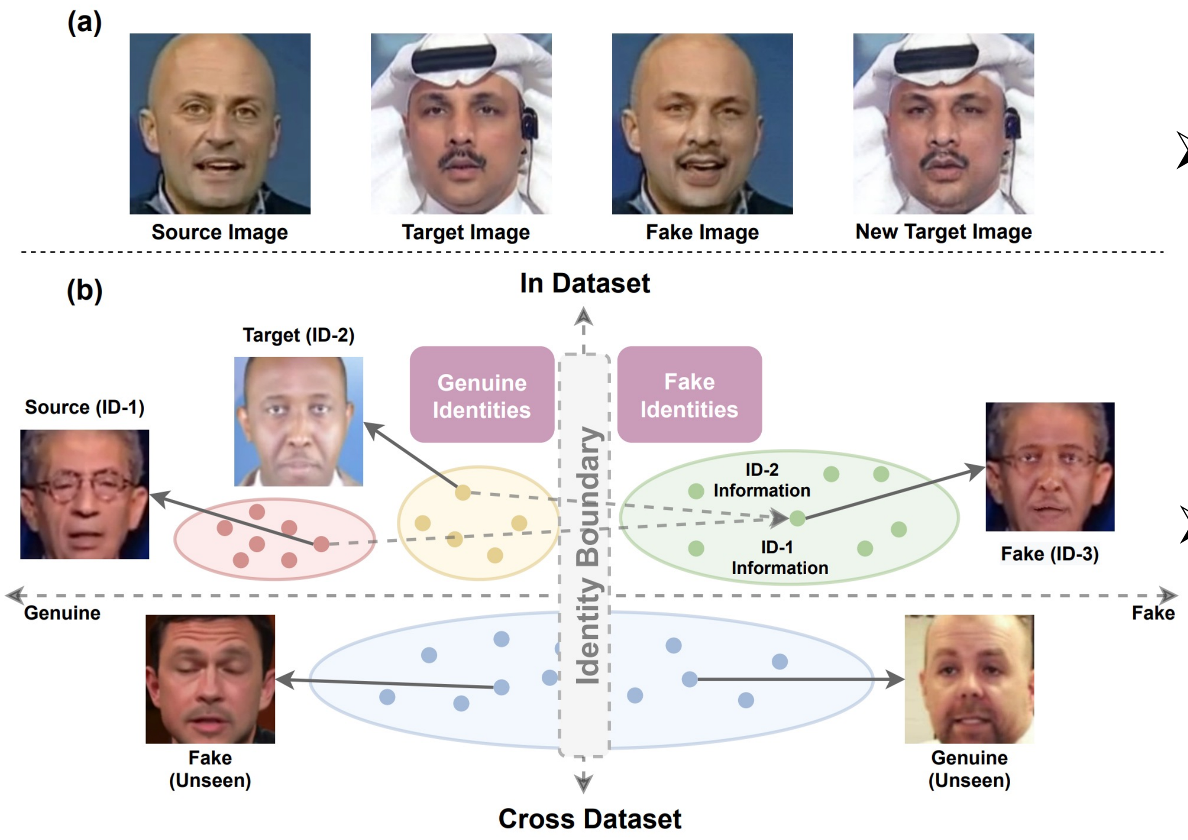
[1] MEGVII Technology

*Equal contributions †Correspondence author

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Preview

- **The Implicit Identity Leakage phenomenon**
  - ➢ The *stumbling block* for the generalization abilities of binary classifiers on deepfake detection

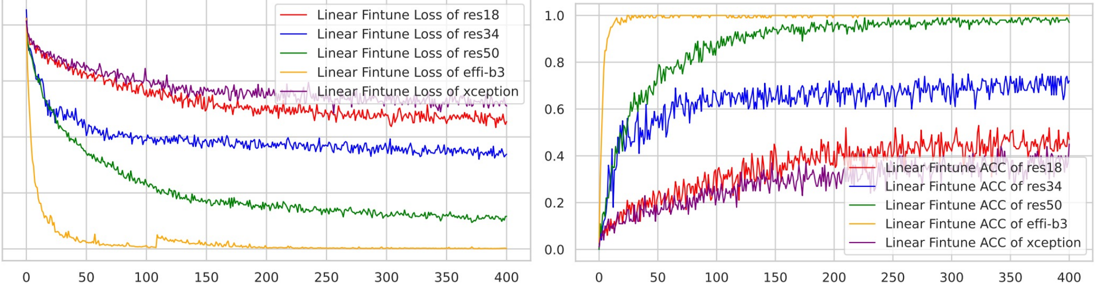


➢ There exists an implicit gap between genuine identities and fake identities in the training set, which is unintentionally captured by binary classifiers.

