UIAMark: Unified Identity and Attribute Watermarking for Source Tracing and Proactive Deepfake Detection

Peiqi Jiang* , **Yuhao Sun*** , **Lingyun Yu** and **Hongtao Xie**
USTC, EEIS
{jpqjiang, syh3327}@mail.ustc.edu.cn {yuly, htxie}@ustc.edu.cn

Abstract

Malicious Deepfakes have posed sharp security risks to our society by generating remarkably realistic forged faces. Although various countermeasures have been developed for the post-detection of Deepfakes, their performance still lacks reliability and stability in real-world applications. Instead of detecting Deepfakes passively, we propose a novel framework called **U**nified **I**dentity and **A**ttribute Water**mark**s (**UIAMark**), which is devised to provide a proactive strategy for source tracing and Deepfake-agnostic detection with a face disentangling and reconstruction architecture. In detail, we embed a robust watermark into the **facial-irrelevant attributes** which is naturally resilient to Deepfakes, accomplishing robust **source tracing**. Meanwhile, pairs of self-correlated semi-fragile watermarks are injected into **facial-relevant attributes** and **identity** features. If either the facial identity or attributes have been manipulated, the correlation between the semi-fragile watermarks will be broken. In this way, we can fully leverage the inherent Deepfake-fragile property of the high-level semantic representations, achieving **Deepfake-agnostic detection**. Extensive experiments demonstrate the effectiveness of our approach, we achieve an average detection accuracy of 97.36% across various Deepfake methods and a low BER of 0.0389% across different distortions.

1 Introduction

With the rapid evolution of deep generative models [Karras *et al.*, 2019; Goodfellow *et al.*, 2020], advanced face forgery algorithms [Thies *et al.*, 2016; Thies *et al.*, 2019] are able to create remarkably realistic face images, posing potential risks for malicious use and raising significant privacy and security concerns. In recent years, there has been a substantial increase in research dedicated to detecting Deepfakes [Zheng *et al.*, 2021; Haliassos *et al.*, 2021]. Most of these works formulated this task as a binary classification problem and focused on the in-dataset performance. However, training a classifier that solely relies on pre-existing forgery artifacts may not be sufficient for detecting unseen manipulations. While several attempts [Shiohara and Yamasaki, 2022; Dong *et al.*, 2023] have been made to improve generalization ability by capturing common forgery clues, nevertheless, their performance is still limited when applied in real-world scenarios and in-the-wild forensics.

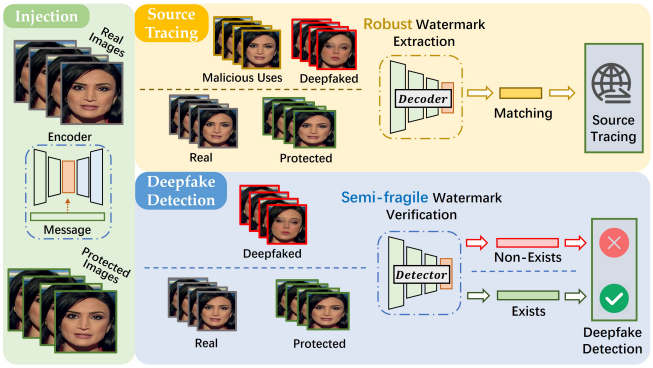


Figure 1: Workflow of source tracing and proactive Deepfake detection. 1) For **Injection**, we embed secret messages into the target image with different robustness levels, and it may be manipulated by malicious users. 2) For **Source Tracing**, we extract the robust watermark and matching with the predefined criteria. 3) For **Deepfake Detection** we verify the existence of the semi-fragile watermark. If the image has been Deepfaked, this watermark will not exist.

Alternatively, a distinct area of research has focused on proactive forensics against malicious tampering, which leverages digital watermarking techniques [Zhu *et al.*, 2018] to safeguard real images by embedding secret messages in advance. Based on the robustness of the watermark, we classify these watermarking methods into two distinct types. 1) The *robust watermark* is resilient against both common distortions and Deepfake manipulations, making it suitable for trustworthy source verification and source tracing [Wang *et al.*, 2020]. 2) The *semi-fragile watermark* is robust to common distortions but sensitive to malicious Deepfakes, thus we can proactively detect Deepfakes by checking whether the watermark exists in the tampered image [Asnani *et al.*, 2022].

