# DiffSwap-Tracker: A Robust Watermark for Tracing Diffusion-Based Face Swapping

Beijing Chen, Yuxuan Chen, Xin Liao, *Senior Member, IEEE*, and Zhangjie Fu, *Member, IEEE*

***Abstract*—Diffusion-based face swapping has become a key deepfake technology due to its high fidelity, posing significant threats to personal reputation and public security. However, existing watermarking methods are predominantly designed for GAN-based face swapping, and lack the robustness to be effective in this new diffusion-based face swapping scenario. To remedy it, a robust watermarking method named DiffSwap-Tracker is proposed for face image protection against diffusion-based face swapping. In the DiffSwap-Tracker, a novel dual-stream attack simulation mechanism is designed in the training process, enabling the watermark to support both tracing and forensics under face swapping attacks, as well as copyright confirmation in unauthorized usage scenarios. One stream simulates the diffusion-based face swapping attack through introducing a surrogate diffusion model into the training. Importantly, it develops a Phased Random Gradient Guidance (PRGG) strategy to reduce memory overhead by randomly selecting a limited number of consecutive denoising steps for gradient backpropagation. The other stream simulates common post-processing operations by introducing seven types of operations as well as none operation. Furthermore, an adaptive post-processing operation selection (APOS) strategy is designed to make the model training adaptively focus on the operations that are more difficult to resist, thereby improving overall robustness. Experimental results demonstrate that the proposed method can extract the watermark from swapped images with a bit error rate as low as 1.73%, achieving the watermarked faces with a PSNR of 43.66. Additionally, the method exhibits robust performance under various common post-processing operations, achieving bit error rates below 1.4% across different operations such as brightness adjustment, blurring, compression, and scaling, highlighting its effectiveness in copyright confirmation scenarios.**

***Index Terms*—Watermark, face swapping, diffusion model, copyright confirmation, forgery traceability.**

## I. Introduction

In recent years, the rapid development of Generative Adversarial Networks (GANs) [1] and diffusion models [2] has significantly advanced face swapping technology. Compared to GAN-based face swapping technologies, diffusion-based technologies offer significant advantages in image realism [2], [3], and demonstrate strong potentiality in semantic control and attribute preservation [4]. For example, Face-Adapter [5] is a popular diffusion-based face swapping method that introduces a lightweight facial editing adapter, which enables precise control over identity and attributes and effectively facilitating face reconstruction and swapping tasks. Consequently, the high-quality swapped faces generated by these technologies can confuse the false with the true. They not only pose serious threats to user privacy, but also have become weapons for a series of criminal activities such as interfering with politics, misleading public opinion, committing financial fraud, and defaming reputations. Addressing the threats posed by diffusion-based face swapping has therefore become a pressing challenge in the field of multimedia security.

To combat deepfake-related threats, existing defense strategies can be broadly categorized into passive detection and proactive defense. Passive detection aims to distinguish manipulated images from authentic ones by leveraging artifacts or inconsistencies in visual content. However, passive detection is a post-hoc defense and cannot prevent the harm caused by deepfakes. In contrast, proactive defense is a pre-hoc defense, protecting users' images from malicious forgery before harm occurs. Proactive defense includes two main approaches: watermarking-based and adversarial example-based defenses. Watermarking-based approach provides proactive defense by encoding a robust watermark representing identity information or unique content identification into the carrier image [6]. If an image is forged, users can extract the watermark to trace its source. Adversarial example-based approach relies on injecting perturbations into original images to disrupt model inference during malicious manipulations, but often suffers from poor robustness when subjected to common post-processing operations. In comparison, deep learning-based watermarking methods provide a more practical solution, offering better imperceptibility and generalizability in real-world scenarios. Therefore, this paper adopts the robust watermarking approach to defend against deepfakes.

In recent years, some watermarking works have been presented to address the detection and traceability of deepfakes, such as Sepmark [7], BiFPro [8], and Dual Defense [9]. However, these works are all specifically designed for GAN-based deepfakes and cannot be effectively transferred to diffusion-based deepfakes. Effective watermarking methods remain scarce. To bridge this gap, a robust watermarking method named DiffSwap-Tracker is proposed for protecting face images against diffusion-based face swapping and for

This paragraph of the first footnote will contain the date on which you submitted your paper for review, which is populated by IEEE. This work was supported by the National Natural Science Foundation of China under Grants 62572251, U22B2062 and U22A2030. *Corresponding author: Zhangjie Fu.* Beijing Chen and Yuxuan Chen are co-first authors.

Beijing Chen, Yuxuan Chen, and Zhangjie Fu are with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China. (e-mail: nbutimage@126.com; cyx19860918870@163.com; fzj@nuist.edu.cn).

Xin Liao is with College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. (e-mail: xinliao@hnu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

enabling copyright confirmation in scenarios of unauthorized usage. The main contributions of this paper are as follows:

- A robust watermarking method named DiffSwap-Tracker is proposed to defend against diffusion-based face swapping. This method designs a dual-stream attack simulation module to mimic both face swapping attack and common post-processing operations during training. To the best of our knowledge, this is the first watermarking work specifically designed for diffusion-based face swapping;
- The face swapping simulation stream introduces a surrogate diffusion-based model into the training and specially develops a Phased Random Gradient Guidance (PRGG) strategy to alleviate GPU memory overhead of the surrogate model. In the early phase of denoising process, a limited number of consecutive denoising steps are randomly selected to allow for gradient backpropagation in the latter half of the denoising process. Then, in the later phase, a similar random selection is made in the first half of the denoising steps;
- The post-processing simulation stream considers eight types of common operations including a none operation and specially develops an adaptive post-processing operation selection (APOS) strategy. This strategy improves robustness against various common post-processing operations, especially more challenging ones, by dynamically assigning a selection probability to each operation according to its difficulty of watermark extraction during each training step.
- Experimental results show that the DiffSwap-Tracker exhibits strong robustness against diffusion-based face swapping while maintaining the high visual quality of watermarked images. It effectively supports both forgery traceability and copyright confirmation.

The rest of this paper is organized as follows. Section II reviews related work on deep face swapping and watermarking against face forgery. Section III presents the framework of the proposed DiffSwap-Tracker, detailing the diffusion-based face swapping simulation and the common post-processing simulation streams. Section IV discusses the experimental results. Then, Section V introduces the limitations and potential social impact of our work, as well as the directions for future work. Finally, Section VI provides a summary of this paper.

## II. Related Work

### A. Deep Face Swapping

Face swapping aims to transfer the identity features from a source image to a target image while preserving other attributes of the target image—such as lighting, hairstyle, background, and pose—as much as possible. This technology has significant application in areas such as virtual avatar generation, film production, and entertainment content creation. Mainstream face swapping methods can be broadly categorized into two types: GAN-based methods and diffusion-based methods.

**GAN-based Methods.** Early face swapping research largely relied on GAN frameworks [10], [11], [12], [13], [14], [15], [16] due to their high generation efficiency and strong image synthesis capabilities of GANs. Among them, SimSwap [10] proposed an efficient and general face swapping framework that achieves identity transfer in the feature space via an identity injection module and introduces a weak feature matching loss to preserve target image attributes. To further enhance the geometric consistency of synthesized images, HifiFace [14] incorporated 3D shape modeling based on 3DMM [17] and proposed a "3D shape-aware identity control" strategy to improve realism while maintaining the structural identity from the source. Despite progress in preserving image consistency, GAN-based methods still suffer from artifacts and unnatural edge blending, particularly under challenging conditions such as extreme expressions, complex lighting, or occlusions [18].