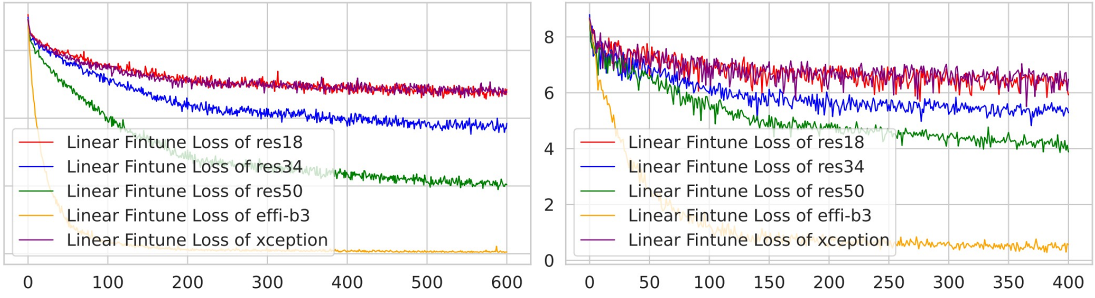
➢ Such biased representations may be mistakenly used by binary classifiers, causing false judgments when tested on the cross-dataset evaluation.

Preview

- **Verifying Implicit Identity Leakage phenomenon**
  - Existence of ID representations.



(a) Celeb-DF

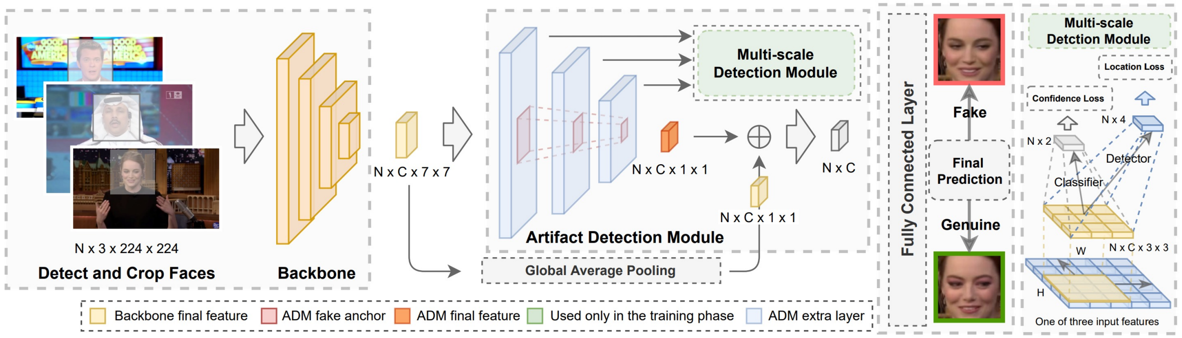


(b) FF++ (c) LFW

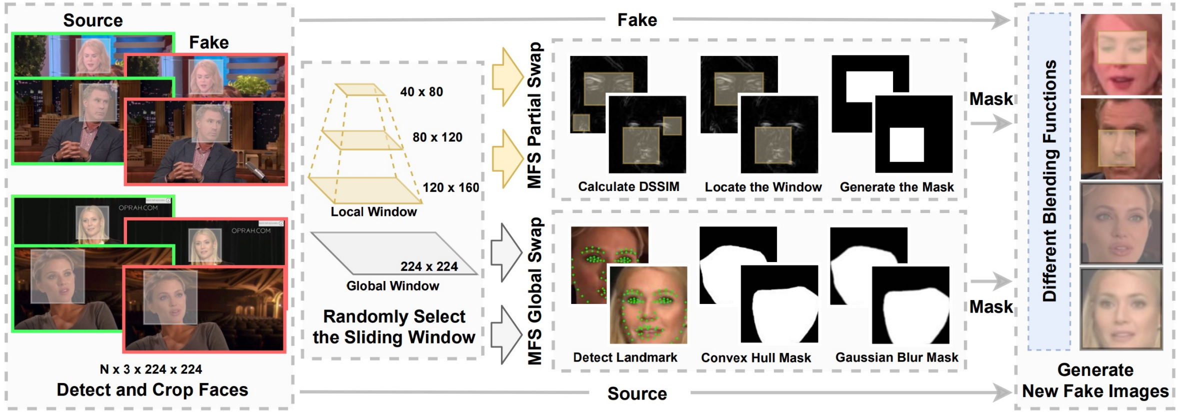
  - Influence of ID representations.

Datasets	ResNet-18	ResNet-34	ResNet-50	Xception	EfficientNet-b3
FF++	81.53	89.77	99.58	97.32	94.87
Celeb-DF	46.88	47.22	49.47	47.23	44.43

