Track 1:
Adversarial Watermarking for Face Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

Watermarking is an essential technique for embedding an identifier (*i.e.*, watermark message) within digital images to assert ownership and monitor unauthorized alterations. In face recognition systems, watermarking plays a pivotal role in ensuring data integrity and security. However, an adversary could potentially interfere with the watermarking process, significantly impairing recognition performance. We explore the interaction between watermarking and adversarial attacks on face recognition models. Our findings reveal that while watermarking or input-level perturbation alone may have a negligible effect on recognition accuracy, the combined effect of watermarking and perturbation can result in an *adversarial watermarking attack*, significantly degrading recognition performance. Specifically, we introduce a novel threat model, the adversarial watermarking attack, which remains stealthy in the absence of watermarking, allowing images to be correctly recognized initially. However, once watermarking is applied, the attack is activated, causing recognition failures. Our study reveals a previously unrecognized vulnerability: *adversarial perturbations can exploit the watermark message to evade face recognition systems.* Evaluated on the CASIA-WebFace dataset, our proposed adversarial watermarking attack reduces face matching accuracy by 67.2% with an ℓ_∞ norm-measured perturbation strength of $2/255$ and by 95.9% with a strength of $4/255$.

1 Introduction

Face recognition systems have become increasingly prevalent in various domains, such as access control and surveillance [1–3]. Ensuring the integrity and ownership of facial images used for training and evaluation in such systems is crucial. Image watermarking has offered a viable solution for proprietary face image protection [4–6]. Watermarking can embed hidden information (also called 'watermark message') in digital faces to assert ownership, authenticate content, and verify data integrity [7–9].

However, as machine learning (ML) models become more sophisticated, they also become susceptible to adversarial attacks. Adversarial perturbations (also known as evasion attacks) are carefully crafted modifications to input data that deceive ML models without noticeable changes in the image to human observers [10–12]. In the context of face recognition, such perturbations can cause recognition errors, leading to security breaches; See the literature review in Section 2.

Although watermarking aims to protect and authenticate images, the interaction between watermarking processes and adversarial attacks remains underexplored. The presence of watermarking and adversarial attacks, along with their interaction, has added substantial complexity to evaluation of face recognition systems. Inspired by the above, we address the following question:

(Q) *How does watermarking affect the adversarial robustness of face recognition systems, and can adversarial attacks exploit watermarking to even degrade face matching performance?*

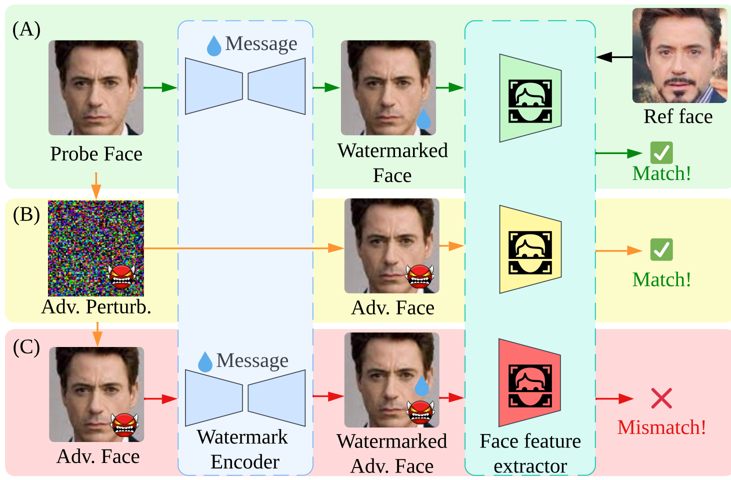

Figure 1: **Overview of the Adversarial Watermarking Attack on Face Recognition.** The green path (A) represents the standard watermarking and face recognition process, where the probe face is watermarked using the watermark encoder and correctly matched with the reference face after feature extraction. The yellow path (B) shows input-level adversarial perturbations applied to evade the face recognition system without watermarking. Subtle adversarial perturbations are added to the probe face, but they do not affect the recognition result without watermarking. The red path (C) demonstrates the adversarial watermarking process, where the adversarially perturbed face image, after being watermarked, fails to match the reference face.