**Diffusion-based Methods.** Given the remarkable performance of diffusion models in image generation tasks, they have quickly become a research focus in the field of face swapping. DiffSwap [18] treated the face swapping task as a conditional image inpainting problem, integrating identity and key point inputs to guide the generation process and thereby achieving precise control over identity and structure. Its proposed midpoint estimation strategy effectively alleviates training inefficiencies caused by multi-step sampling. To further reduce training costs and improve model adaptability, Face-Adapter [5] froze a pretrained diffusion model and trained only lightweight plug-in modules. It achieved precise face swapping through a decoupled control mechanism involving structural conditions, identity encoding, and attribute manipulation. In comparison to various Stable Diffusion models [3], the Face-Adapter significantly lowers training and inference costs while maintaining high-quality identity control and generation.

### B. Watermarking against Face Forgery

Invisible digital watermarking is a technique for embedding watermark information into digital media without significantly affecting the visual or perceptual quality of the content. It can serve various purposes, including copyright confirmation, content authentication, and usage tracking [6], [19], [20], [21], [22], [23], [24].

In response to the rise of deepfakes, recent research has focused on using the watermarking technology to defend against face forgery. These methods can be roughly divided into three categories: robust watermarking for forgery traceability, semi-fragile watermarking for forgery detection, and adversarial watermarking for disrupting forgery.

**Robust Watermarking for Forgery Traceability.** These methods focus on embedding robust watermark into face images to enable forgery traceability. For instance, Fake Tagger [25] utilizes a deep watermarking method to encode a watermark into the carrier face image, which allows for the effective tracking of manipulated face image. Artificial fingerprints [26] have validated their transferability from training data to model generation, enabling direct identification of the source of generated fake images and preventing the misuse of face swapping models.

**Semi-Fragile Watermarking for Forgery Detection.** These methods leverage the key characteristic of semi-fragile watermarks: they are robust to common operations but sensitive

to malicious operations such as facial forgery. This property is then applied to detect forgeries. For example, FaceGuard [27] and FaceSigns [28] proposed semi-fragile watermarking schemes that remain robust under typical pixel-level perturbations but become fragile when subjected to face manipulation, thereby enabling forgery detection and identity authentication. Building on the idea of semi-fragile watermarking, Sepmark [7] introduced a unified framework comprising an embedding encoder, a semi-fragile detector, and a robust decoder (Tracer), demonstrating strong performance in face tampering scenarios. BiFPro [8] distinguished between two face swapping cases, i.e., impersonation and victimization, by applying fragile and robust watermarking respectively for identity verification and forensic traceability.

**Adversarial Watermarking for Disrupting Forgery.** These methods embed adversarial watermarks to disrupt the output of forgery models. Dual-Defense [9] proposed an adversarial robust watermarking mechanism for the dual purpose of copyright tracking and forgery disruption. This was achieved by adopting an original feature fitting-based adversarial watermarking, complemented by a wavelet-domain structural compensation loss and channel attention mechanism.

However, all of the above-mentioned works only focus on GAN-based face swapping scenarios, and they exhibit limited robustness in the context of diffusion-based face swapping (see Table I for details). Therefore, this paper tries to develop a practical watermarking method for both traceability in diffusion-based face swapping and copyright confirmation after unauthorized usage, aiming to further advance the role of watermarking technology in secure image generation.

## III. METHOD

### *A. Problem Statement*

This work aims to protect face images from infringement scenarios such as diffusion-based face swapping and unauthorized image theft, enabling traceability under diffusion-based face swapping and copyright confirmation after unauthorized usage. To achieve this goal, this section first introduces the threat model to clarify the threats that attackers may pose. Then, in order to counter these threats, the mechanism of watermark embedding and the designed dual-stream attack simulation are introduced.

(a) Threat Model

This work assumes that attackers can access the watermarked image, but have no knowledge of the specific watermarking algorithm. The threat model considers two primary threat scenarios.

The first scenario is diffusion-based face swapping attack. Facial images are easily manipulated by malicious users to achieve improper purposes through face swapping. The process of face swapping is easy to disrupt the embedded watermark, as the facial region containing watermark information will be replaced during face swapping, making it difficult to extract complete watermark information from the swapped image.

The second scenario is common post-processing operations, which are relevant to copyright infringement and unauthorized usage. The attacker's actions on the watermarked images may be (i) intentional, aiming at degrading or erasing the watermark to circumvent copyright verification, such as blurring, cropping, color and brightness adjustment, etc., or (ii) unintentional, occurring as a standard part of an image processing procedure, such as compression during social media uploading, scaling during the image acquisition, etc.

(b) Embedding the User Identity Watermark

To enable both traceability in swapped faces and copyright confirmation after unauthorized usage, a watermark message $M$ containing user's identity information is embedded into the original face image $I_{origin}$ using a watermark embedding network $E_m$, generating a watermarked image $I_{wm}$ as:

$$I_{wm} = E_m(I_{origin}, M) \tag{1}$$

(c) Dual-Stream Attack Simulation

After embedding the user identity watermark, to ensure that the watermark remains extractable after both diffusion-based face swapping and common post-processing operations, a dual-stream attack simulation module is introduced during training to mimic realistic attack scenarios. The reason for considering post-processing operations is that the images received for copyright confirmation after unauthorized usage may undergo both non-malicious operations (e.g., compression, noise, scaling during transmission or storage) and malicious ones (e.g., smoothing, cropping aimed at watermark removal). The dual-stream attack simulation is defined as follows:

(i) Diffusion-based Face Swapping Attack Simulation. A diffusion-based face swapping model $F(\cdot)$ is applied to the watermarked source face $I_{wm}$ and a target face $I_{target}$, generating a swapped face $I_{swap}$:

$$I_{swap} = F(I_{wm}, I_{target}) \tag{2}$$

(ii) Common Post-Processing Operation Simulation. A set of post-processing operations $N(\cdot)$, including both non-malicious and malicious transformations, is applied to the watermarked image $I_{wm}$, producing a post-processed image $I_{postp}$:

$$I_{postp} = N(I_{wm}) \tag{3}$$

Therefore, the optimization objective is to jointly improve both the imperceptibility of the watermarked image and the accuracy of watermark extraction by optimizing the embedding network $E_m$ and the decoding network $D_m$. Specifically, the method minimizes three loss functions: $L_{impercept}$, which measures the visual differences between the original and watermarked images, and $L_{extract1}$, which quantifies the error rate in recovering the watermark message $M$ from diffusion-based swapped faces, and $L_{extract2}$, which measures the error under common post-processing operations. The overall optimization goal is defined as:

$$\underset{E_m, D_m}{arg\ min}\ E_I \begin{bmatrix} L_{impercept}\big(E_m(I_{origin}, M), I_{origin}\big) + \\ L_{extract1}\big(D_m(I_{swap}), M\big) + \\ L_{extract2}\big(D_m(I_{postp}), M\big) \end{bmatrix} \tag{4}$$