- **ID-unaware Deepfake Detection Model**
  - Artifact Detection Module



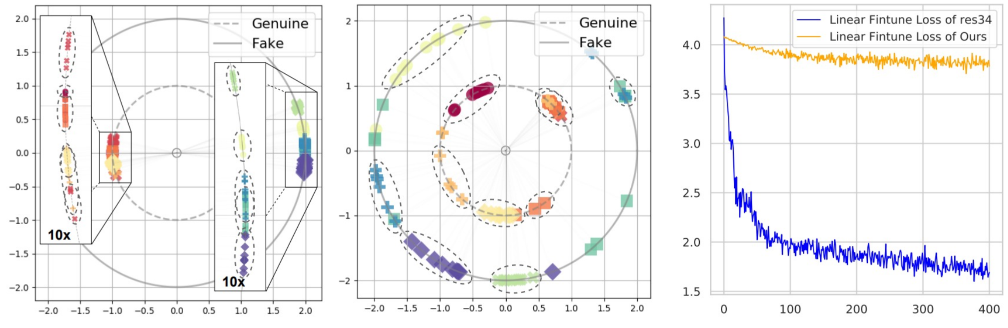
  - Multi-scale Facial Swap



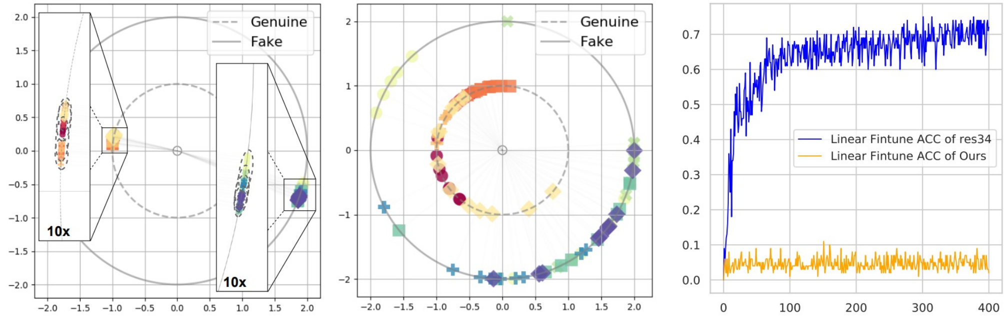
Preview

- **Experiments**

➢ Comparison of Implicit Identity Leakage



(a) Binary classifiers (FF++ (left), Celeb-DF (right)) (b) Training Loss



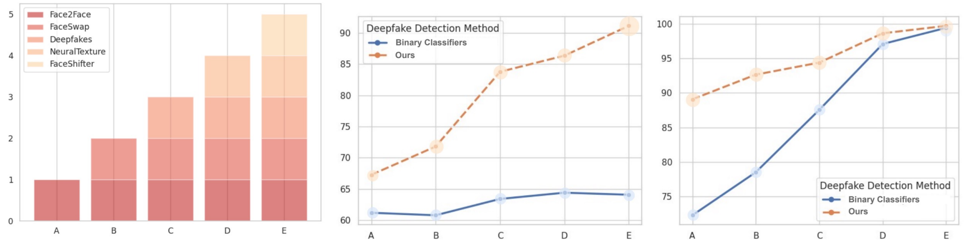
(c) Ours (FF++ (left), Celeb-DF (right)) (d) Training Accuracy

➢ Comparison with state-of-the-art methods

Models	Backbones	Test Set (AUC (%))	
		FF++	Celeb-DF
Multi-task [55]	-	76.30	54.30
Xception [67]	Xception	99.58	49.03
MMMS [81]	Transformer	99.50	65.70
SPSL [46]	Xception	96.91	76.88
Local-Relation [10]	-	-	78.26
Two-branch [52]	DenseNet	93.20	73.40
DSP-FWA [44]	ResNet-50	93.00	64.60
F^3-Net [60]	Xception	98.10	65.17
MAT [88]	Efficient-b4	99.61	68.44
SLADD [8]	Xception	98.40	79.70
Face-x-ray [41]	HRNet	99.17	80.58
PCL+I2G [89]	ResNet-34	99.11	90.03
SBI [70]	Efficient-b4	99.64	93.18
Ours	ResNet-34	**99.70**(↑0.06)	91.15
	Efficient-b3	**99.78**(↑0.14)	93.08
	Efficient-b4	**99.79**(↑0.15)	**93.88**(↑0.70)

Datasets	Xception [67]	SPSL [46]	PCL+I2G [89]	MAT [88]	SBI [70]	Ours		
						Res-34	Effi-b3	Effi-b4
DFDC-V2	45.60	66.16	67.52	70.99	72.42	71.49	**73.74**	**73.85**(↑1.43)

➢ Learning various artifact features in a data-driven scheme



(a) Sub-dataset division (b) AUC on Celeb-DF (c) AUC on FF++

Preview

THANK YOU !

Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization

Shichao Dong[1*], Jin Wang[1*], Renhe Ji[1†], Jiajun Liang[1], Haoqiang Fan[1], Zheng Ge[1]

