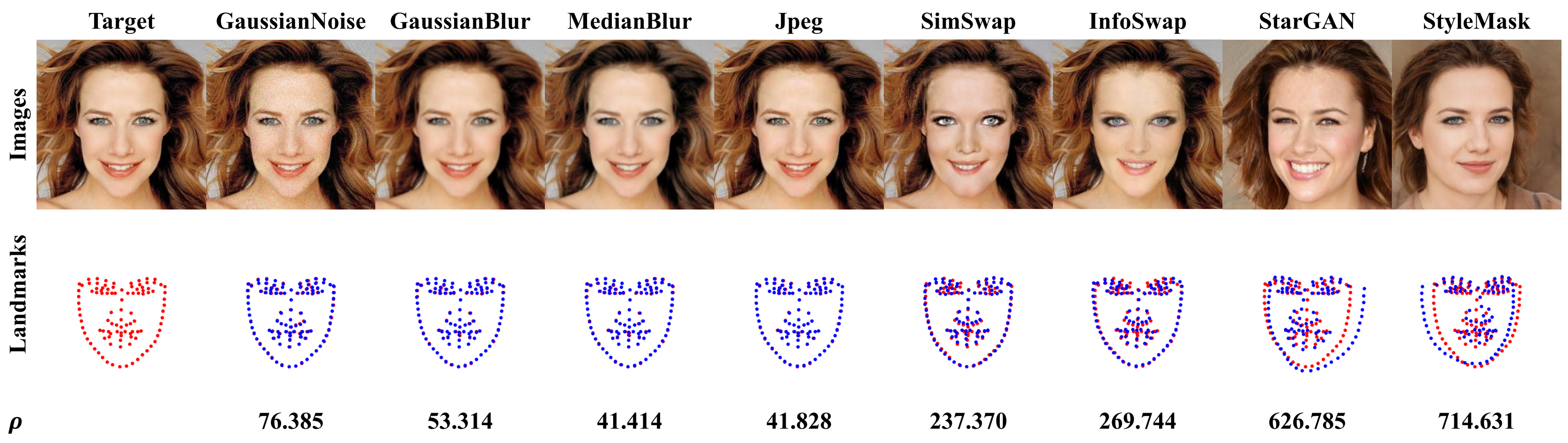
LampMark: Proactive Deepfake Detection via Training-Free Landmark Perceptual Watermarks

Tianyi Wang[1], Mengxiao Huang[2], Harry Cheng[3], Xiao Zhang[2], Zhiqi Shen[1]

[1]College of Computing and Data Science, Nanyang Technological University
[2]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology
[3]School of Computer Science and Technology, Shandong University

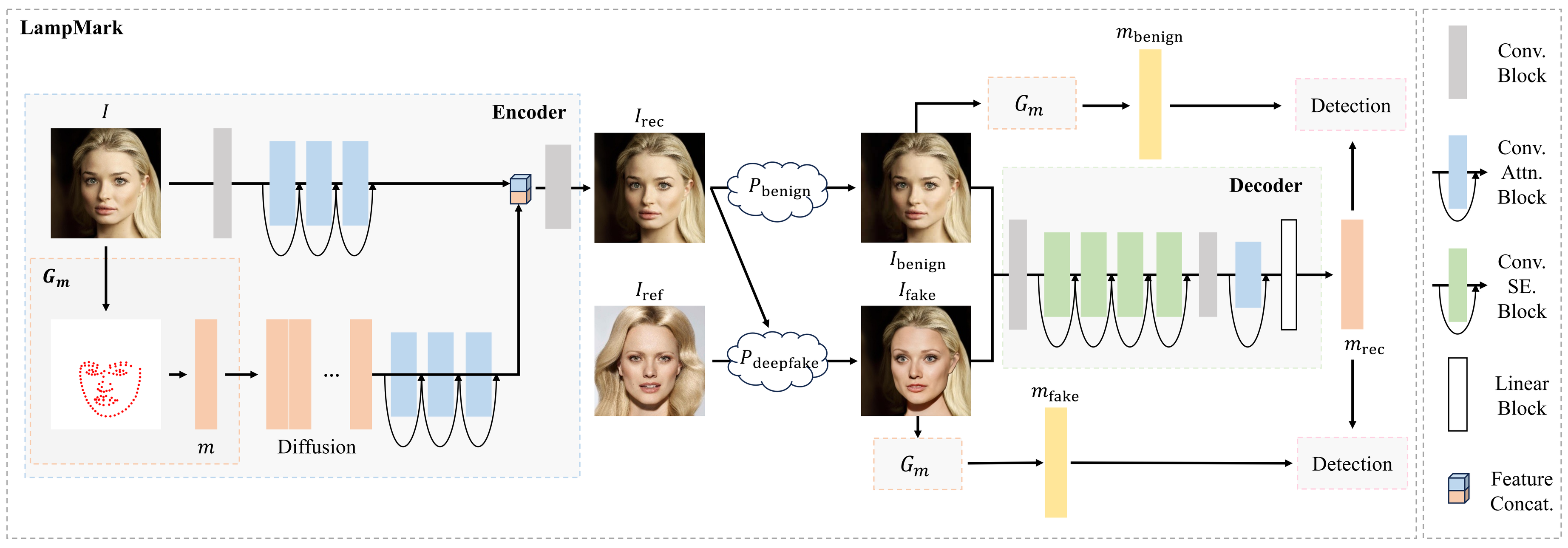
Motivation

- Performance bottlenecks in passive Deepfake detection.
- Unsatisfactory generalizability of existing proactive approaches.
- Structure-sensitive characteristic of Deepfake manipulations: obvious position differences ρ for facial landmarks.
- Benign Deepfake usages shall be allowed.



Main Workflow

- Design the training-free landmark perceptual watermark.
  - Discrimination.
  - Confidentiality.
  - Robustness.
- Construct an auto-encoder for watermark embedding and recovery.
- Perform Deepfake detection based on the consistency between the recovered watermark and the suspect image.



Landmark Perceptual Watermark

- **Discrimination**: no two different facial landmarks corresponds to a same watermark.
  - Facial landmark extraction via Face++.
  - Principle component analysis (PCA) for dimension regulation.
  - Normalization to get binary watermarks.
- **Confidentiality**: watermark encryption to avoid malicious attacks.
  - Cellular automaton encryption system.
  - For an encryption key k_t, the state of each bit i at the next time step $t+1$ is determined by the rule $s_i^{t+1} = R(s_{i-1}^t, s_i^t, s_{i+1}^t)$.
  - $s_i^{t+1} = \begin{cases} s_{l-1}^t \oplus (s_0^t \vee s_1^t), & \text{for } i = 0, \\ s_{i-1}^t \oplus (s_i^t \vee s_{i+1}^t), & 0 < i < l-1, \\ s_{l-2}^t \oplus (s_l^t \vee s_0^t), & i = l-1. \end{cases}$
  - Watermark encryption via XOR operation using selected keys.
- **Robustness**: watermark stays robust against both benign and Deepfake manipulations.
  - Benign image manipulation pool: Gaussian Noise, Gaussian Blur, Median Blur, Jpeg.
  - Malicious Deepfake manipulation pool: SimSwap, InfoSwap, UniFace, E4S, StarGAN, StyleMask, HyperReenact.
  - Model sees only Jpeg and SimSwap during training.

Watermark Embedding and Recovery

- An end-to-end auto-encoder framework.
  - Encoder for watermark embedding.
  - Decoder for watermark recovery.
  - Discriminator for watermarking visual quality improvements.
- Objectives
  - Encoder: $L_I = \|I_{\text{rec}} - I\|_2$.
  - Decoder: $L_m = \|m_{\text{rec}} - m\|_2$
  - Discriminator: $L_{\text{adv}} = -\mathbb{E}\big(\log\big(D(I)\big)\big) + \mathbb{E}(\log(1 - D(I_{\text{rec}})))$.
  - Auxiliary generative loss: $L_G = \|G(I, I_s) - G(I_{\text{rec}}, I_s)\|_2$.

Experimental Results

- Watermarking visual quality evaluation.

- Watermark robustness evaluation via bit-wise recovery accuracy.

