VCIP 2024

FRIDAY: Mitigating Unintentional Facial Identity in Deepfake Detectors Guided by Facial Recognizers

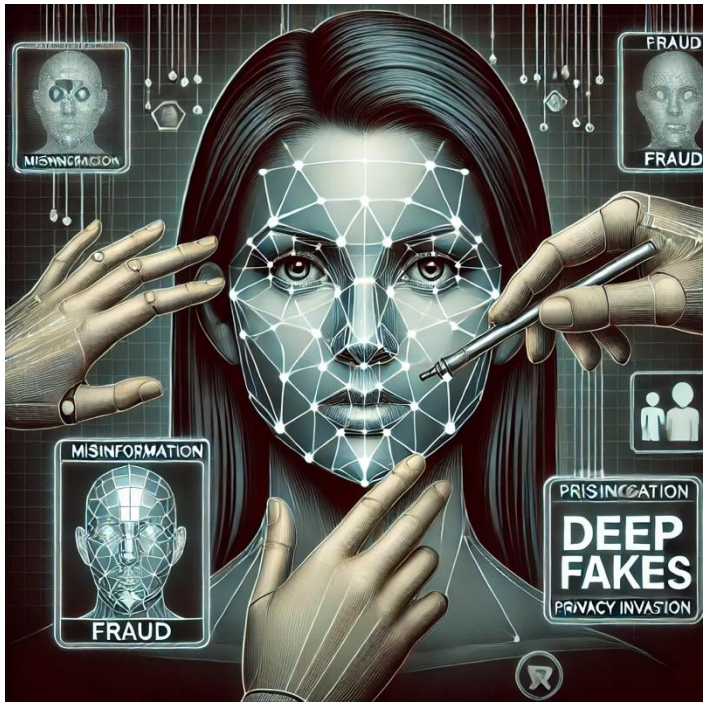
Younghun Kim, Myung-Joon Kwon, Wonjun Lee, and Changick Kim

Korea Advanced Institute of Science and Technology

P-ID 178

Deepfake Detection: Growing Need and Challenges

- **Growing Need**: Deepfake crimes, like fraud, spreading false information, and creating harmful videos of individuals, are **increasing quickly**.

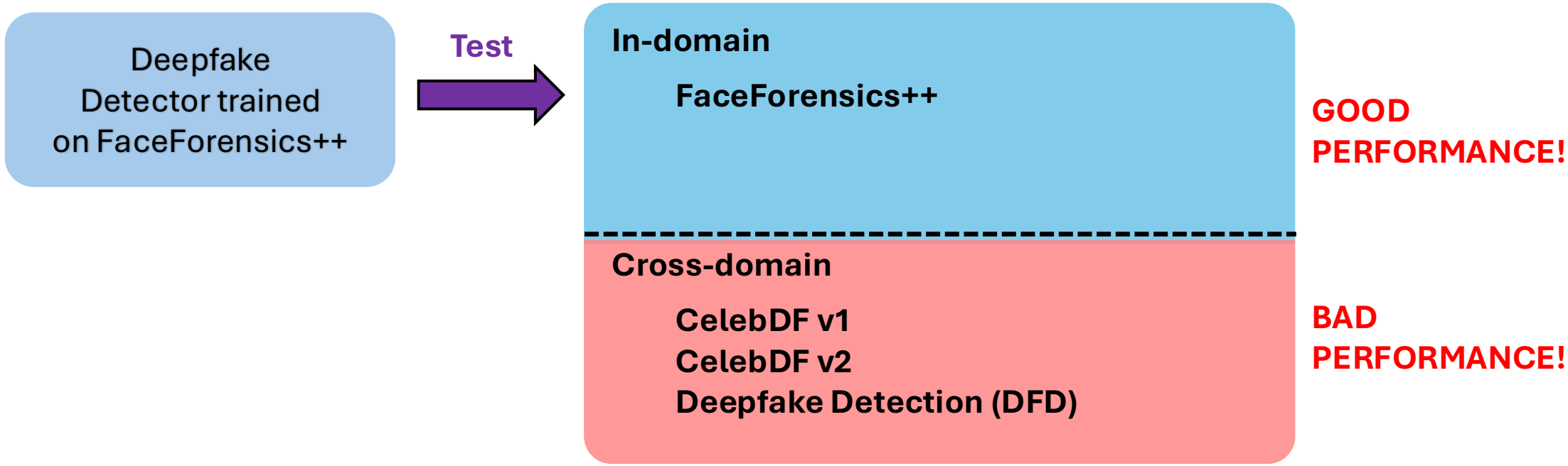


Deepfake Detection: Growing Need and Challenges

- **Challenges**: There are many types of deepfake generators, each producing content with varying levels of quality and unique artifacts. Given this diversity, **building robust deepfake detectors** that can accurately identify deepfakes from new, previously unseen generators is essential.