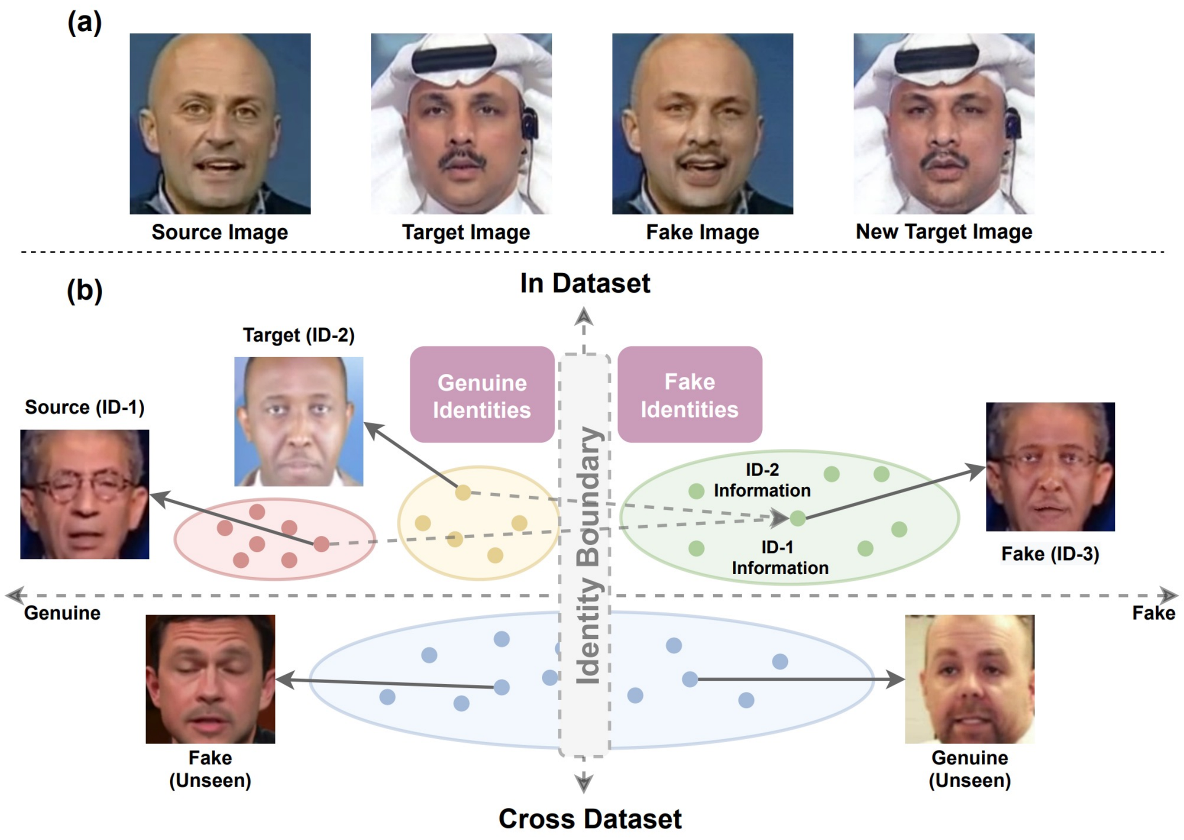
Paper tag: TUE-AM-381

[1] MEGVII Technology
*Equal contributions †Correspondence author

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- **The Implicit Identity Leakage phenomenon**
  - ➢ The *stumbling block* for the generalization abilities of binary classifiers on deepfake detection



- ➢ The identity of the fake image can not be considered as the same as its source/target image due to the information loss of identities.
- ➢ There exists an implicit gap between genuine and fake identities in the training set.
- ➢ Such biased representations may be mistakenly used by binary classifiers, causing false judgments when tested on the cross-dataset evaluation.

Objective

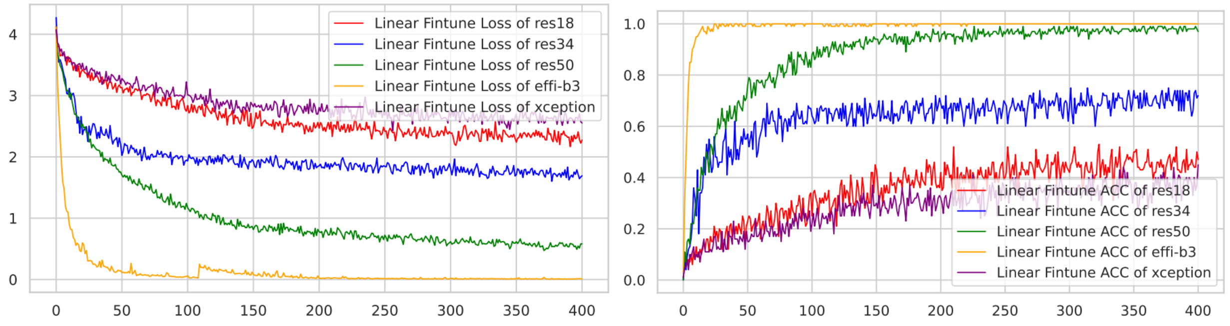
- We aim to accomplish the following two objectives for the task of deepfake detection.
  - **Verifying Implicit Identity Leakage phenomenon**
    - *Verifying the Existence of ID Representation*
    - *Quantifying the Influence of ID Representation*
  - **Improving the generalization abilities of deepfake detection models by reducing the influence of Implicit Identity Leakage phenomenon**
    - *ID-unaware Deepfake Detection Model*
    - *Multi-scale Facial Swap*