Although several attempts have been made in proactive forensics, the field continues to grapple with two key challenges: **1) Unrealistic premise of watermark message**

*Equal contribution.

prior: Most existing methods rely on injecting semi-fragile watermarks, and the detection process involves matching the decoded semi-fragile watermarks with the original watermarks to identify Deepfakes. However, in practical watermark distribution scenarios, the secret message embedded in different images is not a constant value. Therefore, obtaining prior knowledge of the original watermark during the detection process becomes unrealistic. **2) Lack of Generalization:** Current research fails to achieve Deepfake-agnostic detection. For instance, FakeTagger [Wang *et al.*, 2021] uses a GAN simulator to act as a surrogate module to generate Deepfake translations. However, this strategy can lead to inconsistent outcomes when confronted with manipulations beyond the scope of the training process. A similar challenge is evident in [Wu *et al.*, 2023], where their method proves effective against specific types of Deepfake generators but struggles to generalize against others.

To tackle the aforementioned issues, we propose a **U**nified **I**dentity and **A**ttribute Water**mark**ing (**UIAMark**) framework to provide robust source tracing and self-verifiable Deepfake-agnostic detection. The workflow of our approach is illustrated in Figure 1. To realize the proposed objective, we analyze the watermark embedding process in two critical aspects: 1) For reliable source tracing, we embed a decodable (i.e., specific binary sequence can be extracted) robust watermark that is resilient against various types of tampering. Moreover, to hide the robust watermark in the facial-irrelevant attributes, we adopt a learning-based encoder-decoder watermarking architecture to embed the watermark adaptively. 2) To achieve proactive detection without the watermark message prior, a pair of correlated semi-fragile watermarks that can be self-verified are embedded. Moreover, to accomplish Deepfake-agnostic detection, we inject the correlated watermark sequences into the facial-relevant attributes and identity features in an addictive way. If either the facial identity or attributes are manipulated, it will disrupt the correlation between the semi-fragile watermarks, making our approach effective against both face-swapping and attribute-editing Deepfake manipulations.

Building upon these insights, UIAMark embeds two aspects of watermarks within the facial attributes and identities through a modified face reconstruction network. Specifically, given an original image, we first generate a pair of k-bit control sequence and $2^k - 1$-bit pseudo-random noise (PN) sequence with unique autocorrelation properties tied to each specific control sequence. Then we embed the binary control sequence into the face attributes in a deep separable manner, where a single encoder is used for injecting and two decoders are used to extract the secret message at different robustness levels. In this way, we can obtain the robust watermark on facial-irrelevant attributes for **source tracing** and a semi-fragile watermark on facial-relevant attributes which are inherently fragile to attribute-edit manipulations. Meanwhile, the identity watermark is injected by entangling the facial identity feature with the PN sequence, making the identity watermark inherently fragile to face-swap translations. Therefore, we can accomplish the **Deepfake-agnostic detection** by verifying the correlation status between the semi-fragile attribute watermark and identity watermark, any tampering with either facial attributes or identity will disrupt the inherent auto-correlation property of the PN sequence. Finally, based on the watermarked attributes and watermarked identity, we leverage a face generator to generate the final watermarked image. To conclude, our contributions can be summarized as:

- We propose a unified attribute and identity watermarking framework, which brings a new paradigm to source tracing and proactive Deepfake detection.
- We accomplish Deepfake-agnostic detection by leveraging the inherent fragility of high-level facial semantic information (i.e., facial-relevant attributes and identity features) and utilizing the correlation property between the control and pseudo-random sequences.
- Extensive experiments demonstrate the effectiveness of our approach in source tracing and proactive Deepfake detection. In particular, we achieve a BER of 0.0389% across different distortions and an average detection accuracy of 97.36% across various Deepfake methods in the black-box setting.