To the best of our knowledge, our work unveils the joint effects of watermarking and adversarial attacks on face recognition models for the first time. We summarize our contributions below.

• We propose a testbed (Figure 1) that integrates watermarking techniques into face recognition systems. This framework embeds watermarks into facial images to assert ownership while facilitating the study of adversarial attacks (Figure 1-(B) and (C)).

• We introduce a new threat model (Figure 1-(C)) called the Adversarial Watermarking attack, which differs from conventional evasion attacks against image classifiers [10, 13, 14]. This attack is designed to remain stealthy when watermarking is absent (Figure 1-(B)), allowing images to be correctly recognized initially. However, once watermarking is applied, the attack is triggered, causing recognition failures and exposing a critical vulnerability in the watermarking process.

• We validate our proposed attack through extensive experiments on the open-source CASIA-WebFace dataset. Our results demonstrate a significant degradation in face matching performance under small adversarial perturbations (*e.g.*, $\frac{2}{255}$ and $\frac{4}{255}$) when the watermarking is applied (Figure 1).

2 Related Work

Watermarking in Face Recognition. Watermarking techniques have long been used to embed imperceptible information into digital images for purposes such as copyright protection, authentication, and integrity verification [6, 15, 16]. In the realm of face recognition, watermarking serves as a tool to protect personally identifiable images from unauthorized use and tampering [4, 17–21]. Various methods have been proposed to integrate watermarking into facial images without significantly affecting recognition performance. Traditional watermarking approaches use frequency domain transformations such as Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) to embed watermarks in images, with the aim of robustness against common image processing attacks [22, 23]. In contrast, recent methods leverage deep neural networks (DNNs) for watermarking, such as the HiDDeN framework, which employs end-to-end trainable networks to embed and extract watermarks, enhancing resilience against various attacks [5]. Other recent studies have focused on ensuring that the watermarking process preserves critical facial features essential for accurate recognition [7, 8, 17]. However, these methods mainly focus on robustness against non-adversarial distortions and fail to account for the impact of adversarial perturbations specifically designed to deceive ML models, particularly when watermarking is applied.

Adversarial Attacks in Face Recognition. Adversarial attacks involve introducing subtle, often imperceptible perturbations to input data with the intent of deceiving ML models [10, 13, 24]. In face recognition systems, adversarial examples can lead to recognition errors, impersonation, or evasion, posing significant security risks [25–27]. For example, attackers can manipulate facial images to bypass authentication systems or to impersonate other enrolled individuals in the system. Various attack generation algorithms, such as Fast Gradient Sign Method (FGSM) [10] and Projected Gradient Descent (PGD) [14], have been employed to generate adversarial examples against face recognition models. Meanwhile, defense mechanisms such as adversarial training and input pre-processing have been proposed to mitigate these attacks [14, 24, 28]. The ongoing arms race between attack and defense persists. However, existing studies have primarily focused on evading or improving the robustness of model performance, without considering the impact of watermarking whose use is growing, *e.g.,* for labeling computer generated images. To the best of our knowledge, the interaction between adversarial perturbations and watermarking in face recognition is largely unexplored, with no prior work investigating how adversarial attacks leverage watermarking to degrade recognition performance.

3 Methods

Watermarking System. We start by introducing the technique used for generating watermarked face images and its application in the subsequent face recognition task, as shown in Figure 1-(A). To formalize the watermarking problem, let the input image be denoted as $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$, and a binary watermark message as $\mathbf{m} \in \{0,1\}^L$ (an L-bit digital signature) embedded into the facial images [5, 7, 29]. Our goal is to produce a watermarked image $\mathbf{I}_{\text{w}}$ that maintains visual similarity to the original image $\mathbf{I}$ containing the watermark message $\mathbf{m}$. Furthermore, the watermarked image should allow extraction of $\mathbf{m}$, allowing provenance of the image.