Method

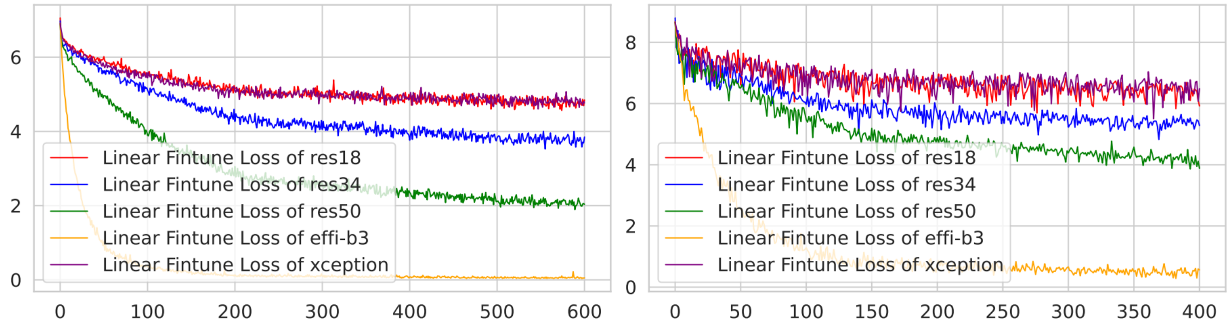
- **Verifying Implicit Identity Leakage phenomenon**

➢ *Verifying the Existence of ID Representation*

Hypothesis 1: *The ID representation in the deepfake dataset is accidentally captured by binary classifiers during the training phase when without explicit supervision.*



(a) Celeb-DF



(b) FF++ (c) LFW

➢ We measured the linear classification accuracy of identities on frozen features extracted from the classifier for FF++, Celeb-DF and a face recognition dataset LFW.

➢ Linear classification on features of different classifiers converged to varying degrees and achieved varying degrees of accuracy for identity classification

Method

- **Verifying Implicit Identity Leakage phenomenon**

  - *Quantifying the Influence of ID Representation*

> ***Hypothesis 2***: *Although the accidentally learned ID representation may enhance the performance on the in-dataset evaluation, it tends to mislead the model on the cross-dataset evaluation.*

  - We used the multivariate interaction metric [1] to quantify the influence of the ID representation

$$I([S]) = \phi\left([S] \mid N_{[S]}\right) - \sum_{i \in S} \phi\left(i \mid N_i\right)$$

  - Effect of ID representation on deepfake detection measured by AUC.

Datasets	ResNet-18	ResNet-34	ResNet-50	Xception	EfficientNet-b3
FF++	81.53	89.77	99.58	97.32	94.87
Celeb-DF	46.88	47.22	49.47	47.23	44.43

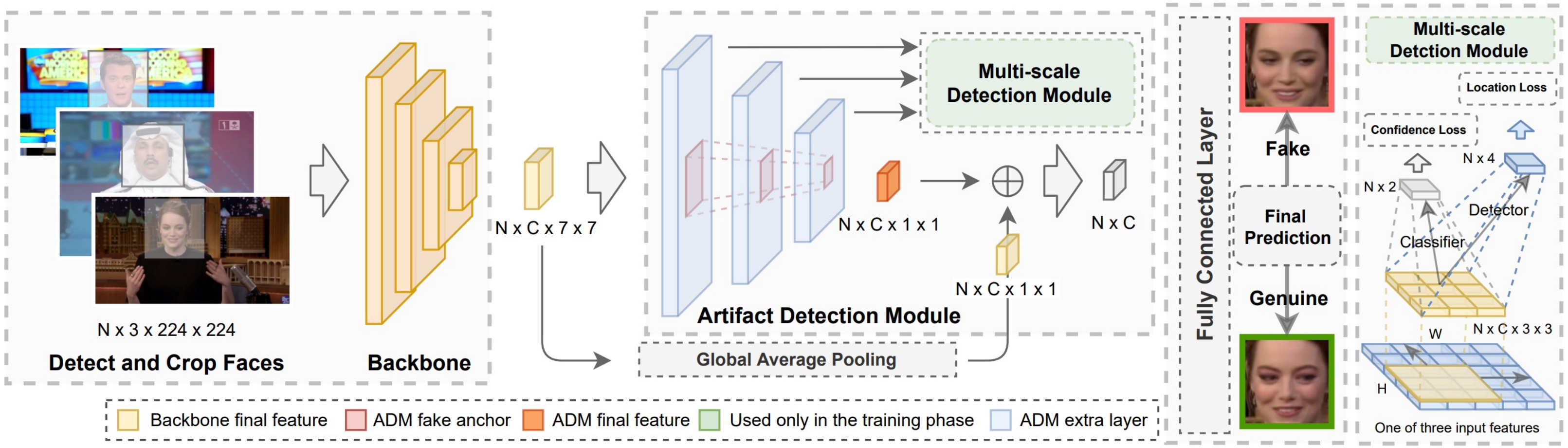
  - Such results indicate the enhancement on the in-dataset evaluations and the misguidance on the cross-dataset evaluations in terms of the influence of ID representation.

[1] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 10877–10886, 2021.