2 Related Work

2.1 Passive Deepfake Detection

With the rise of malicious Deepfake abuses, numerous attempts have been in the field of passive Deepfake detection. These methods all follow a similar paradigm that is trained on a given dataset with binary labels and detect Deepfakes after the manipulations have already happened. Earlier research primarily focused on designing a powerful network to detect artifacts from RGB representations [Amerini *et al.*, 2019]. Besides, researchers have also begun to notice the vital problem of cross-dataset generalization. For instance, Face X-ray [Li *et al.*, 2020] focused on detecting blending boundary artifacts, PCL+I2G [Zhao *et al.*, 2021] detecting Deepfakes by capturing the intra-frame inconsistency, and SBIs [Shiohara and Yamasaki, 2022] reproduce common Deepfake artifacts by seamlessly stitching two images.

2.2 Proactive Defence against Deepfake

The research on proactive defense mainly leverages digital watermarking techniques to safeguard real images by embedding secret messages in advance. [Wang *et al.*, 2021] proposed a Deepfake-robust watermarking network to identify the Deepfake source. [Asnani *et al.*, 2022] introduced a semi-fragile watermarking strategy through a combination of Deepfake surrogate models. [Zhao *et al.*, 2023] leveraged identity watermarking to detect face-swap manipulations. [Wu *et al.*, 2023] proposed a separable deep watermarking approach that contains both robust watermark and semi-fragile watermark, and widens the application scenarios of watermark protection methods. However, most prevalent methods rely on conventional digital watermarking frameworks and substitute the noise layer with a Deepfake surrogate module in their design. This training approach inherently restricts the capacity for achieving generalizable Deepfake detection. In contrast, our framework is devised from a

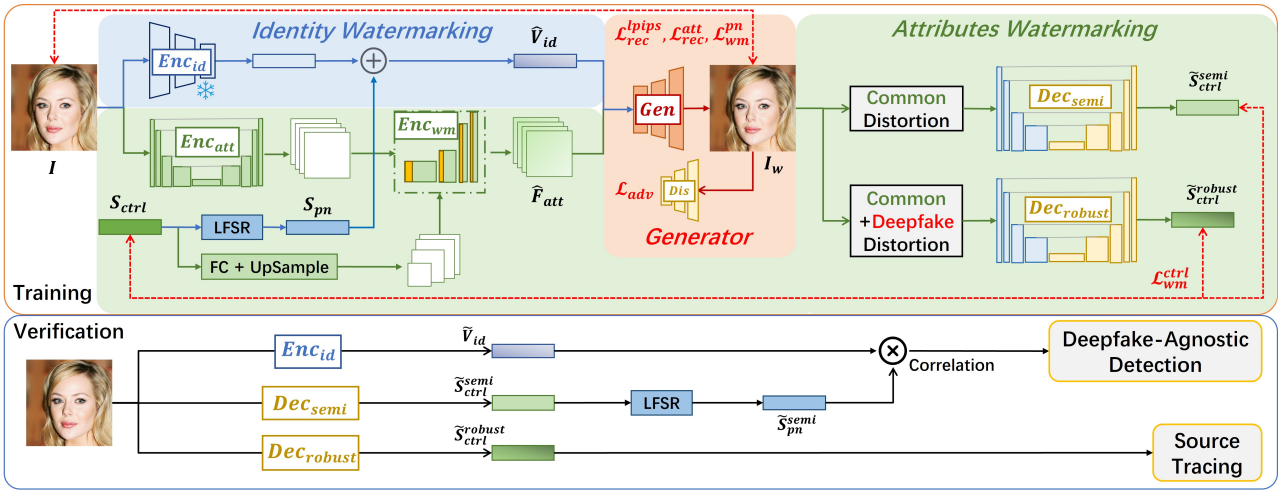


Figure 2: The overview of UIAMark. 1) In the training stage, we disentangle the input image into the attributes and identity representations. Then, we embed the control sequence into attribute features through a watermark encoder, and it can be separately extracted by two decoders at different robustness levels. Meanwhile, we inject the PN sequence into the identity feature, making the identity watermark inherently fragile to face-swap translations. 2) In the verification stage, we leverage the robust watermark for **source tracing** and **detect Deepfakes** based on the correlation result between the identity features and the control sequence decoded from the semi-fragile watermark.

face reconstruction network, allowing us to leverage the inherent fragility of high-level facial semantic information for Deepfake-agnostic detection.

2.3 Digital Watermarking

Digital watermarking is a broadly studied approach to imperceptibly concealing secret messages within media content [Cox, 2007]. However, earlier hand-crafted watermarking methods suffer from limited robustness and generalization. Recently, DNN-based approaches have been introduced in digital watermarking for stronger robustness. For example, HiDDeN [Zhu *et al.*, 2018] is the first end-to-end framework for robust watermarking, where an encoder-decoder network is utilized for watermark injection and extraction, along with a noise layer to simulate perturbations for robustness. [Luo *et al.*, 2020] employs an adversarial noise layer to further improve the robustness against unseen distortions. Inspired by the encoder-decoder-based watermarking paradigm, our method designs a single encoder and two separate decoders which are integrated into a face reconstruction network for the attribute watermarking.

3 Method

3.1 Model Architecture

We illustrate the overview of the proposed UIAMark in Figure 2. The whole pipeline is based on a face reconstruction network and we elaborate on the watermark injection and extraction process as follows: 1) Given an real image I, we first randomly generate a pair of k-bit control sequence S_{ctrl} and $2^k - 1$-bit PN sequence S_{pn} as the secret messages. 2) Then, we use an identity encoder and an attribute decoder to disentangle the input image into two independent representations: attributes and identity. 3) Next, S_{ctrl} and S_{pn} are embedded into the attributes and identity separately. For the attribute watermark, we leverage a learnable watermark encoder to embed the control sequence S_{ctrl} into the attributes and obtain the watermarked attributes. For the identity watermark, we directly add the scaled S_{pn} into the identity vector to generate the watermarked identity. 4) Finally, we utilize a face reconstruction generator to synthesize the corresponding watermarked image I_w.

In the watermark extraction stage, we first utilize two separate watermark decoders to extract the attribute watermark at two robustness levels, i.e., a robust attribute watermark S_{ctrl}^{robust} and a semi-fragile attribute watermark S_{ctrl}^{semi}, where the robust watermark can be used for **source-tracing**. Meanwhile, we extract the identity vector from the watermarked image and compute the correlation with the S_{ctrl}^{semi} sequence, if Deepfakes manipulate either attribute or identity, the correlation property will break. Thus, we can leverage this feature to **detect Deepfakes**.

Next, we will introduce the details of each component.

Sequence Generation. In the sequence generation stage, we aim to create a pair of messages with certain correlations for the attribute and identity watermarking, therefore we can leverage this correlation to verify the existence of the identity watermark. Inspired by the telecommunication system, we fully leverage the features of pseudo-random noise (PN) sequences. We first generate a random k-bit binary sequence denoted as the control sequence S_{ctrl}. Next, we feed the control sequence into the linear feedback shift registers (LFSR) to generate a $2^k - 1$-bit PN sequence S_{pn}. The PN sequence has the same statistical property as the Gaussian noise but has a special correlation property. The auto-correlation of the PN sequence is approximate to an impulse function, with the impulse peaks occurring only at the zero point, it can be formulated as:

$$Corr[l] = \begin{cases} 1 & \text{if } l = 0 \\ -\frac{1}{N} & \text{otherwise} \end{cases}, \quad (1)$$

where the period N equals to $2^k - 1$. With this autocorrelation property, we can verify whether the control sequence and the PN sequence match with each other.

Feature Disentanglement. To entangle the message with the original facial attributes and identity, we first employ two encoders, namely identity encoder $\mathbf{Enc_{id}}(\cdot)$ and attributes encoder $\mathbf{Enc_{att}}(\cdot)$, to extract the respective representations from the input image I.

The attributes representation of a face image encompasses features including style, pose, expression, background, etc. These features contain different levels of detail, ranging from coarse features like the overall facial outline to finer details such as hair color. Therefore, we devise a modified U-net style network to encode k levels of attribute features $F_{att} = \left\{F_{att}^1, F_{att}^2, ..., F_{att}^n\right\} = \mathbf{Enc_{att}}(I)$, where $F_{att}^i \in \mathbb{R}^{C_i \times H_i \times W_i}$ denotes the i-th level of feature map in the encoder $\mathbf{Enc_{att}}(\cdot)$.