We implement the watermarking system using the open source neural network-based HiDDeN framework [5]. This system consists of an encoder network $f_{\boldsymbol{\theta}}$ and a decoder network $g_{\boldsymbol{\phi}}$. The encoder takes the input image $\mathbf{I}$ and the watermark message $\mathbf{m}$ as inputs and generates the watermarked image $\mathbf{I}_{\text{w}} = f_{\boldsymbol{\theta}}(\mathbf{I}, \mathbf{m})$. The decoder takes the watermarked image $\mathbf{I}_{\text{w}}$ as input and reconstructs the embedded watermark message $\hat{\mathbf{m}} = g_{\boldsymbol{\phi}}(\mathbf{I}_{\text{w}})$. The encoder and decoder networks are jointly trained using a combination of image reconstruction loss and message decoding loss. The loss of image reconstruction ℓ_{recons} (*e.g.*, mean squared error) ensures that the watermarked image is visually similar to the original, while the loss of message decoding ℓ_{decode} (*e.g.*, bitwise binary cross-entropy loss) minimizes the difference between embedded and extracted watermark messages. The overall training objective for watermarking encoder and decoder is:

Table 1: The robustness of watermarking evaluated using the reconstructed watermark bit accuracy (%) against various (post-watermarking) data transformations at different scaling strengths. Each value is averaged over 1000 face images, with an image size of 112×112 and a watermark string bit length of 48. See more setup details in Section 4.

Transformation	**Scaling ratio**					
	1	0.95	0.9	0.85	0.8	0.75
Crop	98.39	97.22	93.7	95.12	94.77	94.3
Resize	98.39	92.47	92.0	91.58	89.62	85.93
Transformation	**Scaling factor**					
	1	1.5	2	2.5	3	3.5
Brightness	98.39	98.48	96.65	94.21	91.6	88.87
Contrast	98.39	98.81	98.15	96.82	94.92	92.62
Transformation	**JPEG quality factor**					
	100	95	90	85	80	75
JPEG compression	98.39	90.36	85.0	80.8	76.65	73.06

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{\mathbf{I},\mathbf{m}} \left[\ell_{\text{recons}}(\mathbf{I}_{\text{w}}, \mathbf{I}) + \lambda \ell_{\text{decode}}(\hat{\mathbf{m}}, \mathbf{m})\right] \tag{1}$$

where λ is a regularization parameter balancing the two losses. During training, a random message generator produces random bits for $\mathbf{m}$. This randomness allows the network to generalize to any watermark message, enabling us to embed user-defined messages in face images later on. Table 1 shows that our watermarking system is fairly robust against different data transformations. However, as demonstrated later, this does not guarantee adversarial robustness for the downstream task when using watermarked data.

Face Recognition on Watermarked Images. With watermarked face images acquired above, we proceed to face recognition to assess the impact of the watermarking. In what follows, we provide a brief background on face recognition. Given an input face image $\mathbf{I}$, the face recognition model $h_{\boldsymbol{\psi}}$ maps the image to a feature representation $\mathbf{z}$: $\mathbf{z} = h_{\boldsymbol{\psi}}(\mathbf{I})$, where $\boldsymbol{\psi}$ represents the learnable parameters of the model. The feature $\mathbf{z}$ is typically extracted from the penultimate layer of a convolutional neural network (CNN), such as ResNet [30]. The model is trained to minimize a classification loss, such as the softmax loss [31] or margin-based losses [31–33], which encourage facial features from the same identity to be close in the embedding space while pushing apart facial features from different identities. During inference, the model extracts feature representations for a probe face $\mathbf{I}_\mathrm{p}$ and a reference face $\mathbf{I}_\mathrm{r}$, denoted as $\mathbf{z}_\mathrm{p}$ and $\mathbf{z}_\mathrm{r}$, respectively. The similarity between the probe and reference faces is computed using the cosine similarity:

$$s(\mathbf{z}_\mathrm{p}, \mathbf{z}_\mathrm{r}) = \frac{\mathbf{z}_\mathrm{p}^\top \mathbf{z}_\mathrm{r}}{|\mathbf{z}_\mathrm{p}||\mathbf{z}_\mathrm{r}|} \tag{2}$$

where $|\cdot|$ denotes the Euclidean norm. A match is determined based on whether the similarity score exceeds a predefined threshold τ:

$$\mathrm{match}(\mathbf{z}_\mathrm{p}, \mathbf{z}_\mathrm{r}) = \begin{cases} 1, & \text{if } s(\mathbf{z}_\mathrm{p}, \mathbf{z}_\mathrm{r}) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Our experiments later verify that the watermarking process does not significantly degrade face recognition performance in the absence of adversarial perturbations.

Adversarial Watermarking Attack for Face Recognition. We introduce an adversarial watermarking attack that exploits the interaction between adversarial perturbations and the watermarking process to degrade face recognition performance. The adversary aims to craft a minimal perturbation $\boldsymbol{\delta}$ added to a probe face image $\mathbf{I}_\mathrm{p}$ and find a specific watermark message $\mathbf{m} \in \{0,1\}^L$ such that:

1. **Pre-watermark recognition success:** The perturbed image $\mathbf{I}'_\mathrm{p} = \mathbf{I}_\mathrm{p} + \boldsymbol{\delta}$ is correctly matched with the reference image $\mathbf{I}_\mathrm{r}$ by the face recognition model $h_{\boldsymbol{\psi}}$, *i.e.*, the similarity between their feature representations remains high. Here $\boldsymbol{\delta} \in \mathbb{R}^{H\times W\times C}$ denotes adversarial perturbations bounded by $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, where ϵ is the perturbation strength ensuring imperceptibility.
2. **Post-watermark recognition failure:** After applying the watermarking encoder $f_{\boldsymbol{\theta}}$ with the adversary-learned watermark message $\mathbf{m}$, the perturbed input image $\mathbf{I}'_\mathrm{p}$ and its watermarked counterpart $\mathbf{I}'_\mathrm{w} = f_{\boldsymbol{\theta}}(\mathbf{I}'_\mathrm{p}, \mathbf{m})$ lead to a low similarity with the reference image $\mathbf{I}_\mathrm{r}$, causing the face recognition model $h_{\boldsymbol{\psi}}$ to fail.

Our rationale has two key aspects. First, satisfying both conditions 1 and 2 ensures that the adversarial attack ($\boldsymbol{\delta}$) stays stealthy when watermarking is absent, but is triggered upon watermark application, leading to recognition failures. Second, this design reveals a unique adversarial challenge in face recognition with watermarking, where the optimization of the watermark message in condition 2 interacts synergistically with the input perturbations $\boldsymbol{\delta}$ to amplify the adversarial effect.

We propose the following joint optimization problem to find the adversarial perturbation $\boldsymbol{\delta}$ and the watermark message $\mathbf{m}$:

$$\min_{\mathbf{m}\in\{0,1\}^L} \min_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} -s(\mathbf{z}'_\mathrm{p}, \mathbf{z}_\mathrm{r}) + s(\mathbf{z}'_\mathrm{w}, \mathbf{z}_\mathrm{r}) \tag{4}$$