Method

- **Improving the generalization abilities of deepfake detection models by reducing the influence of Implicit Identity Leakage phenomenon**

➢ *ID-unaware Deepfake Detection Model*



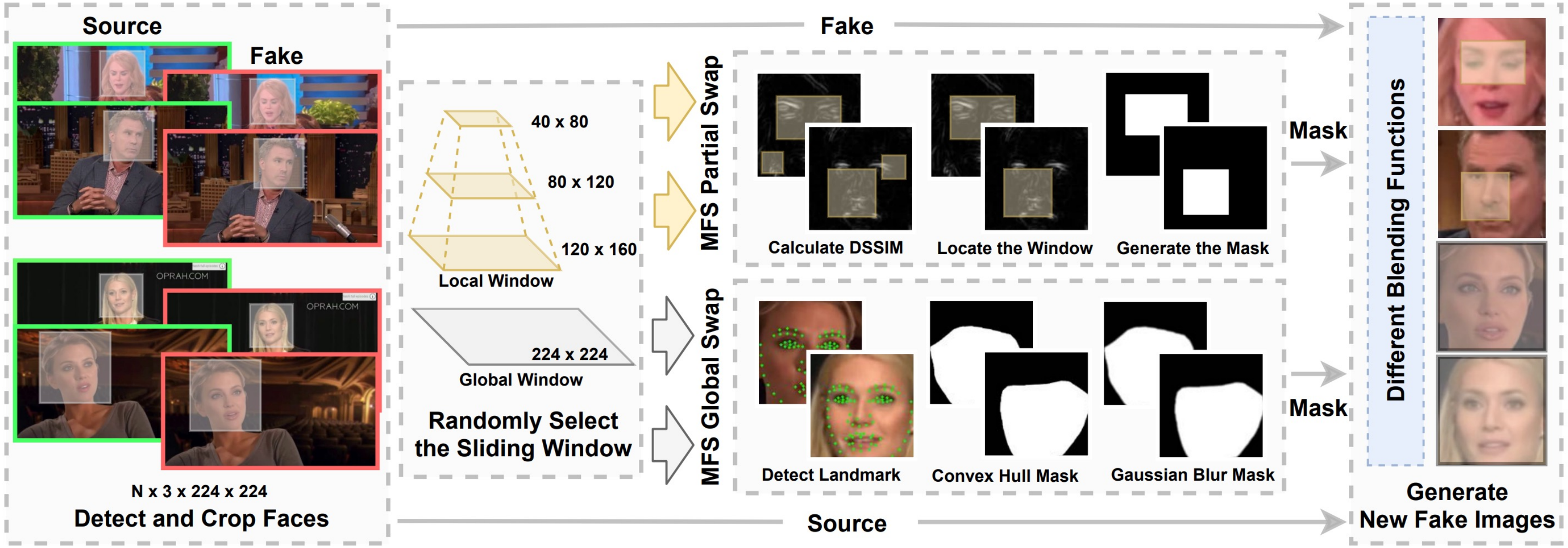
➢ Motivation: Inspired by the fact that local areas usually do not reflect the identity of images, we proposed the *ID-unaware Deepfake Detection Model* to improve the generalization ability of binary classifiers.

➢ *Artifact Detection Module* is designed to detect the position of artifact areas based on multi-scale anchors.

Method

- **Improving the generalization abilities of deepfake detection models by reducing the influence of Implicit Identity Leakage phenomenon**

➢ *Multi-scale Facial Swap*



➢ To facilitate the training of the *Artifact Detection Module*, we propose the *Multi-scale Facial Swap* method to generate fake images with the ground truth of artifact area positions.

Method

- **Improving the generalization abilities of deepfake detection models by reducing the influence of Implicit Identity Leakage phenomenon**

➢ *Loss function*

$$L = \beta L_{det} + L_{cls}$$

➢ The overall loss function is a weighted sum of the global classification loss L_{cls} and detection loss L_{det}.

➢ L_{cls} is the cross-entropy loss to measure the accuracy of the final prediction, i.e., fake or genuine images.

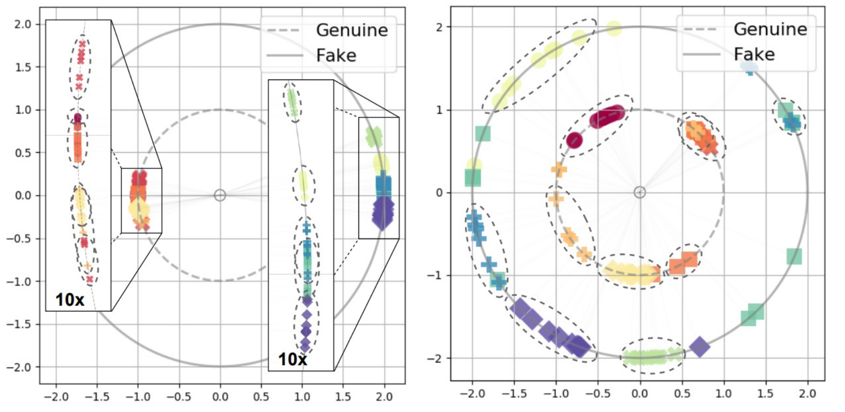
➢ L_{det} is the detection loss to guide the learning of ADM. Similar to SSD [2], it contains confidence loss (L_{conf}) and location loss (L_{loc}).

$$L_{det} = \frac{1}{N}(L_{\text{conf}}(x,c) + \alpha L_{loc}(x,l,g))$$

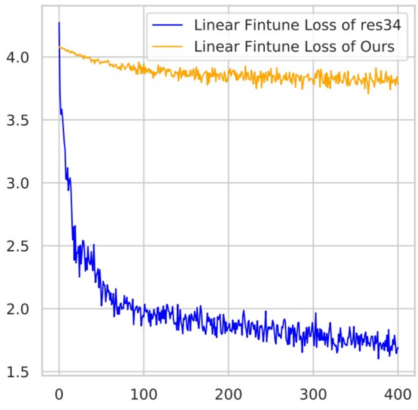
[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.