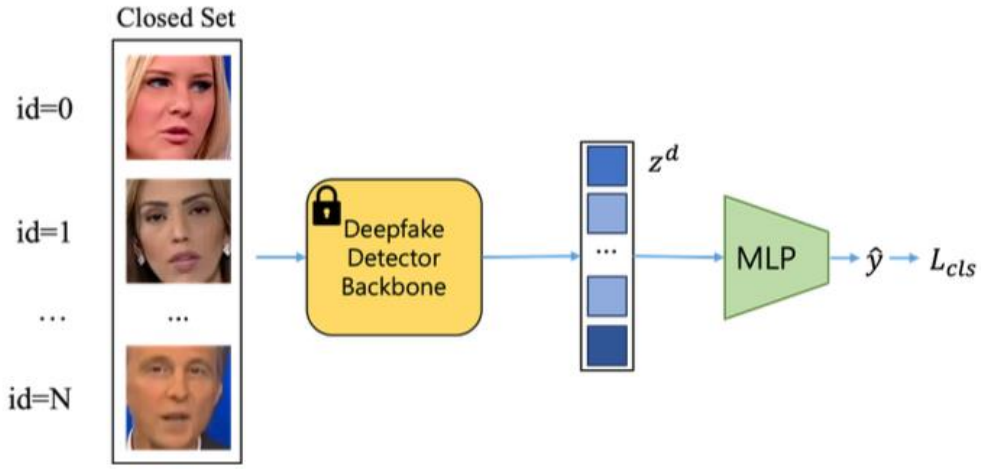
Limitations of Current Deepfake Detection Methods

- **Lack of Robustness**: Existing deep learning models for computer vision perform well on the datasets they're trained on (in-domain) **but struggle with new, unseen datasets (cross-domain)**.

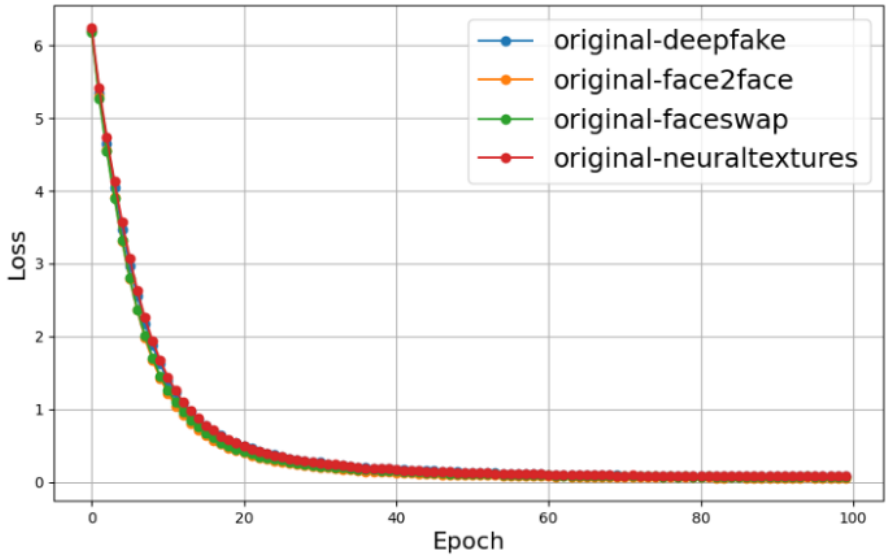


Causes of Lack of Robustness

- **Unintentional Facial Identity**: One emerging insight is that, instead of focusing on detecting artifacts, deepfake detector models **unintentionally learn irrelevant information, like facial identity,** during training.



(a) Unintentional identity learning check method.



(b) Epoch Loss.

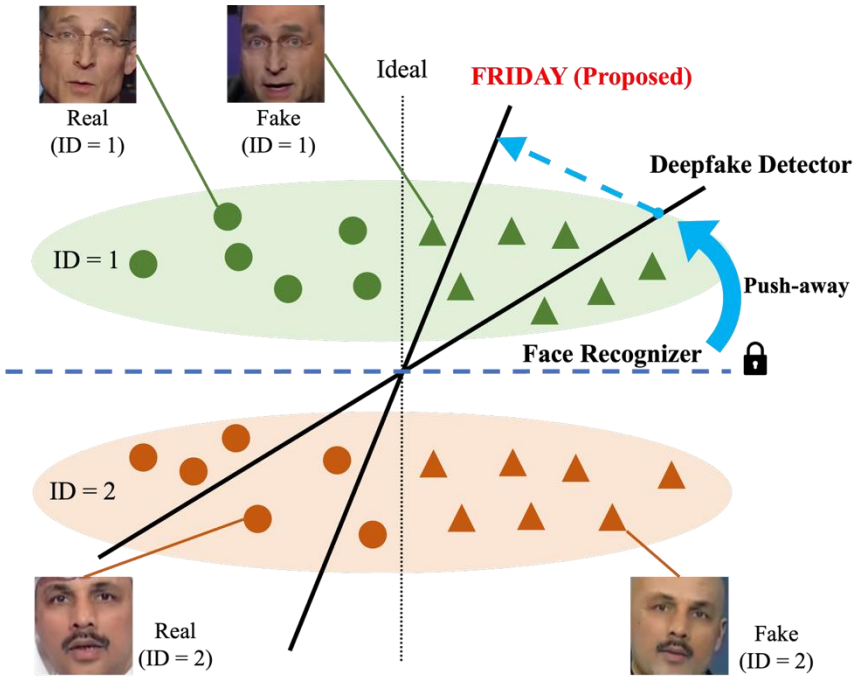
Objective of FRIDAY (**F**ace **R**ecognizer **ID**-**A**ttenuating Methodolog**Y**)

- **Reducing Unintentional Identity Learning**: **Prevent the model from learning irrelevant facial identity information**, focusing instead on artifact-based features critical for detecting deepfakes.

- **Enhancing Cross-Domain Performance**: Improve the generalization of deepfake detectors across different datasets **by minimizing reliance on facial identity features**.

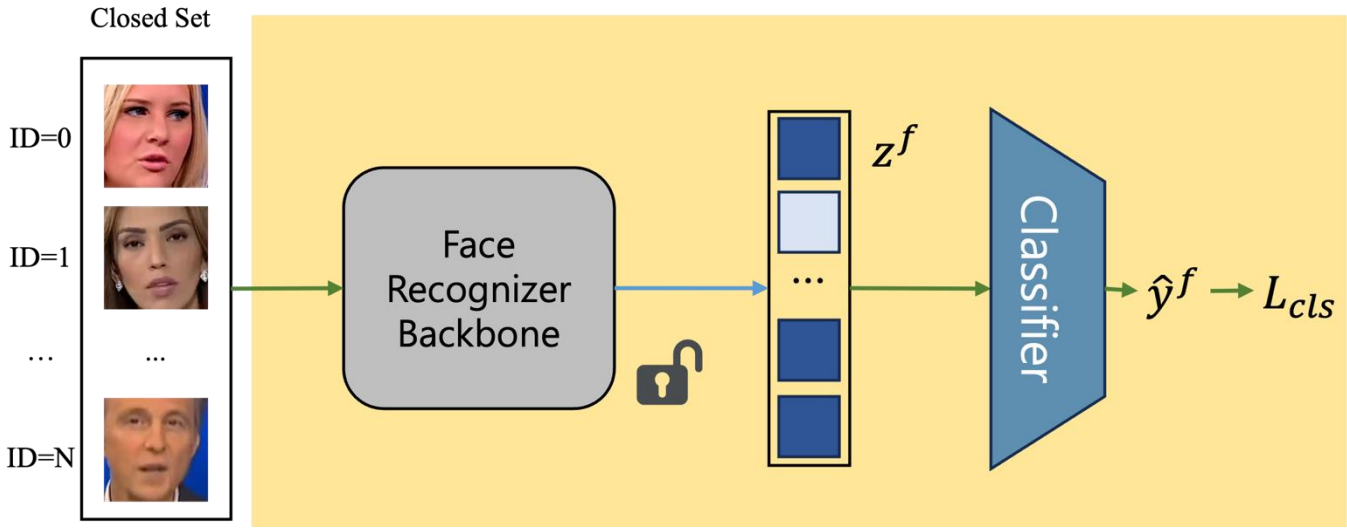
Overview of FRIDAY (**F**ace **R**ecognizer **ID**-**A**ttenuating Methodolog**Y**)

- **Intuition**: The Face Recognizer inherently captures rich facial identity features.
- **FRIDAY (ours)**: We leverages the Face Recognizer to guide the **deepfake detector away from learning identity-related information**, ensuring it focuses on detecting deepfake artifacts instead.



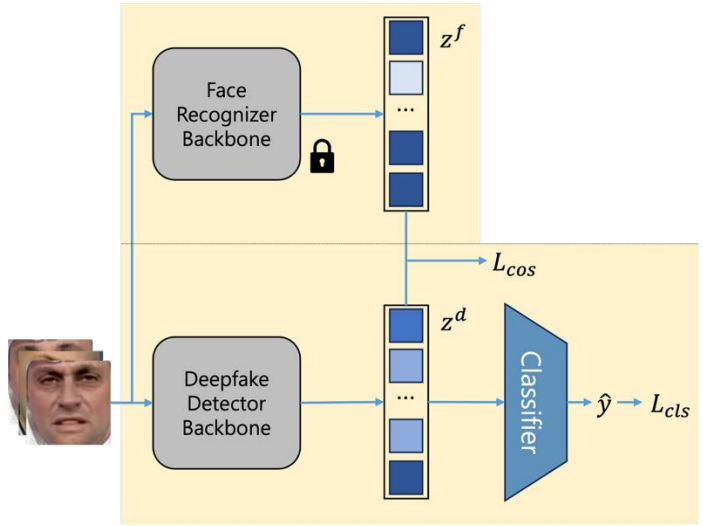
Phase 1

- **Face Recognizer Training**: We train **a face recognizer to distinguish between a total of N individuals**, with the intention of leveraging it later in the deepfake detector training process.



Phase 2

- **Deepfake Detector Training**: In Phase 2, we freeze the face recognizer trained in Phase 1 and use it during the deepfake detector training. When an input is processed, we **encourage the features from the face recognizer and the deepfake detector to diverge**, ensuring they focus on different aspects of the image.



$$L_{\text{fia}} = \left| \frac{z^f \cdot z^d}{\|z^f\|_2 \|z^d\|_2} \right|$$

$$L_{\text{cls}} = -[y \log(\hat{y}_d) + (1 - y) \log(1 - \hat{y}_d)]$$

$$L_{\text{total}} = L_{\text{cls}} + \lambda \cdot L_{\text{fia}}$$

Experiments Settings

- **Datasets**
  - FaceForensics++ [In-domain]
  - Celeb-DF v1 & v2 [Cross-domain]
  - DeepfakeDetection (DFD) [Cross-domain]

In Domain (Train & Test)

FaceForensics++

Cross-Domain (Test only)

CelebDF v1 & v2

DeepfakeDetection (DFD)

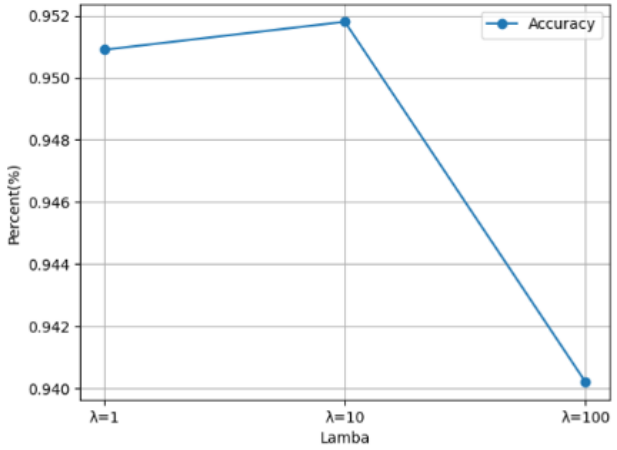
Performance Comparison

TABLE I: Performance Comparison Across Different Test Datasets (%)