where the optimization variables are the binary watermark message $\mathbf{m}$ and the input perturbations $\boldsymbol{\delta}$, and $s(\cdot,\cdot)$ and $\mathbf{z}_\mathrm{r}$ are defined in (2). Recall that $\mathbf{z}'_\mathrm{p} = h_{\boldsymbol{\psi}}(\mathbf{I}'_\mathrm{p})$ and $\mathbf{z}'_\mathrm{w} = h_{\boldsymbol{\psi}}(\mathbf{I}'_\mathrm{w})$ are the feature representations given the probe image $\mathbf{I}'_\mathrm{p} = \mathbf{I}_\mathrm{p} + \boldsymbol{\delta}$ and $\mathbf{I}'_\mathrm{w} = f_{\boldsymbol{\theta}}(\mathbf{I}'_\mathrm{p}, \mathbf{m})$, respectively. In (4), the original similarity term $s(\mathbf{z}'_\mathrm{p}, \mathbf{z}_\mathrm{r})$ ensures that the perturbed face is still recognized as the same identity in the absence of watermarking. And the watermarked similarity term $s(\mathbf{z}'_\mathrm{w}, \mathbf{z}_\mathrm{r})$ minimizes the similarity between the watermarked, perturbed image and the reference image, causing face recognition failure post-watermarking.

To solve the optimization in (4), we then adopt an alternative optimization procedure to jointly optimize $\boldsymbol{\delta}$ and $\mathbf{m}$. Specifically, we use the PGD (projected gradient descent) method [14] to iteratively minimize one variable while keeping the other fixed. In the optimization process, we face the challenge of the discrete nature of the watermark message $\mathbf{m}$. Direct optimization over binary variables is computationally intractable for large dimensionality L. To address this, we relax $\mathbf{m}$ to be continuous in the range $[0,1]^L$ during the optimization. This relaxation allows us to employ PGD in an efficient way. That is, after performing gradient descent on the relaxed $\mathbf{m}$, we project back onto the binary set $\{0,1\}^L$ by rounding each element to 0 or 1. This ensures the watermark message remains valid for the encoder. By alternately optimizing over $\boldsymbol{\delta}$ and $\mathbf{m}$, we minimize the joint objective. This approach finds a combination of adversarial perturbation and a watermark message that maintains high genuine similarity before watermarking and cause misrecognition afterward.

4 Experiments

Experimental Setup. We use the **CASIA-WebFace** dataset [34], containing face images of 10,575 individuals, for evaluating face recognition models. We extract 1,000 individuals with two matching face images for each identity ($\mathbf{I}_{\mathrm{p}}$ and $\mathbf{I}_{\mathrm{r}}$), and pre-processed them by aligning and resizing the images to 112×112 pixels. We adopt our face recognition model from the **AdaFace** framework [35]. AdaFace is known for its adaptive margin loss that accounts for the quality of the face images, improving recognition performance. The model is trained on MS-Celeb-1M dataset [31] using standard training protocols with a ResNet-50 backbone [30]. For watermarking, we follow the **HiDDeN** framework [5] to solve the problem (1). The encoder and decoder networks are trained on the MS-COCO dataset [36] with random 48-bit watermark messages. The trained encoder is then used to embed watermarks in the CASIA-WebFace face images. In generating the adversarial watermarking attack (4), the step sizes for optimizing $\boldsymbol{\delta}$ and $\mathbf{m}$ are set to $\alpha = \frac{\epsilon}{T}$ and $\beta = \frac{1}{T}$, respectively, where $T = 10$ represents the number of iterations for the PGD-10 attack.

Evaluation. We assess the effectiveness of the adversarial watermarking attack by analyzing face recognition performance under two key conditions. First, in the case of recognition with adversarial perturbations, adversarial perturbations are applied to the probe images *without watermarking*. Next, in the case of recognition with the adversarial watermarking attack (*with watermarking*), both adversarial perturbations and an optimized watermark message are applied, following the joint optimization in (4).