Experiments

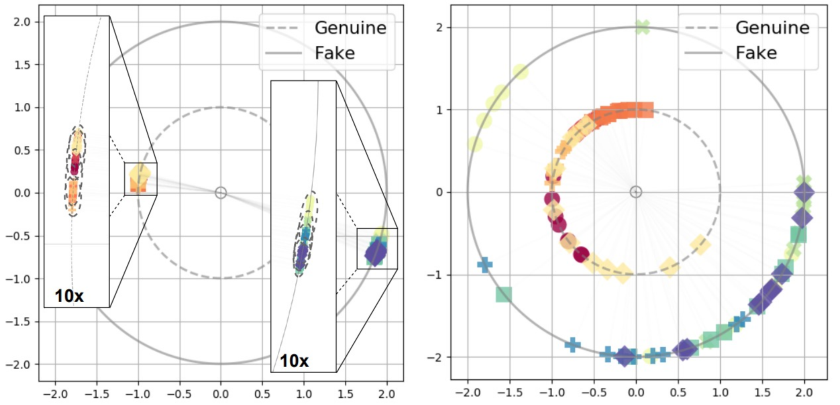
- **Comparison of Implicit Identity Leakage**



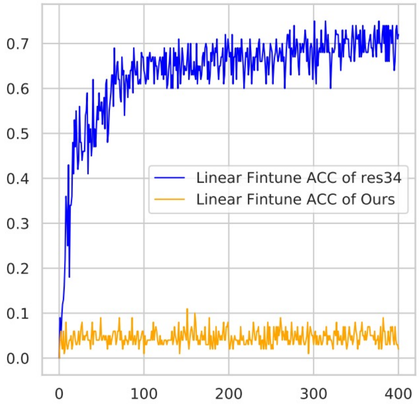
(a) Binary classifiers (FF++ (left), Celeb-DF (right))



(b) Training Loss



(c) Ours (FF++ (left), Celeb-DF (right))



(d) Training Accuracy

- In Fig (a) and (c), we used t-SNE to visualize the high dimensional features extracted from the final layer of different models in 2D.

- In Fig (b) and (d), we conduct the same ID linear classification experiment as before to compare the existence of ID representation in features of our model and the binary classifier.

Experiments

- **Comparison with state-of-the-art methods**

Models	Backbones	Test Set (AUC (%)) FF++	Test Set (AUC (%)) Celeb-DF
Multi-task [55]	-	76.30	54.30
Xception [67]	Xception	99.58	49.03
MMMS [81]	Transformer	99.50	65.70
SPSL [46]	Xception	96.91	76.88
Local-Relation [10]	-	-	78.26
Two-branch [52]	DenseNet	93.20	73.40
DSP-FWA [44]	ResNet-50	93.00	64.60
F^3-Net [60]	Xception	98.10	65.17
MAT [88]	Efficient-b4	99.61	68.44
SLADD [8]	Xception	98.40	79.70
Face-x-ray [41]	HRNet	99.17	80.58
PCL+I2G [89]	ResNet-34	99.11	90.03
SBI [70]	Efficient-b4	99.64	93.18
Ours	ResNet-34	**99.70**(↑0.06)	91.15
	Efficient-b3	**99.78**(↑0.14)	93.08
	Efficient-b4	**99.79**(↑0.15)	**93.88**(↑0.70)

Datasets	Xception [67]	SPSL [46]	PCL+I2G [89]	MAT [88]	SBI [70]	Ours Res-34	Ours Effi-b3	Ours Effi-b4
DFDC-V2	45.60	66.16	67.52	70.99	72.42	71.49	**73.74**	**73.85**(↑1.43)

- Compared with previous methods, our method significantly improved the performance on both the in-dataset and cross-dataset evaluations.
- Such results show the effectiveness of reducing the influence of Implicit Identity Leakage to learn generalized artifact features on face forgeries.

Experiments

- **Robustness evaluation**

Method	Saturation	Contrast	Block	Noise	Blur	Pixel	**Avg**
Xception [67]	99.3	98.6	99.7	53.8	60.2	74.2	81.0
Face-x-ray [41]	97.6	88.5	99.1	49.8	63.8	88.6	81.2
LipForensices [26]	**99.9**	99.6	87.4	73.8	96.1	95.6	92.1
Ours	99.6	**99.8**	**99.8**	**87.4**	**99.0**	**98.8**	**97.4**