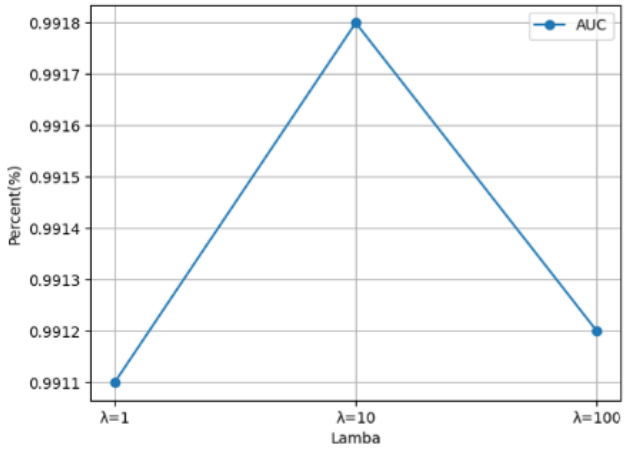
Model	Test Datasets									
	In-domain		Cross-domain							
	FF++		Celeb-DF V1		Celeb-DF V2		DFD		Average	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
CapsuleNet [7]	89.55	96.92	63.00	69.78	70.66	74.41	67.71	71.67	67.12	71.95
Xception [5]	<u>94.47</u>	<u>98.49</u>	64.00	68.21	71.24	72.82	78.20	76.38	71.15	72.47
CViT [15]	93.84	98.26	**74.00**	<u>82.98</u>	74.90	79.60	77.41	78.93	75.44	80.50
UIA-ViT [9]	93.57	98.46	**74.00**	82.32	<u>75.28</u>	<u>80.22</u>	<u>87.41</u>	**86.14**	<u>78.90</u>	<u>82.89</u>
FRIDAY ($\lambda = 10$) (Ours)	**95.18**	**99.18**	**74.00**	**85.27**	**76.25**	**83.88**	**90.12**	<u>83.95</u>	**80.12**	**84.37**

- FRIDAY outperforms recent deepfake detection models in accuracy and robustness.

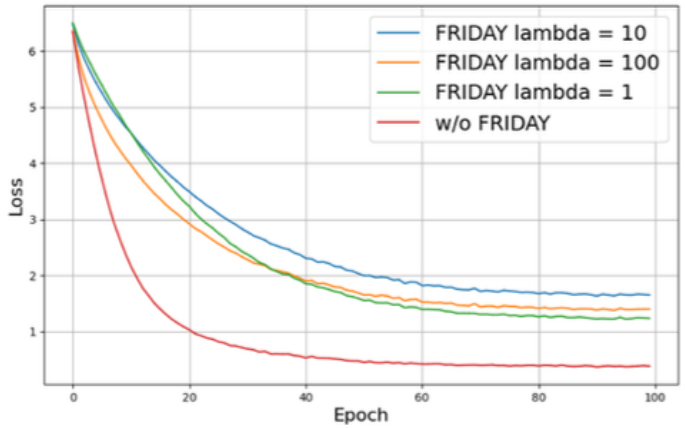
Ablation Study



(a) ACC changes along with λ



(b) Lambda AUC changes



(c) FRIDAY effectiveness

- Setting **Lambda to 10** gives the best results.

- **Effectiveness**: Increasing Lambda in the Facial Identity Attenuating loss **reduces the model's focus on facial identity and increase the performance**.

Conclusion

- **Introduction of FRIDAY**: We proposed FRIDAY, a new method aimed at **improving the generalization of deepfake detectors** by reducing reliance on facial identity features.

- **Use of Facial Identity Attenuating (FIA) Loss**: Leveraged a pre-trained face recognizer to apply FIA loss, encouraging the model to **focus on deepfake-specific artifacts instead of identity-related cues**.

- **Demonstrated Effectiveness in Ablation Study**: The ablation study **shows that FRIDAY effectively reduces facial identity reliance**, validating the impact of our approach in minimizing identity-related features within the detector.

- **Performance Improvements**: Demonstrated that FRIDAY outperforms existing models in both in-domain and cross-domain settings, showing robust results across various datasets.

Thank You for your attention!

CILAB

Email: younghun1664@kaist.ac.kr