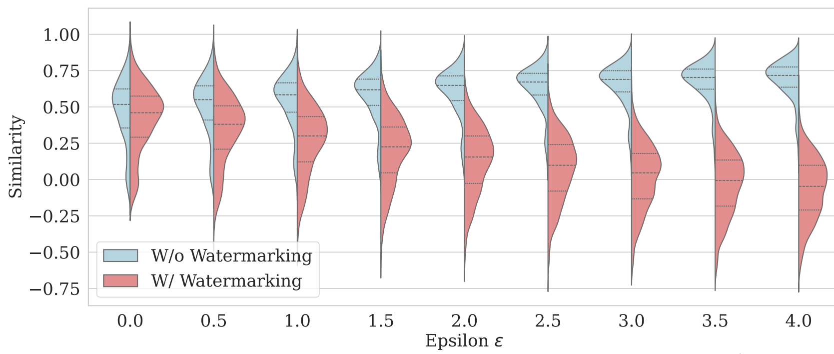

Figure 2: Violin plots of similarity scores in (2) at different ϵ values (scaled by $1/255$). For each ϵ, the violin plot shows the distribution of similarity scores between perturbed probe and reference images under two conditions: with watermarking (blue) and without watermarking (red). By $\|\delta_\infty\| \leq \epsilon$, we change ϵ to control the perturbation strength.

Table 2: Face matching accuracy (%) with and without watermarking at different perturbation levels (ϵ, scaled by $1/255$), where the matching threshold is set to $\tau = 0.3$ in (3). Performance reduction by watermarking attack is highlighted in blue.

ϵ	**0.0**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**
W/o Watermarking	81.8	85.4	88.5	90.9	92.2	94.1	95.7	97.5	98.3
W/ Watermarking	73.9	63.5	50.0	35.7	25.0	16.5	8.4	4.5	2.4
Reduction	7.9	21.9	38.5	55.2	67.2	77.6	87.3	93.0	95.9

Adversarial Watermarking: Joint Effects of Watermarking and Adversarial Perturbations. To analyze the effect of the adversarial watermarking attack on face recognition performance, we first examine the similarity scores between probe and reference images across different perturbation strengths ϵ. **Figure 2** shows violin plots of the similarity distributions for face recognition, both with and without watermarking, when evaluated using input perturbations $\boldsymbol{\delta}$ from the proposed adversarial watermarking attack. As the perturbation strength ϵ increases, the similarity between probe and reference images decreases significantly in the presence of watermarking, while it remains largely unaffected without watermarking. This is because in the absence of watermarking, the first loss term in (4) aims to maximize the similarity between the probe image and the reference image for the applied perturbations $\boldsymbol{\delta}$. With watermarking in the face recognition process, the similarity score quickly drops with increased perturbation strength. In fact, when $\epsilon = 0.5/255$, the similarity has tended to be smaller than the matching threshold τ (commonly set at $\tau = 0.3$). This shows that even a small adversarial perturbation can disrupt face recognition after watermarking, although performance remains stable without watermarking.

Table 2 shows that watermarking reduces face matching accuracy at all perturbation levels (ϵ). For example, at $\epsilon = 0.0$, accuracy drops by **7.9%** from 81.8% to 73.9% after watermarking. This indicates that the adversarial watermark message alone, as found by (4), reduces recognition accuracy. As the perturbation magnitude ϵ increases, the accuracy reduction intensifies. At $\epsilon = 2/255$, the accuracy decreases by **67.2%**, from 92.2% to 25.0%, and at $\epsilon = 4/255$, the reduction reaches **95.9%**, with the accuracy dropping from 98.3% to just 2.4%. These drastic reductions illustrate the adversarial watermarking attack's effectiveness in significantly degrading face recognition, especially at higher perturbation magnitudes. The results demonstrate that adversarial watermarking exploits the interaction with perturbations, significantly reducing face matching accuracy.