- **Cross-method evaluation**

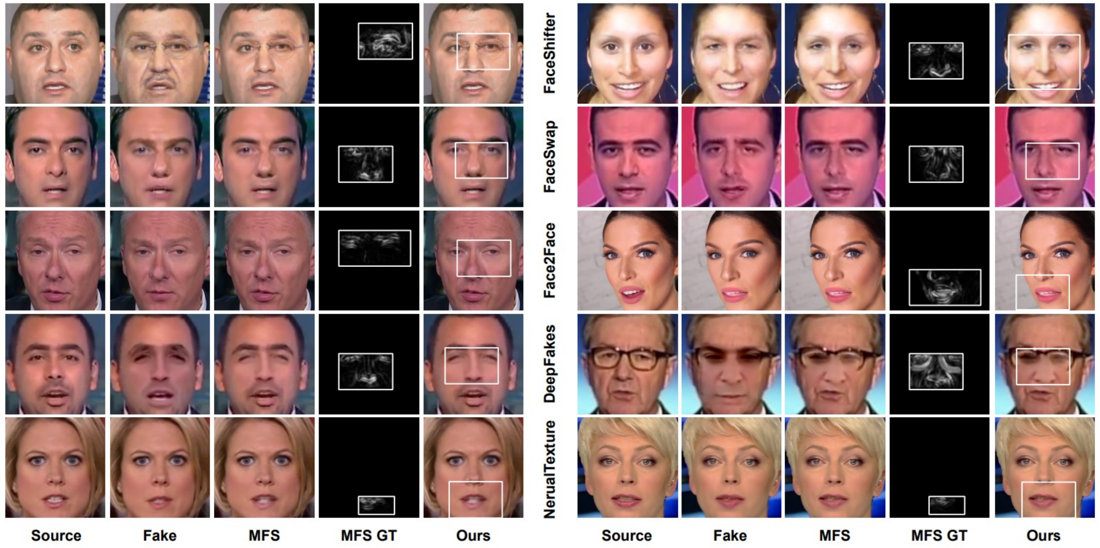
Training set	Model	DF	F2F	FS	NT	FF++
DF	Xception [67]	99.38	75.05	49.13	80.39	76.34
	Ours+Xception [67]	**100.00**	**83.94**	**58.33**	**68.98**	**77.81** (↑1.47)
F2F	Xception [67]	87.56	99.53	65.23	65.90	79.55
	Ours+Xception [67]	**99.88**	**99.97**	**79.40**	**82.38**	**90.41** (↑10.86)
FS	Xception [67]	70.12	61.70	99.36	**68.71**	74.91
	Ours+Xception [67]	**93.42**	**74.00**	**99.92**	49.86	**79.30** (↑4.39)
NT	Xception [67]	93.09	84.82	47.98	99.50	83.42
	Ours+Xception [67]	**100.00**	**97.93**	**86.76**	**99.46**	**96.04** (↑12.62)

Training set	Model	Test Set		Training set	Model	Test Set	
		FF++	DFDC-V2			DFDC-V2	FF++
FF++	ResNet-34	**99.88**	48.73	DFDC-V2	ResNet-34	92.49	60.56
	Ours+ResNet-34	99.70	**71.49**		Ours+ResNet-34	**94.85**	**77.32**
	Effi-b3	99.75	54,12		Effi-b3	94.31	60.87
	Ours+Effi-b3	**99.78**	**73.74**		Ours+Effi-b3	**95.67**	**84.43**

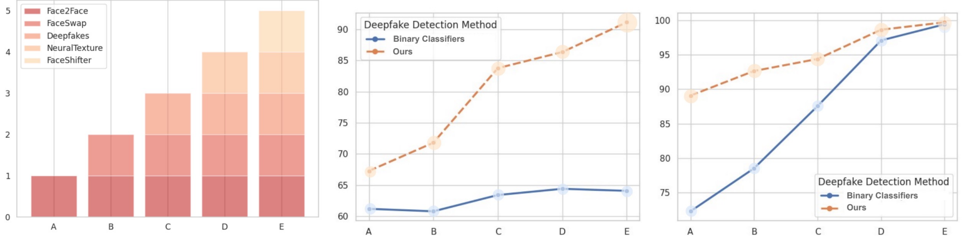
- **Potential Applicability**

Models	FF++	Celeb-DF	DFDC-V2
SBI [70]	**99.64**	93.18	72.42
Ours+SBI [70]	99.33(↓ 0.31)	**94.15**(↑ 0.97)	**79.57**(↑ 7.15)

- **Visual results**



- **Learning various artifacts**



(a) Sub-dataset division (b) AUC on Celeb-DF (c) AUC on FF++

➢ Our method can automatically learn various artifact features in a data-driven scheme.

Contribution

In this paper, we discover the phenomenon termed as *Implicit Identity Leakage* through experimental verification: the deepfake detection model is sensitive to the identity information of the data, which reduces the model generalization ability on unseen datasets. To this end, we propose *ID-unaware Deepfake Detection Model* to alleviate the Implicit Identity Leakage phenomenon. Extensive experiments demonstrate that by reducing the influence of *Implicit Identity Leakage*, our model successfully learns generalized artifact features and outperforms the state-of-the-art methods. In summary, this research provides a new perspective to understand the generalization of deepfake detection models, which sheds new light on the development of the field.

MEGVII 旷视

THANK YOU !

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA