

Supplementary Materials: Proactive Deepfake Detection via Training-Free Landmark Perceptual Watermarks

Anonymous Authors

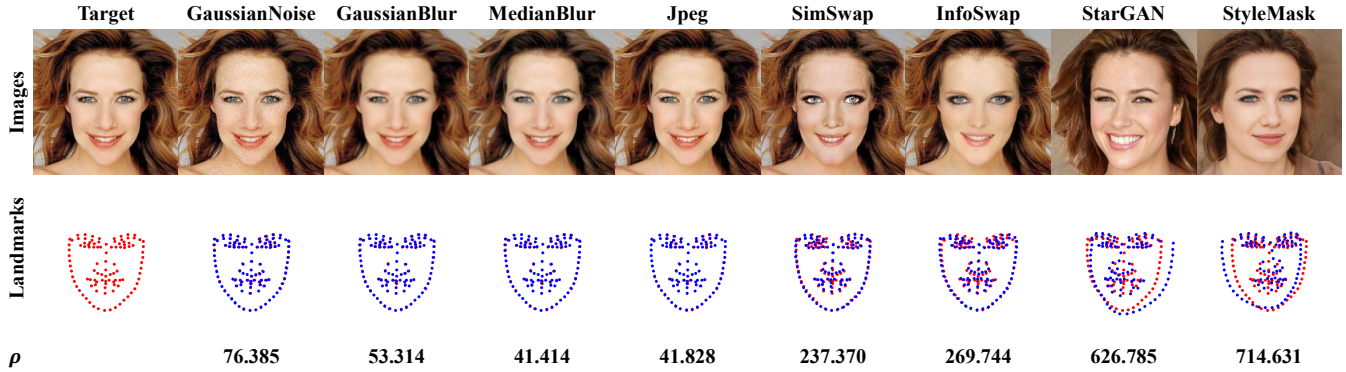


Figure 1: Full demonstration of the structure-sensitive characteristic of Deepfake manipulations regarding facial landmarks. Four benign and four Deepfake manipulations are included. Notation ρ refers to the Euclidean distances between landmarks of the raw and manipulated faces.

1 DETAILED STATISTICS ON LANDMARK OFFSET DISTRIBUTIONS

In the main paper content, we displayed the offset distributions of facial landmarks regarding four benign and four Deepfake manipulations on 10K sampled images at the 256 resolution. Specifically, for every sampled raw image, a manipulated image is generated for each image manipulation and is computed the Euclidean distance with the raw image. In this section, we exhibited the landmark offsets on the sampled target image for all four benign and four Deepfake manipulations in Figure 1. Additionally, in Table 1, the averaged Euclidean distances at 128 and 256 resolutions, denoted as ρ_{128} and ρ_{256} , are reported. The distances consistently demonstrate similar findings as the main paper content discloses such that Deepfake manipulations drastically modify image structures. In particular, swapping an identity brings changes in shape and layout within a face and can lead to relatively critical Euclidean distances away from the original facial landmarks. Meanwhile, even larger ρ values can be observed with respect to the reenactment manipulations that modify the entire expression and head pose. On the contrary, when encountering benign manipulations, the average ρ values at both resolution levels are consistently below 100, regardless of specific operation types.

2 EXTENSIVE ROBUSTNESS EVALUATION ON MORE BENIGN MANIPULATIONS

In the main content of this paper, the model is trained adversarially against Jpeg and SimSwap [2] and evaluated on four benign and seven Deepfake manipulations. While the four benign manipulations (GaussianNoise, GaussianBlur, MedianBlur, and Jpeg) are the most commonly seen ones that can best imitate real-life scenarios upon spreading images on the internet, in this section, we further

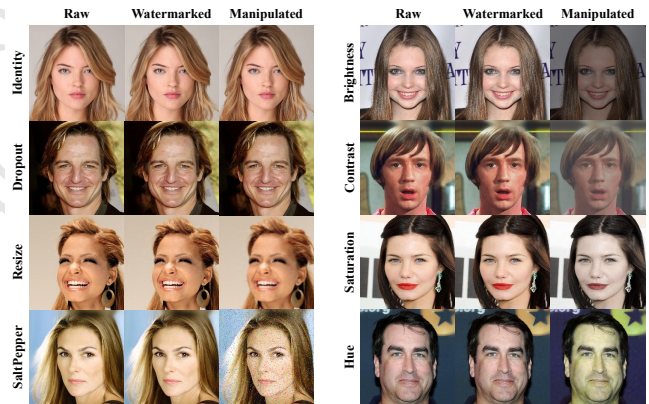


Figure 2: Visualizations of the effects for each benign manipulation.

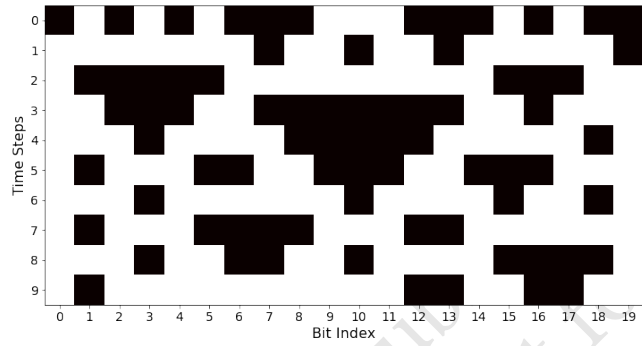
conducted evaluations on the trained model at the 128 resolution on CelebA-HQ [7] with more benign manipulations that are also popularly discussed in recent studies, namely, Identity, Dropout, Resize, SaltPepper, Brightness, Contrast, Saturation, and Hue. The visualizations of the effects for each manipulation are displayed in Figure 2 and the comparative results regarding the bit-wise watermark recovery accuracy are listed in Table 2. As a result, while the contrastive models are mostly vulnerable regarding some of the manipulations, SepMark [10] and our approach favorably maintain the robustness with the highest average watermark recovery accuracies of 99.98% and 99.97%, respectively. Although slightly surpassed by SepMark, the proposed LampMark promisingly maintains accuracies above 99.90% for all benign manipulations discussed in this section.

Table 1: Landmark offsets between manipulated and original images, where ρ_{128} and ρ_{256} refer to the average Euclidean distances at 128 and 256 resolutions, respectively.

	GaussianNoise	GaussianBlur	MedianBlur	Jpeg	SimSwap [2]	InfoSwap [4]	StarGAN [3]	StyleMask [1]
ρ_{128}	60.854	57.454	55.312	55.870	184.772	183.624	454.812	339.513
ρ_{256}	91.770	63.443	63.069	73.341	346.389	344.951	691.076	565.585

Table 2: Watermark robustness evaluation against further benign manipulations. Bit-wise watermark recovery accuracies are computed regarding each manipulation.

Manipulations	HiDDeN [12]	MBRS [6]	RDA [11]	CIN [8]	ARWGAN [5]	SepMark [10]	Ours
Identity	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%
Dropout	82.44%	99.99%	94.76%	99.99%	97.25%	99.99%	99.99%
Resize	82.01%	99.99%	99.94%	99.17%	93.73%	99.99%	99.99%
SaltPepper	52.30%	99.37%	66.62%	93.95%	64.55%	99.96%	99.97%
Brightness	75.34%	99.96%	99.96%	98.47%	99.02%	99.99%	99.97%
Contrast	70.02%	98.43%	99.68%	99.99%	98.97%	99.99%	99.95%
Saturation	78.10%	99.92%	99.83%	99.99%	99.01%	99.99%	99.97%
Hue	71.30%	31.56%	80.85%	99.98%	87.93%	99.99%	99.91%
Average	76.43%	91.15%	92.70%	98.94%	92.56%	99.98%	99.97%

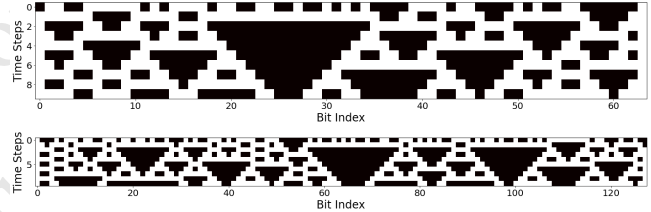
**Figure 3: Visualization of the binary key values from time steps 0 to 9 in white and dark grids with a key length of 20.**

3 EXPLANATION ON WATERMARK CONFIDENTIALITY

In order to ensure watermark confidentiality, a cellular automaton encryption system is devised following Rule 30 [9] in the main content, with equations denoted as follows,

$$s_i^{t+1} = \begin{cases} s_{l-1}^t \oplus (s_0^t \vee s_1^t), & \text{for } i = 0, \\ s_{i-1}^t \oplus (s_i^t \vee s_{i+1}^t), & \text{for } 0 < i < l-1, \\ s_{l-2}^t \oplus (s_l^t \vee s_0^t), & \text{for } i = l-1. \end{cases} \quad (1)$$

In specific, given an initial key k_0 with binary values, the bit values of the succeeding keys at each time step are determined by every three consecutive bit values iteratively at the preceding time step. In this section, we demonstrated the unpredictable and complex characteristics of this encryption system by transforming a sample

**Figure 4: Visualization of the binary key values in white and dark grids in real cases with key lengths of 64 and 128.**

initial key k_0 of length 20 for nine time steps. To begin with, we randomly initialized an initial key k_0 of length 20.

k_0 : 0 1 0 1 0 1 0 0 0 1 1 1 0 0 0 1 0 1 0 0

Then, based on the rule and equation, the next key k_1 is denoted accordingly.

k_1 : 1 1 1 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0

Particularly, the bit value 1 at the first index of k_1 is determined by values at the last index and the first two indices, 0 0 1, of k_0 . Similarly, all bit values are computed iteratively over the indices. Thereafter, the keys k_i for $2 \leq i \leq 9$ are derived as follows.

k_2 : 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 1

k_3 : 1 1 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 1 1 1

k_4 : 1 1 1 0 1 1 1 1 0 0 0 0 0 1 1 1 1 1 0 1

k_5 : 1 0 1 1 1 0 0 1 1 0 0 0 1 1 0 0 0 1 1 1

k_6 : 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1

k_7 : 1 0 1 1 1 0 0 0 0 1 1 1 0 0 1 1 1 1 1 1

k_8 : 1 1 1 0 1 1 0 0 1 1 0 1 1 1 1 0 0 0 0 1

k_9 : 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 0 0 1 1

Following the rule and equation, the nine keys are nonrepeated as expected. To visualize, we plotted the ten keys, including k_0 , in white and dark grids for binary values 1 and 0, respectively. It can be easily observed in Figure 3 that there are no identical rows, implying the uniqueness of each key. Suppose k_3 , k_5 , and k_8 are randomly drafted as the encryption keys applied to the raw watermarks, the derived encrypted watermarks after sequential XOR operations are unpredictable, and it is complex to recover the raw watermarks without knowing the encryption keys. Lastly, we visualized the keys in real cases of our experiments, setting the key lengths to 64 and 128. As a result, in Figure 4, the key sequences with both lengths successfully maintain the key uniqueness, demonstrating watermark confidentiality.

REFERENCES

- [1] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos. 2023. StyleMask: Disentangling the Style Space of StyleGAN2 for Neural Face Reenactment. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, Los Alamitos, CA, USA, 1–8. <https://doi.org/10.1109/FG57933.2023.10042744>
- [2] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 2003–2011. <https://doi.org/10.1145/3394171.3413630>
- [3] Yunjei Choi, Youngjung Uh, Jaesun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8185–8194. <https://doi.org/10.1109/CVPR42600.2020.00821>
- [4] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. 2021. Information Bottleneck Disentanglement for Identity Swapping. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3403–3412. <https://doi.org/10.1109/CVPR46437.2021.00341>
- [5] Jiangtao Huang, Ting Luo, Li Li, Gaobo Yang, Haiyong Xu, and Chin-Chen Chang. 2023. ARWGAN: Attention-Guided Robust Image Watermarking Model Based on GAN. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–17. <https://doi.org/10.1109/TIM.2023.3285981>
- [6] Zhaoyang Jia, Han Fang, and Weiming Zhang. 2021. MBRS: Enhancing Robustness of DNN-based Watermarking by Mini-Batch of Real and Simulated JPEG Compression. In *Proceedings of the 29th ACM International Conference on Multimedia (Virtual Event, China)*. Association for Computing Machinery, New York, NY, USA, 41–49. <https://doi.org/10.1145/3474085.3475324>
- [7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Int. Conf. Learn. Represent.*
- [8] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. 2022. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia (<conf-loc>, <city>Lisboa</city>, <country>Portugal</country>, </conf-loc>)*. Association for Computing Machinery, New York, NY, USA, 1532–1542. <https://doi.org/10.1145/3503161.3547950>
- [9] Stephen Wolfram. 2002. *A New Kind of Science*. Wolfram Media, Inc.
- [10] Xiaoshuai Wu, Xin Liao, and Bo Ou. 2023. SepMark: Deep Separable Watermarking for Unified Source Tracing and Deepfake Detection. In *ACM Int. Conf. Multimedia*.
- [11] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz. 2021. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 14428–14437. <https://doi.org/10.1109/ICCV48922.2021.01418>
- [12] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. HiDDeN: Hiding Data With Deep Networks. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 682–697.