	SimSwap [6]	InfoSwap [9]	UniFace [47]	E4S [23]	StarGAN [7]	StyleMask [3]	HyperReenact [2]	Average
HiDDeN [56]	50.02%	50.07%	54.98%	49.19%	50.24%	49.99%	50.15%	50.66%
MBRS [17]	49.98%	50.82%	50.22%	50.07%	49.95%	50.08%	50.08%	50.17%
RDA [51]	50.00%	50.01%	71.15%	63.03%	47.45%	48.94%	56.65%	55.32%
CIN [25]	50.28%	50.60%	46.01%	50.55%	50.05%	50.24%	50.43%	49.74%
ARWGAN [15]	52.06%	47.94%	59.30%	49.81%	50.51%	50.10%	49.86%	51.37%
SepMark [46]	86.17%	77.27%	66.13%	81.62%	49.05%	50.16%	50.05%	65.78%
Ours	**99.95%**	**97.99%**	**99.72%**	**92.09%**	**73.12%**	**74.19%**	**73.53%**	**87.23%**
MBRS [17]	50.00%	50.71%	49.98%	50.07%	49.95%	50.00%	50.07%	50.11%
FaceSigns [26]	49.74%	50.00%	50.59%	49.73%	50.51%	49.10%	49.28%	49.85%
SepMark [46]	92.09%	81.49%	57.44%	77.32%	50.11%	50.06%	50.02%	65.50%
Ours	**99.98%**	**98.31%**	**94.28%**	**93.27%**	**74.66%**	**75.83%**	**74.18%**	**87.21%**

Deepfake Detection

- Given watermarked image I_{rec} embedded with watermark m.
- Common and Deepfake manipulations on I_{rec}, derives I_{benign} and I_{fake}.
- Generate landmark perceptual watermarks regarding I_{benign} and I_{fake}, deriving m_{benign} and m_{fake}.
- The robust watermark m_{rec} can be recovered from I_{benign} and I_{fake}, faithfully similar to m.
- Comparing m_{rec} and m_{benign} leads to high similarity, indicating real.
- Comparing m_{rec} and m_{fake} leads to low similarity, indicating fake.

	Xception [48]		SBIs [32]		RECCE [4]		CADDM [8]		Ours	
Resolution	128	256	128	256	128	256	128	256	128	256
SimSwap [6]	39.37%	71.15%	75.30%	88.94%	60.37%	69.01%	55.91%	87.66%	**97.80%**	**99.01%**
InfoSwap [9]	60.82%	65.50%	85.11%	80.50%	55.51%	52.13%	48.29%	61.39%	**98.59%**	**99.18%**
UniFace [47]	71.79%	70.34%	72.45%	79.41%	61.58%	67.35%	82.16%	82.73%	**96.76%**	**97.03%**
E4S [23]	43.40%	53.70%	63.63%	61.05%	60.88%	47.19%	64.93%	73.13%	**98.99%**	**99.10%**
StarGAN [7]	37.14%	40.30%	48.98%	65.86%	35.82%	41.55%	37.41%	44.34%	**98.96%**	**99.32%**
StyleMask [3]	29.41%	40.23%	38.45%	48.45%	31.08%	23.87%	34.87%	39.73%	**98.62%**	**98.98%**
HyperReenact [2]	38.96%	76.27%	52.36%	53.35%	82.23%	78.23%	35.87%	42.87%	**98.87%**	**99.02%**
Mixed	41.28%	41.42%	60.39%	68.62%	54.09%	52.51%	52.04%	59.84%	**98.39%**	**98.55%**

Summary

- We analyzed the structure sensitivity of images derived by Deepfake manipulations.
- We proposed a training-free landmark perceptual watermark that maintains the original uniqueness of facial landmarks.
- We devised a sophisticated cellular automaton encryption system to securely protect the watermarks.
- We constructed an auto-encoder to robustly embed and recover watermarks.
- Our method outperform the SOTAs across in-dataset, cross-dataset, and cross-manipulation scenarios.

Insights

- Assigning semantics to the robust watermarks completes the detection pipeline without requiring the ground-truth.
- Watermark robustness is achieved for cross-manipulation since the generative goals of Deepfake models are the same.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Contact

Tianyi Wang (王天一)
Web: https://www.tianyiwang.xyz
Email: terry.ai.wang@gmail.com

ACM Multimedia 2024
Melbourne, Australia