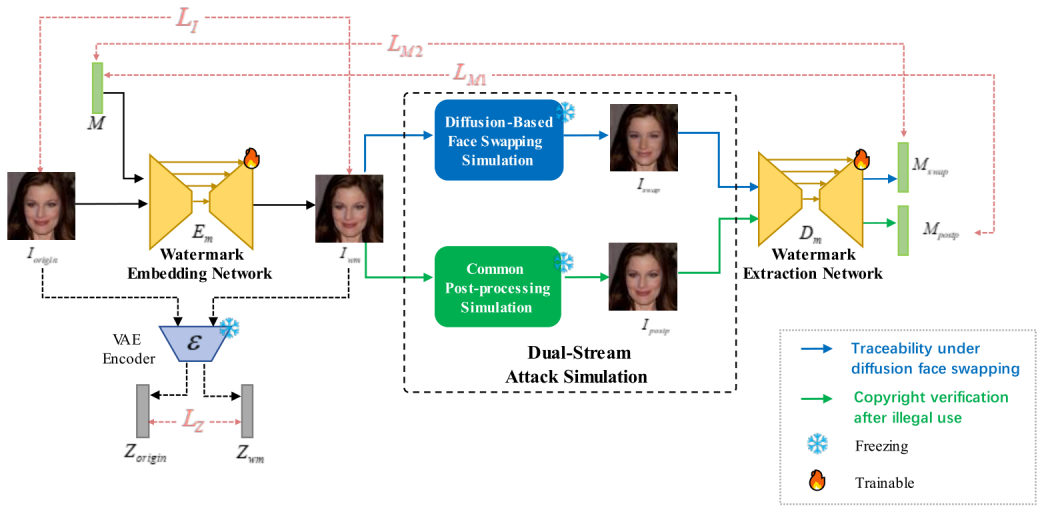


**Fig. 1.** Overall Framework of DiffSwap-Tracker.

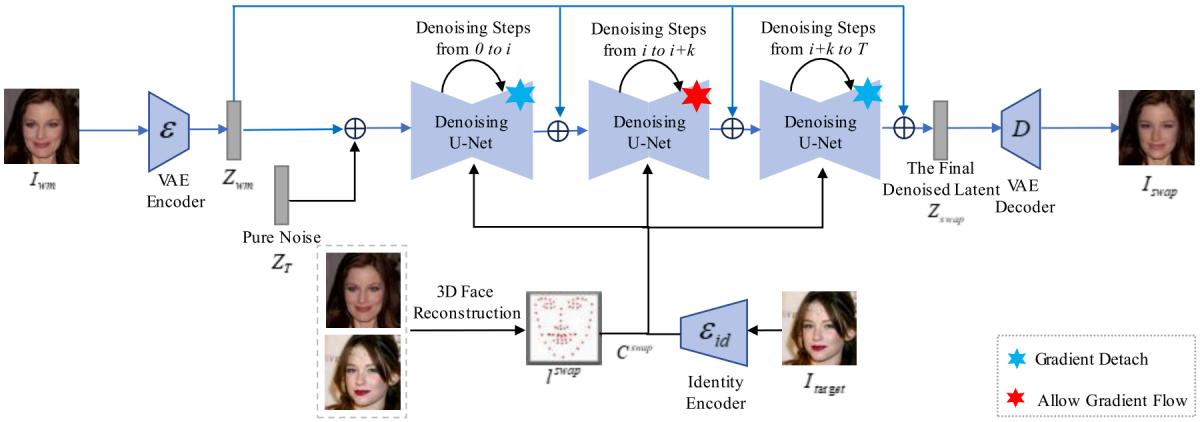


**Fig. 2.** Stream of Diffusion-based Face Swapping Simulation.

*B. Overall Framework*

Fig. 1 illustrates the framework of DiffSwap-Tracker, which is designed to defend against diffusion-based face swapping. The DiffSwap-Tracker consists of three primary modules: watermark embedding, dual-stream attack simulation, and watermark extraction. The dual-stream attack simulation is the core contribution and will be discussed in detail in Section III-C. A brief overview of each module is provided below.

Watermark Embedding Module: The watermark embedding module employs an embedding network $E_m$ to integrate the watermark information $M$ with the source face image $I_{origin}$, obtaining a watermarked image $I_{wm}$. The watermark message $M \in \{0,1\}^L$ is a binary vector of length $L$, representing the user's identity information. The embedding process effectively fuses $M$ into $I_{origin}$ using a U-Net [29] architecture.

Dual-Stream Attack Simulation Module: The dual-stream attack simulation module is designed to improve the robustness of the watermark against two major threats: diffusion-based face swapping and common post-processing. Therefore, one stream is designed for one threat. The frameworks of two streams are shown in Fig. 2 and Fig. 3, respectively. The details of these two streams will be shown in the following Section III.C. Furthermore, the PRGG strategy is designed in this

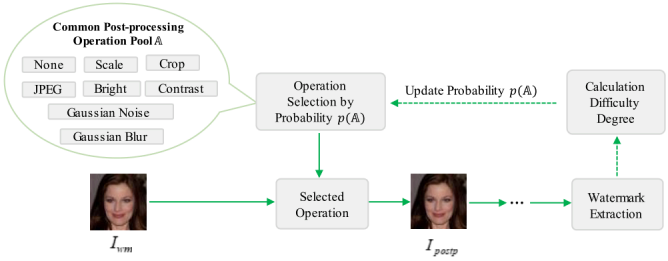


**Fig. 3.** Stream of Common Post-Processing Simulation.

module to address the GPU memory overhead of the diffusion-based face swapping simulation. It allows gradient backpropagation through only a random subset of denoising time steps. The details can be found in Section III-C. The dual-stream attack simulation enables the system to support traceability after face swapping and copyright confirmation after unauthorized usage.

Watermark Extraction Module: Under the dual-stream attack setting, the watermark extraction module is also bifurcated. When the protected image undergoes face swapping, the generated image $I_{swap}$ is input into the extraction network $D_m$ to recover the watermark message $M_{swap}$, serving as evidence for traceability. When the image is subjected to unauthorized usage or common attacks, the processed image $I_{postp}$ is fed into

$D_m$ to extract the watermark message $M_{postp}$ for copyright authentication. The decoder network $D_m$ is constructed using a series of residual blocks to improve reconstruction accuracy.

Pretraining: Prior to the full framework training, it is necessary to pretrain the model, where the attack simulation module is simplified to a single-stream variant that includes only common post-processing operations. Pretraining is necessary because the diffusion-based face swapping simulation stream requires the operations of face alignment and identity extraction on watermarked images, but untrained embedding network tends to produce poor-quality outputs that easily cause these operations to fail. Therefore, pretraining helps stabilize the performance of both the embedding and extraction modules before introducing face swapping simulations in the full training stage.

### *C. Dual-Stream Attack Simulation*

(a) Diffusion-Based Face Swapping Simulation

To simulate diffusion-based face swapping attacks, Face-Adapter [5] is adopted as a surrogate diffusion model. Given a watermarked source face image $I_{wm}$ and a target face image $I_{target}$, the model outputs a swapped face $I_{swap}$.

However, integrating a diffusion-based face swapping model into training requires enabling gradient backpropagation through consecutive denoising steps, which imposes a significant memory burden and increases a large amount of training time. To address this issue, Robust-Wide [30] proposed a strategy of allowing gradient backpropagation only during a few fixed denoising steps in the later phase of instruction-driven image editing. Although this strategy ensures the embedding of watermark details in the later phase of denoising, it tends to neglect high-level semantic features captured in the early phase [2], [3], [31]. Moreover, Robust-Wide is specifically designed for instruction-based editing and lacks robustness when applied to diffusion-based face swapping. To overcome these limitations, an improved strategy named Phased Random Gradient Guidance (PRGG) is designed.

The PRGG strategy selects time steps for gradient backpropagation in two phases with the order of the later phase and early phase of the denoising process. This enhances the watermark model's learning of both fine-grained local textures and high-level semantics. In addition, the PRGG strategy randomly selects $k$ consecutive steps in each phase to allow gradient backpropagation, which enables the watermark model to learn a wide range of denoising steps in diffusion-based face swapping models. Specifically, during the training of the watermarking model, assuming the total number of denoising time steps in the diffusion process is $T$, and gradient backpropagation is enabled from steps $i$ to $i+k$. The PRGG strategy selects time steps in the following two phases:

(i) Early phase. In each denoising iteration, $k$ consecutive steps are randomly selected within the second half of the denoising time steps (i.e., $i \in (\frac{T}{2}, T)$) to allow gradient backpropagation. This ensures the embedding of watermark signals into fine-grained local textures and details.

(ii) Later phase. Similarly, $k$ consecutive steps are randomly selected within the first half of the timeline (i.e., $i \in (0, \frac{T}{2})$) to enable gradient backpropagation, promoting the embedding of watermark features into high-level semantic representations.

Fig. 2 depicts the framework of the diffusion-based face swapping simulation. The Face-Adapter [5] model comprises a VAE [32], a U-Net [29], and an identity encoder [5], all of which are kept frozen during training. The watermarked image $I_{wm}$ is first encoded into latent features $Z_{wm}$ via the VAE encoder $\varepsilon$. These features combined with pure noise $Z_T$ are then input into the U-Net for denoising. Here, gradients are truncated in steps prior to $i$, while steps from $i$ to $i+k$ allow gradient backpropagation. Subsequent steps are again detached. After this process, the final denoised latent $\mathrm{Z_{swap}}$ is obtained and decoded by the VAE decoder $D$ into the swapped image $I_{swap}$. Additionally, Face-Adapter uses identity features extracted from the target face by the identity encoder and facial masks generated from 3D face reconstruction (using both the source and target images) as conditional inputs $C_{swap}$, guiding the entire sampling process.

(b) Common Post-processing Simulation

Fig. 3 illustrates the framework for simulating common post-processing operations. These operations reflect real conditions that images submitted for copyright confirmation may undergo both non-malicious operations, such as compression, noise, or rescaling caused by storage or transmission, as well as malicious operations aimed at watermark removal, including blurring and cropping. To enhance the watermark's robustness under such conditions, these operations are simulated as one stream of the dual-stream attack module during training. However, the model's robustness against different operations is different. Therefore, in order to improve the overall robustness, an APOS strategy is designed. The core idea is that the watermark model will evaluate which operation is the most difficult during each training step, and dynamically prioritize the simulation of such attacks. This forces the watermark model to continuously strengthen its weakest defense, thereby improving its overall robustness. In addition, the success rate of watermark extraction is used to measure the difficulty degree, and the Exponential Moving Average (EMA) [33] algorithm is used to smooth the changing trend of the success rate over training, thereby avoiding the influence of sudden changes in any single training step.

Let $\mathbb{A} = \{A_1, A_2, \dots, A_K\}$ be a set of $K$ post-processing operations. For each operation $A_i$ at each training step $t$, the EMA of its success rate $s_i(t)$ is calculated by:

$$\hat{s}_i(t) = (1-\beta)\hat{s}_i(t-1) + \beta s_i(t) \tag{5}$$

where $\beta \in (0,1]$ is the EMA decay factor.

Then, a "difficulty degree" $g_i(t)$ is defined by using the failure rate, i.e., $1-\hat{s}_i(t)$, as:

$$g_i(t) = \left(1-\hat{s}_i(t)\right)^{\alpha} \tag{6}$$

Here, the failure rate is amplified by an exponent $\alpha$ (>1) to increase the gap between the difficulty degrees of difficult and easy operations.

Finally, the probability $p_i(t)$ for selecting operation $A_i$ at the next step is calculated by adding its minimum sampling probability $\theta$ to the normalized score $g_i(t)$:

$$p_i(t) = \theta + (1 - K\theta)\frac{g_i(t)}{\sum_{j=1}^{K} g_j(t)} \quad (7)$$

During training, the operations are sampled according to these dynamically updated probabilities.

*D. Loss Functions*

The proposed method aims to ensure the visual quality of the watermarked images, the robustness of watermark extraction after diffusion-based face swapping, and the resilience to common post-processing operations. To this end, three specific types of loss functions are designed as follows:

(a) Visual Quality Loss of Watermarked Images

To ensure that the watermarked image $I_{wm}$ remains visually similar to the original face image $I_{origin}$, the constraints are imposed on both the latent feature space and the pixel space:

(i) Latent Space Constraint. Both $I_{origin}$ and $I_{wm}$ are passed through the VAE encoder $\varepsilon$ of Face-Adapter to obtain their corresponding latent features $Z_{origin}$ and $Z_{wm}$. An L2 loss is applied to these features:

$$\begin{aligned} L_Z &= L_2\left(Z_{origin}, Z_{wm}\right) \\ &= L_2\left(\varepsilon\left(I_{origin}\right), \varepsilon\left(E_m\left(I_{origin}, M\right)\right)\right) \end{aligned} \quad (8)$$

(ii) Pixel Space Constraint. An additional pixel-level L2 loss is applied between $I_{origin}$ and $I_{wm}$:

$$L_I = L_2\left(I_{origin}, I_{wm}\right) \quad (9)$$

Then, the total visual quality loss is then:

$$L_{impercept} = \lambda_1 L_I + \lambda_2 L_Z \quad (10)$$

(b) Watermark Reconstruction Loss after Face Swapping

To ensure that the embedded watermark information $M$ can be accurately extracted from the swapped face $I_{swap}$ ,a reconstruction loss is introduced based on Mean Squared Error (MSE) between the original message $M$ and the extracted message $M_{swap}$ from the decoding network $D_m$:

$$L_{extract1} = MSE\left(M, M_{swap}\right) = MSE\left(M, D_m\left(I_{swap}\right)\right) \quad (11)$$

(c) Watermark Reconstruction Loss after Post-Processing

To enhance the robustness of watermark extraction against post-processing attacks, a similar MSE loss is applied between the original message $M$ and the message $M_{postp}$ extracted from the post-processed image $I_{postp}$:

$$L_{extract2} = MSE\left(M, M_{postp}\right) = MSE\left(M, D_m\left(I_{postp}\right)\right) \quad (12)$$

(d) Total Loss

The final training objective combines the above-mentioned loss terms with appropriate weighting factors $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. The total loss is defined as:

$$\begin{aligned} L_{total} &= L_{impercept} + \lambda_3 L_{extract1} + \lambda_4 L_{extract2} \\ &= \lambda_1 L_I + \lambda_2 L_Z + \lambda_3 L_{extract1} + \lambda_4 L_{extract2} \end{aligned} \quad (13)$$

This total loss is minimized to jointly optimize the watermark embedding network $E_m$ and the extraction network $D_m$ as shown in Eq. (4).

## IV. Experimental Results and Analysis

*A. Experimental Settings*

**Datasets.** This study primarily utilizes two face image datasets: CelebA-HQ [34] and FFHQ [35]. In addition, real-world facial dataset Wider Face [36] was used for compression experiments and forensic integrity experiments. CelebA-HQ is a high-quality version of CelebA [37], containing 30,000 high-resolution face images, which are used for training. FFHQ is a more diverse dataset with 70,000 1024×1024 images, covering a broader range of age, pose, lighting, and background variations. In this paper, 1,000 images are randomly selected from FFHQ as the test set. Since the training and testing sets are from different datasets, the cross-dataset generalization is actually tested. The Wider Face dataset is a benchmark dataset for face detection. This dataset contains a large number of highly variable faces in terms of proportion, posture, and occlusion. All images are resized to 512×512 resolution for consistency during experiments.

**Implementation Details.** All experiments are implemented using PyTorch on a workstation equipped with an H800 GPU. During pretraining of the watermark embedding and extraction networks, the AdamW optimizer is used, with a cosine annealing learning rate scheduler and a 400-step warm-up, the learning rate of 1e-3, batch size of 1, and a total of 10,000 training steps. 1,000 64-bit binary sequences, simulating user identity information, are randomly generated as watermark messages for 1,000 testing cover images. In the main training stage, the AdamW optimizer is continued with the same set. The learning rate decays to 1e-4, the batch size remains at 1, and the training proceeds for 20,000 steps. The loss function weights are set as follows: $\lambda_1 = 1.0, \lambda_2 = 0.001, \lambda_3 = 1.5, \lambda_4 = 1.5$.

For diffusion-based face swapping simulation, the Euler sampler is employed with 50 inference steps, where the text


**Fig. 4.** Examples from three experimental datasets.

guidance scale $sT = 5.0$ and the image guidance scale $sI = 1.5$. The value of $k$ (the number of steps allowing gradient backpropagation) is set to 4 during training. For specific selection details, please refer to Section IV-D. Details of common post-processing operation simulation are as follows: Brightness & Contrast with intensity randomly selected in the range $[0.8,1.2]$; Gaussian Blur with kernel size 5×5 and standard deviation 0.5; Gaussian Noise with standard deviation 0.05; JPEG Compression quality gradually reduced from 100 to 50; Cropping and Scale random strength between 0.6 to 1.0. These setting are also used for the following robustness experiment. In the adaptive post-processing operation selection strategy, the EMA decay factor $\beta = 0.1$, the failure rate amplification factors $\alpha = 2.0$, and the minimum sampling probability $\theta$ for each operation is 0.01. Note that, the $\theta$ is set to 0.1 for the no operation scenario to ensure the performance under such important scenario.

**Baseline Methods.** Five recent watermarking methods are introduced as baselines for comparison: MBRS [20], CIN [38], PIMoG [39], Sepmark [7], and Robust-Wide [30]. All baselines are implemented using their official code. Since CIN only supports a 128×128 input resolution, the watermarked image is first up-sampled to 512×512 for face swapping, then down-sampled to 128×128 for watermark extraction. PIMoG has no public model weights, so it is retrained with its official source code. For Sepmark, watermark extraction after diffusion-based face swapping is performed using its Tracer decoder.

**Metrics.** PSNR and SSIM are used to evaluate the visual quality of the watermarked images: PSNR reflects pixel-level error, with higher values indicate better image quality. SSIM measures structural similarity on a scale of [0,1] where higher SSIM are better.

To assess watermark extraction accuracy, Bit Error Rate (BER) is adopted as the core metric. It is defined as:

$$BER(M_E, M_D) = \frac{1}{L}\sum_{i=1}^{L}\left(M_E^i \neq M_D^i\right) \times 100\% \qquad (14)$$

where $M_E = \{M_E^1, M_E^2, \ldots, M_E^L\}$ represents the embedded watermark, and $M_D = \{M_D^1, M_D^2, \ldots, M_D^L\}$ is the extracted watermark, $M_E^i, M_D^i \in \{0,1\}$, and $L$ is the watermark length. A BER close to 0 indicates high watermark extraction accuracy.

Furthermore, to evaluate the forensic reliability and prevent false accusations, the False Positive Rate (FPR) is introduced. The *FPR* is defined as the proportion of unwatermarked samples that are incorrectly identified as watermarked. To this end, the extracted message from each unwatermarked sample is compared with each one in the set of watermarks (1,000 randomly generated watermarks in this paper), and the case with the lowest BER is selected for judgment. If the *BER* in this case is no more than the predefined decoding threshold $\tau_{ber}$, the corresponding unwatermarked sample is counted as a False Positive (FP). Conversely, if the *BER* exceeds $\tau_{ber}$, the sample is correctly identified as unwatermarked, constituting a True Negative (TN). The *FPR* is then calculated by:

TABLE I
QUANTITATIVE COMPARISON UNDER DIFFUSION-BASED FACE SWAPPING ATTACKS (↓:LOWER IS BETTER; ↑:HIGHER IS BETTER; BOLD: BEST RESULT; UNDERLINE: SECOND BEST)

| Methods | BER(%)↓ | | PSNR↑ | SSIM↑ |
|---|---|---|---|---|
| | w/o Face Swapping | w/ Face Swapping | | |
| MBRS[20] | **0.0000** | 45.4311 | **43.9780** | 0.9870 |
| CIN[38] | **0.0000** | 47.2796 | 35.3183 | 0.9212 |
| PIMoG[39] | **0.0000** | 44.8694 | 35.3183 | 0.9212 |
| Sepmark[7] | **0.0000** | <u>14.6656</u> | 36.4341 | 0.9194 |
| Robust-Wide[30] | **0.0000** | 15.6406 | 41.9142 | **0.9910** |
| DiffSwap-Tracker | **0.0000** | **1.7319** | <u>43.6628</u> | <u>0.9870</u> |

$$FPR = \frac{FP}{FP + TN} \qquad (15)$$

A low *FPR* is crucial for ensuring the integrity and trustworthiness of the watermarking system.

### *B. Performance against Diffusion-based Face Swapping*

Table I shows the quantitative results of the proposed method and five baseline methods in terms of effectiveness and imperceptibility under diffusion-based face swapping attacks. It can be observed that all baseline methods suffer from a BER higher than 14.67% after the diffusion-based face swapping process, indicating their limited ability to withstand such attacks, while the proposed DiffSwap-Tracker maintains BER within 1.73%, demonstrating superior robustness. This is attributed to the dual-stream attack simulation strategy in the DiffSwap-Tracker, particularly the simulation of diffusion-based face swapping during training. Furthermore, the DiffSwap-Tracker maintains high PSNR and SSIM scores, and its embedded watermark is imperceptible. Please find the following qualitative experiments and visual results in Fig. 5.

Then, Fig. 5 shows the qualitative results of six compared methods. It can be observed from the residual images that DiffSwap-Tracker has few visual artifacts and has highly competitive image quality compared to the other five methods. In addition, it can be seen from Fig. 6 that there is not much difference in the face swapping results between the watermarked image generated by DiffSwap-Tracker and the origin image, indicating that the DiffSwap-Tracker does not affect the normal face synthesis process.

In order to test the DiffSwap-Tracker's cross-model generalization, seven face swapping models are evaluated in this experiment, i.e., the diffusion-based models DiffSwap [18], IP-Adapter [40], REFace [41], and Face-Adapter [5] (surrogate

TABLE II

CROSS-MODEL GENERALIZATION OF DIFFSWAP-TRACKER UNDER DIFFERENT FACE SWAPPING MODELS IN TERMS OF BER (%) (※ INDICATES THE SURROGATE MODEL IN TRAINING)

| SimSwap | HifiFace | FaceShifter | DiffSwap | IP-Adapter | REFace | Face-Adapter※ |
|---|---|---|---|---|---|---|
| 5.6503 | 1.8974 | 2.4718 | 0.9981 | 0.9453 | 1.4437 | 1.7319 |

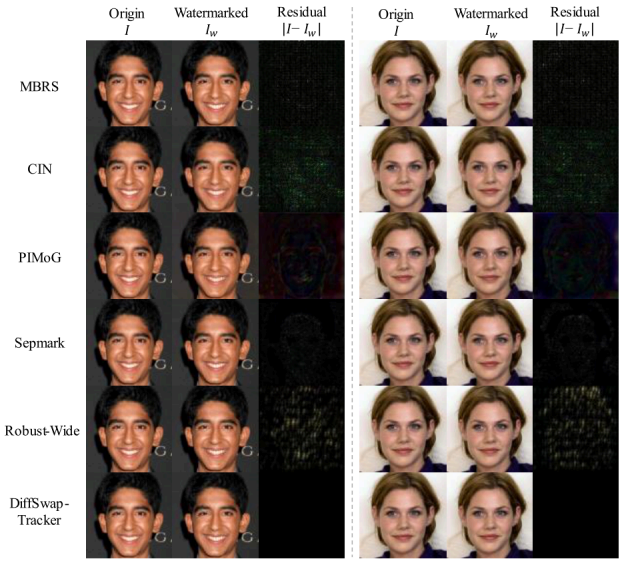


**Fig. 5.** Visual comparison of watermarked images.

model), as well as the GAN-based method SimSwap [10], HifiFace [14] and FaceShifter [42]. The results are presented in Table II. Notice that the five baselines are not compared with the DiffSwap-Tracker here because Table I has shown that they are generally not robust to diffusion-based face swapping, making it meaningless to test their cross-model generalization. The results show that: (a) Although Face-Adapter is used as a surrogate model, the DiffSwap-Tracker also achieves low BERs in the other six models, demonstrating its strong cross-model generalization capability; (b) Although the performance drops a little when transferred to the GAN-based models, the BER remains low, especially for the models HifiFace [14] and FaceShifter [42]. The drop is mainly attributed to the differences between the theoretical foundations of GANs and diffusion models; (c) The BERs of the other three diffusion-based models are even lower than that of the surrogate model Face-Adapter. A possible explanation is that Face-Adapter adopts a more complex and refined guidance condition

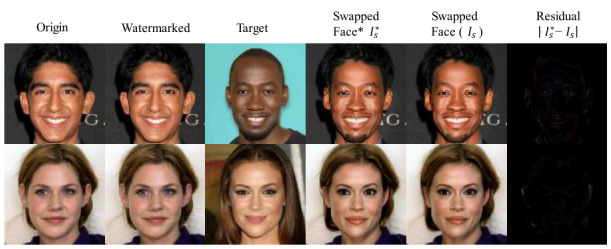


**Fig. 6.** Face swapping results of the origin image and the watermarked image generated by DiffSwap-Tracker (* indicates the results of watermarked image).

TABLE III

FORENSIC INTEGRITY VERIFICATION OF DIFFSWAP-TRACKER IN TERMS OF FPR (THE THRESHOLD $\tau_{ber}$ IS SET TO 20%)

| Datasets | No Noise | Gaussian Noise | Salt and Pepper Noise |
|---|---|---|---|
| FFHQ | 0.0000 | 0.0000 | 0.0000 |
| Wider Face | 0.0000 | 0.0010 | 0.0000 |

compared to the other three models. The Face-Adapter uses a specially designed spatial condition generator, identity encoder, and attribute controller to improve motion control accuracy, ID retention capability, and generation quality, which is also why the Face-Adapter is selected as the surrogate model.

In addition, Table III presents an experiment of the forensic integrity and reliability of DiffSwap-Tracker. In this experiment, 1,000 images are randomly selected from the FFHQ and Wide Face datasets, respectively. In addition, Gaussian noise with standard deviation 0.05 and salt-and-pepper noise with density 0.05 are added to these images for testing performance under different conditions. It can be seen from the results in Table III that the FPR remains at 0 or very close to 0 under different datasets and conditions. These results strongly confirm the forensic integrity of the DiffSwap-Tracker and ensure that it does not pose a risk of misidentifying unwatermarked content in practical applications.

### *C. Performance against Common Post-Processing Operations and Real-World Compression Scenarios*

Table IV shows the watermark extraction performance of the DiffSwap-Tracker and five baselines under various common post-processing operations, simulating unauthorized usage or copyright infringement. For most of these operations, DiffSwap-Tracker achieves BER that are highly competitive with the baseline methods, especially for three operations with no loss in the BER, remaining 0.000. This confirms its effectiveness in standard copyright protection contexts. In addition, the DiffSwap-Tracker is also robust to the diffusion-based face swapping attack as shown in Table I. It is not the case for the baselines.

In addition, the more challenging Wider Face dataset is used to test the applicability of the watermarking method in the real world. 100 facial images with complex lighting and partial occlusion are selected to test the effectiveness in complex scenes. In order to simulate the transmission of images on the social media platforms, the watermarked image is uploaded to X (Twitter) and MicroBlog for compression, and then downloaded for watermark extraction. As shown in Table V, the watermark can be extracted almost perfectly on both the original watermarked images and the

TABLE IV
QUANTITATIVE COMPARISON UNDER DIFFERENT POST-PROCESSING OPERATIONS IN TERMS OF BER(%)

| Methods | Bright | Contrast | JPEG | Gaussian Noise | Gaussian Blur | Scale | Cropping |
|---|---|---|---|---|---|---|---|
| MBRS[20] | 0.0047 | 0.0379 | 3.5518 | 0.0002 | **0.0000** | 0.0617 | 1.5240 |
| CIN[38] | 0.0064 | 0.0027 | 1.2513 | 0.0010 | 0.0001 | 0.0352 | **0.8362** |
| PIMoG[39] | 0.0021 | 0.0315 | 3.6209 | 0.0002 | **0.0000** | 0.0593 | 2.8624 |
| Sepmark[7] | 0.0152 | **0.0026** | 2.4280 | 0.0002 | **0.0000** | **0.0058** | 2.1758 |
| Robust-Wide[30] | 0.0268 | 0.0150 | 4.0241 | 0.0012 | **0.0000** | 0.0816 | 3.6408 |
| DiffSwap-Tracker | **0.0000** | 0.0099 | **1.1899** | **0.0000** | **0.0000** | 0.0269 | 1.3746 |

TABLE V
COMPRESSION ROBUSTNESS ON WIDER FACE DATASET IN TERMS OF BER(%)

| No Compression | JPEG | MicroBlog | X (Twitter) |
|---|---|---|---|
| 0.0961 | 1.1899 | 0.9731 | 1.0478 |

compressed ones. One possible reason is that there are some commonalities among different compression processes, and the robustness to JPEG compression obtained through the simulation in training has been successfully transferred to other compression processes. These experimental results demonstrate the reliability of the DiffSwap-Tracker in real-world applications.

*D. Ablation Study*

**Impact of Diffusion-Based Face Swapping Simulation.** Table VI evaluates the impact of incorporating the diffusion-based face swapping simulation into the training process. Here, the "One Stream Model without Face-Swapping Simulation" means no face swapping simulation. From Table VI, it can be seen that the watermarking method using face swapping simulation has significantly improved the robustness against diffusion-based face swapping attack. It confirms the effectiveness of incorporating face swapping simulation into training. Meanwhile, the significant improvement in robustness is accompanied by a trade-off in training resources. Since the diffusion-based face swapping model needs to complete multi-step denoising during face swapping requiring a large number of complex calculations, integrating it into training increases the total training time and GPU memory usage. However, this trade-off is valuable due to the great threats posed by diffusion-based face swapping and the weak robustness without considering the face swapping simulation as shown in Table VI. Furthermore, Table VI shows that the inference time that users are more concerned about does not increase. This indicates that the trained model can be effectively deployed to practical use cases without more time burden.

**Impact of Gradient Backpropagation with Different Number of Steps *k*.** Table VII compares the performance of the watermarking model under different values of $k$, which is the number of consecutive denoising steps that allow for the gradient backpropagation in the PRGG strategy. The experimental results show that with the increase of $k$, the BER first decreases and then increases, reaching its minimum value at $k = 4$. This shows that although more back propagation steps are initially beneficial for watermarking extraction, too many steps may be counterproductive. One possible explanation is that the watermark model obtains sufficient knowledge from diffusion-based face swapping model to learn robust feature embedding when $k$ increases to 4. Certainly, the image quality decreases slightly when $k = 4$. However, the great values of PSNR and SSIM demonstrate the watermark imperceptibility. It is also verified by the visual results in Fig. 5. Therefore, this paper selects $k = 4$ when training the watermark model because it maintains the best robustness with the smallest BER while maintaining watermark imperceptibility.

**Effect of Different Gradient Guidance Strategies.** Table VIII evaluates the performance of six different gradient guidance strategies in the simulation of diffusion-based face swapping with a fixed setting $k = 4$. Here, the strategy "Fixed Gradient Backpropagation" of the method Robust-Wide [30] selects the

TABLE VI
EFFECTIVENESS OF FACE SWAPPING SIMULATION

| Methods | BER(%)↓ | PSNR↑ | SSIM↑ | Training Time↓ | Inference Time↓ | GPU Usage↓ |
|---|---|---|---|---|---|---|
| One Stream without Face-Swapping Simulation | 20.5902 | **55.2247** | **0.9981** | **3h** | 121ms | **13257MiB** |
| Dual-Stream Attack Simulation | **1.7319** | 43.6628 | 0.9870 | 10.5h | **120ms** | 33745MiB |

TABLE VII
IMPACT OF GRADIENT BACKPROPAGATION WITH DIFFERENT NUMBER OF STEPS $k$

| $k$ | BER(%)↓ | PSNR↑ | SSIM↑ |
|---|---|---|---|
| 1 | 5.7952 | **45.8514** | **0.9931** |
| 2 | 3.9687 | 45.2683 | 0.9897 |
| 3 | 2.1667 | 44.3207 | 0.9847 |
| 4 | **1.7319** | 43.6628 | 0.9870 |
| 5 | 2.3174 | 44.4637 | 0.9884 |

TABLE VIII
IMPACT OF DIFFERENT GRADIENT GUIDANCE STRATEGIES

| Methods | BER(%)↓ | PSNR↑ | SSIM↑ |
|---|---|---|---|
| Fixed Gradient Backpropagation | 6.8521 | 42.1688 | 0.9836 |
| Random Non-Consecutive | 3.6625 | 42.5394 | 0.9839 |
| Random Later Phase Only | 4.0159 | 42.9641 | 0.9847 |
| Random Early Phase Only | 4.1090 | 43.1687 | 0.9854 |
| PRGG | **1.7319** | **43.6628** | **0.9870** |

last $k$ sampling steps' gradients to backpropagation; the strategy "Random Non-Consecutive" randomly selects $k$ non-consecutive steps across the whole denoising process; the strategy "Random Later Phase Only" randomly selects $k$ consecutive steps from the second half of the denoising process; the "Random Early Phase Only" randomly selects $k$ consecutive steps from the first half of the denoising process; the PRGG is t he proposed strategy that selects $k$ consecutive steps in the order of the second half and first half of the denoising process.

It can be observed from Table VIII that: Firstly, compared with the "Fixed Gradient Backpropagation" strategy, "Random Later Phase Only" strategy introducing randomness further improves robustness. The reason is that random selecting gradient steps helps the watermark model learn over almost the whole of the later phase, equivalent to simulating the later phase of the surrogate model; Secondly, the proposed PRGG strategy achieves the best robustness with the lowest BER while maintaining good imperceptibility. This is because single-phase random strategies fail to help watermarking model balance both texture- and semantic-level embedding, whereas PRGG guides the watermark model to focus on texture-level embedding in early phase and on semantic-level embedding in later phase.

**Effect of APOS Strategy.** Table IX compares the performance of the standard model with equal probability sampling for all post-processing operations (W/o APOS) and the model with APOS strategy (W/ APOS). The results in Table IX clearly indicate the significant effect of APOS strategy. The most significant improvement is under JPEG compression operation, where the adaptive method reduces the BER from 5.0991% to 1.1899%. It attributes to the continuous selection of JPEG compression in the training process because JPEG compression is the most difficult operation (please refer to the results of "W/o APOS").

TABLE IX
EFFECT OF APOS STRATEGY IN TERMS OF BER(%)

| Operations | W/o APOS | W/ APOS |
|---|---|---|
| Bright | **0.0000** | **0.0000** |
| Contrast | 0.0107 | **0.0099** |
| JPEG | 5.0991 | **1.1899** |
| Gaussian Noise | **0.0000** | **0.0000** |
| Gaussian Blur | **0.0000** | **0.0000** |
| Scale | 0.0285 | **0.0269** |
| Cropping | 1.4864 | **1.3746** |

## V. LIMITATIONS AND SOCIAL IMPACT

Although using powerful facial watermarking technology is beneficial for security, it is worth carefully discussing its limitations and potential social impacts.

Regarding the social impact, since embedding any message into a person's facial image may modify his biometric information, if this technology is deployed on a large scale without explicit user approval, it will raise ethical concerns about privacy. In addition, the embedding of mandatory watermarks may turn protection tools into monitoring or censorship tools, resulting in a dual dilemma.

Regarding the technical limitations, the proposed method mainly targets diffusion swapping attacks and conventional post-processing operations. More professional attackers may design an adaptive watermark erasure attack, which can learn the characteristics of watermark embedding and erase it, posing a significant challenge to the robustness of watermark models. Perhaps the robustness of watermarks against unknown attacks can be enhanced by adversarial training schemes or introducing diffusion denoising surrogate models to simulate watermark erasure. In addition, the proposed method also has some weaknesses, such as robustness against extreme cropping (the BER reaches 15.16% when cropping ratio 50%), and the long training time after the introduction of diffusion-based face swapping simulation.

Therefore, future work will be dedicated to addressing these challenges. Resolving these technical and ethical issues is crucial for the steady development of watermarking technology and social stability.

## VI. CONCLUSIONS

This paper proposes the DiffSwap-Tracker for face images to defend against diffusion-based face swapping attacks. By introducing a dual-stream attack simulation during training, the method enables both traceability under face swapping attacks and copyright confirmation after unauthorized usage. To

address the challenge of high memory consumption in training, the Phased Random Gradient Guidance (PRGG) strategy is designed to alleviate heavy memory overhead in simulation diffusion-based face swapping and enhance the robustness of the watermarking model under diffusion-based face swapping scenarios. Furthermore, to enhance the overall robustness against various common post-processing operations, the Adaptive Post-processing Operation Selection (APOS) strategy is developed to make the model training focus on the more difficult operations. The experimental results verify the superior performance of DiffSwap-Tracker under both diffusion-based face swapping attacks and common post-processing operations.

## REFERENCES

[1] I. Goodfellow *et al*., "Generative adversarial networks," *Commun. ACM,* vol. 63, no. 11, pp. 139-144, 2020.

[2] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv:2010.02502,* 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*(*CVPR*), 2022, pp. 10684-10695.

[4] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.* (*NeurIPS*), 2021, vol. 34, pp. 8780-8794.

[5] Y. Han *et al.*, "Face-adapter for pre-trained diffusion models with fine-grained id and attribute control," in *Proc. Eur. Conf. Comput. Vis.* (*ECCV*), 2024, pp. 20-36.

[6] Z. Meng, B. Peng, and J. Dong, "Latent watermark: Inject and detect watermarks in latent diffusion space," *IEEE Trans. Multimedia,* 2025. DOI: 10.1109/TMM.2025.3535300.

[7] X. Wu, X. Liao, and B. Ou, "SepMark: Deep separable watermarking for unified source tracing and deepfake detection," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 1190-1201.

[8] H. Liu *et al*., "BiFPro: A bidirectional facial-data protection framework against DeepFake," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 7075-7084.

[9] Y. Zhang *et al*., "Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping," *IEEE Trans. Inf. Forensics Secur.,* vol. 19, pp. 4628-4641, 2024.

[10] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2003–2011.

[11] J. Kim, J. Lee, and B.T. Zhang, "Smooth-swap: A simple enhancement for face-swapping with smoothness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2022, pp. 10779-10788.

[12] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*(*ICCV*), 2019, pp. 7184-7193.

[13] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (*CVPR*), 2020, pp. 5074--5083.

[14] Y. Wang *et al*., "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *arXiv:2106.09965,* 2021.

[15] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (*CVPR*), 2022, pp. 7642-7651.

[16] W. Xie, W. Lu, Z. Peng, and L. Shen, "Consistency preservation and feature entropy regularization for GAN based face editing," *IEEE Trans. Multimedia,* vol. 25, pp. 8892-8905, 2023.

[17] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 157-164.

[18] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (*CVPR*), 2023, pp. 8568-8577.

[19] Q. Guan, P. Liu, W. Zhang, W. Lu, and X. Zhang, "Double-layered dual-syndrome trellis codes utilizing channel knowledge for robust steganography," *IEEE Trans. Inf. Forensics Secur.,* vol. 18, pp. 501-516, 2022.

[20] Z. Jia, H. Fang, and W. Zhang, "MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 41-49.

[21] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proc. Eur. Conf. Comput. Vis.* (*ECCV*), 2018, pp. 657-672.

[22] H. Fang, Z. Jia, Y. Qiu, J. Zhang, W. Zhang, and E.-C. Chang, "De-END: Decoder-driven watermarking network," *IEEE Trans. Multimedia*, vol. 25, pp. 7571-7581, 2023.

[23] C. Qin, X. Li, Z. Zhang, F. Li, X. Zhang, and G. Feng, "Print-camera resistant image watermarking with deep noise simulation and constrained learning," *IEEE Trans. Multimedia*, vol. 26, pp. 2164-2177, 2024.

[24] S. Wu, W. Lu, and X. Luo, "Robust watermarking based on multi-layer watermark feature fusion," *IEEE Trans. Multimedia*, 2025. DOI: 10.1109/TMM.2025.3543079.

[25] R. Wang, F. Juefei-Xu, M. Luo, Y. Liu, and L. Wang, "FakeTagger: Robust safeguards against DeepFake dissemination via provenance tracking," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3546–3555.

[26] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (*ICCV*), 2021, pp. 14448-14457.

[27] Y. Yang, C. Liang, H. He, X. Cao, and N. Z. Gong, "Faceguard: Proactive deepfake detection," *arXiv:2109.05673,* 2021.

[28] P. Neekhara, S. Hussain, X. Zhang, K. Huang, J. McAuley, and F. Koushanfar, "FaceSigns: Semi-fragile neural watermarks for media authentication and countering deepfakes," *arXiv:2204.01960,* 2022.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.* (*MICCAI*), 2015, pp. 234-241.

[30] R. Hu, J. Zhang, T. Xu, J. Li, and T. Zhang, "Robust-wide: Robust watermarking against instruction-driven image editing," in *Proc. Eur. Conf. Comput. Vis.* (E*CCV*), 2024, pp. 20-37.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.* (*NeurIPS*), 2020, vol. 33, pp. 6840-6851.

[32] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (*CVPR*), 2021, pp. 12873-12883.

[33] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.* (*NeurIPS*), 2017, vol. 30.

[34] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv:1710.10196,* 2017.

[35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2019, pp. 4401-4410.

[36] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (*CVPR*), 2016, pp. 5525-5533.

[37] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (*ICCV*), 2015, pp. 3730-3738.

[38] R. Ma *et al.*, "Towards blind watermarking: Combining invertible and non-invertible mechanisms," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1532–1542.

[39] H. Fang, Z. Jia, Z. Ma, E.-C. Chang, and W. Zhang, "PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 2267–2275.

[40] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text compatible image prompt adapter for text-to-Image diffusion models," *arXiv:2308.06721,* 2023.

[41] S. Baliah, Q. Lin, S. Liao, X. Liang, and M. H. Khan, "Realistic and efficient face swapping: A unified approach with diffusion models," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*(*WACV*), 2025, pp. 1062-1071.

[42] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (*CVPR*), 2020, pp. 5074-5083.