Visualizations of Face Images vs. Watermarking and Perturbations. Figure 3 examines the combination of watermarking and perturbations (with strength ϵ at $4/255$) on face images. To compare with reference faces (a), original faces (b) are visualized along with similarity scores by (2). Watermarked faces (c) are added with message m by (4), along with similar scores to (b), exhibiting minor effects by watermarking. Perturbed faces (d) are added with perturbation $\boldsymbol{\delta}$ by (4), along with larger scores than (b), maintaining the face matching performance. Adversarial watermarked faces (g) have extremely low similarity scores, exhibiting the joint adversarial effect of watermarking and perturbation. Element-wise absolute differences are visualized in (d), (f), and (h) respectively for (c), (e), (g) to show the imperceptibility of watermark/perturbation. It should be noted that adversarial watermarking difference (h) shows more focus on the edges and corners, *i.e.*, the high frequency area than (d) and (f), illustrating why the attack works while watermakring or perturbation alone does not.

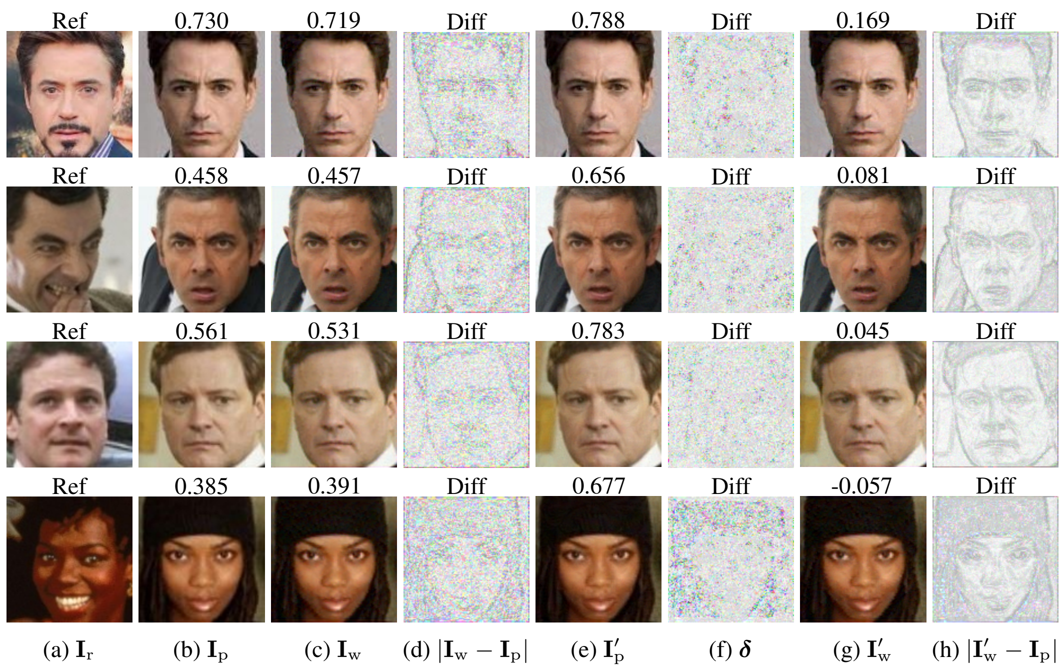

Figure 3: Visualization of reference, probe, and perturbed/watermarked face images along with perturbation/watermark for four identities. (**a**) Reference face. (**b**) Probe face. (**c**) Watermarked face. (**d**) Difference between (b) and (c). (**e**) Perturbed face. (**f**) Difference between (b) and (e). (**g**) Adversarial watermarked face by watermarking perturbed face. (**h**) Difference between (b) and (g). All element-wise absolute differences are scaled by $\times 10$ and color reverted. All probe faces are marked with their similarity score compared with reference faces at the top of images.

5 Conclusion

Our study investigated the vulnerabilities of face recognition systems when adversarial perturbations are combined with watermarking. While watermarking alone had a minimal effect on recognition accuracy, the introduction of adversarial perturbations before watermarking caused significant performance degradation. Our findings show that adversarial watermarking attacks could severely undermine recognition systems even if they remain stealthy when watermarking is absent, highlighting the need for improved defenses in both watermarking and face recognition models.

References

[1] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.

[2] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*, 2018.

[3] Muhammad Irwan Padli Nasution, Nurbaiti Nurbaiti, Nurlaila Nurlaila, Tri Inda Fadhila Rahma, and Kamilah Kamilah. Face recognition login authentication for digital payment solution at covid-19 pandemic. In *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, pages 48–51. IEEE, 2020.

[4] Anil K Jain and Umut Uludag. Hiding biometric data. *IEEE transactions on pattern analysis and machine intelligence*, 25(11):1494–1498, 2003.

[5] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018.

[6] Mahbuba Begum and Mohammad Shorif Uddin. Digital image watermarking techniques: a review. *Information*, 11(2):110, 2020.

[7] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[8] Yunming Zhang, Dengpan Ye, Sipeng Shen, Caiyun Xie, Ziyi Liu, Jiacheng Deng, and Long Tang. Double privacy guard: Robust traceable adversarial watermarking against face recognition. *arXiv preprint arXiv:2404.14693*, 2024.

[9] Koushik Pal, G Ghosh, and M Bhattacharya. Biomedical image watermarking in wavelet domain for data integrity using bit majority algorithm and multiple copies of hidden information. *American Journal of Biomedical Engineering*, 2(2):29–37, 2012.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Yifan Gong, Yuguang Yao, Yize Li, Yimeng Zhang, Xiaoming Liu, Xue Lin, and Sijia Liu. Reverse engineering of imperceptible adversarial image perturbations. *arXiv preprint arXiv:2203.14145*, 2022.

[12] Pu Zhao, Parikshit Ram, Songtao Lu, Yuguang Yao, Djallel Bouneffouf, Xue Lin, and Sijia Liu. Learning to generate image source-agnostic universal adversarial perturbations. *arXiv preprint arXiv:2009.13714*, 2020.

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[15] Frank Hartung and Martin Kutter. Multimedia watermarking techniques. *Proceedings of the IEEE*, 87(7):1079–1107, 1999.

[16] François Cayre, Caroline Fontaine, and Teddy Furon. Watermarking security: theory and practice. *IEEE Transactions on signal processing*, 53(10):3976–3987, 2005.

[17] Yuguang Yao, Steven Grosz, Sijia Liu, and Anil Jain. Hide and seek: How does watermarking impact face recognition? *arXiv preprint arXiv:2404.18890*, 2024.

[18] Mohd Rizal Mohd Isa, Salem Aljareh, and Zaharin Yusoff. A watermarking technique to improve the security level in face recognition systems. *Multimedia Tools and Applications*, 76:23805–23833, 2017.

[19] Mayank Vatsa, Richa Singh, Afzel Noore, Max M Houck, and Keith Morris. Robust biometric image watermarking for fingerprint and face template protection. *IEICE Electronics Express*, 3(2):23–28, 2006.

[20] Mohammed AM Abdullah, Satnam S Dlay, Wai L Woo, and Jonathon A Chambers. A framework for iris biometrics protection: a marriage between watermarking and visual cryptography. *IEEE Access*, 4:10180–10193, 2016.

[21] Gaurav Bhatnagar and QM Jonathan Wu. Biometrics inspired watermarking based on a fractional dual tree complex wavelet transform. *Future Generation Computer Systems*, 29(1):182–195, 2013.

[22] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital watermarking and steganography*. Morgan kaufmann, 2007.

[23] Deepika Singh and Satnam Singh. Dwt-dct and svd based robust and blind watermarking scheme for copyright protection. *Multimedia Tools and Applications*, 76(11):13001–13024, 2017.

[24] Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jinfeng Yi, Mingyi Hong, Shiyu Chang, and Sijia Liu. How to robustify black-box ml models? a zeroth-order optimization perspective. *arXiv preprint arXiv:2203.14195*, 2022.

[25] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1528–1540. ACM, 2016.

[26] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7714–7722, 2019.

[27] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.

[28] Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pages 26693–26712. PMLR, 2022.

[29] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[33] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[34] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[35] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.