The identity representation refers to the high-level semantic feature used to characterize a specific person. We utilize a pretrained face recognition model to act as the identity encoder. We extract the identity $V_{id} = \mathbf{Enc_{id}}(I)$, where the identity vector $V_{id} \in \mathbb{R}^l$, and l is the vector length.

During the training process, the attribute encoder $\mathbf{Enc_{att}}(\cdot)$ is trained from scratch and we freeze the pretrained identity encoder $\mathbf{Enc_{id}}(\cdot)$. Thus, the disentanglement process does not require extra attribute annotations.

Watermark Encoding. Given the disentangled facial representations, we aim to embed a pair of control sequence S_{ctrl} and PN sequence S_{pn} into the attribute features F_{att} and identity vector V_{id}.

Attributes Watermarking: Considering that attribute features have different detail levels, and their feature maps are at different sizes, we adopt an encoder-decoder-based watermark embedding architecture and aim to inject two robustness levels of watermarks on different attribute spaces. Specifically, for each level of attribute features, the attribute watermark encoder $\mathbf{Enc_{wm}}$ maps the k-bit control sequence S_{ctrl} into the corresponding feature map size through independent *FC+Upsample* layers. Then we concatenate the upsampled message with F_{att}^i and fuse them through a lightweighted Coordinate Attention [Hou *et al.*, 2021] mechanism, which makes the message embedding process become position-sensitive and direction-aware. The process is illustrated as $\widehat{F}_{att} = \left\{\widehat{F}_{att}^1, \widehat{F}_{att}^2, ..., \widehat{F}_{att}^n\right\} = \mathbf{Enc_{wm}}(F_{att})$, where $\widehat{F}_{att}$ is the watermarked attributes with n levels of watermarked feature maps $\widehat{F}_{att}^i$ at different resolution.

Identity Watermarking: Observing that facial identity vectors exhibit statistical characteristics similar to Gaussian noise, we inject autocorrelated pseudo-noise (PN) sequence S_{pn} with the same statistical properties into the identity vector, as formalized as $\widehat{V}_{id} = V_{id} + \gamma S_{pn}$, where $\widehat{V}_{id}$ is the watermarked identity and γ is the scale factor.

Face Reconstruction. To integrate watermarked attributes and identity and generate watermarked images with high visual quality. We employ the widely used *Adaptively Attentional Denormalization Generator (AAD-Generator)* [Li *et al.*, 2019] in the face-swapping domain for facial reconstruction. Specifically, the AAD-Generator contains a cascade of AAD Residual Blocks, and each block consists of two AAD layers with residue convolution connections. AAD layer first takes the normalized activation from the previous level as input h_{in}^i. The attentional mask M_i is generated from h_i with convolutions and a Sigmoid operation. The attribute and identity are integrated by denormalizing h_i, it can be formulated as $A_i = \gamma_{att}^i \otimes h_{in}^i + \beta_{att}^i$ and $I_i = \gamma_{id}^i \otimes h_{in}^i + \beta_{id}^i$, where A_i and I_i is the integrated attribute and identity in the i-th layer, γ_{att}^i and β_{att}^i are two modulation parameters both convolved from $\widehat{F}_{att}$, they share the same tensor dimensions with $h_{in}^i \in \mathbb{R}^{C_i \times H_i \times W_i}$. Similarly, $\gamma_{id}^i \in \mathbb{R}^{C_i}$ and $\beta_{id}^i \in \mathbb{R}^{C_i}$ are another two modulation parameters generated from $\widehat{V}_{id}$ through FC layers.

Finally, after the integration of n level of attribute and identity features, the AAD-Generator $\mathbf{Gen_{AAD}}(\cdot)$ outputs the final output image as $I_w = \mathbf{Gen_{AAD}}\left(\widehat{F}_{att}, \widehat{V}_{id}\right)$.

Watermark Extraction. Given a watermarked image I_w, the decoding processes for facial attributes and identity watermarks differ based on their respective embedding methods.

Attribute Watermark Decoding: For the attribute watermark, considering that different levels of attribute features include coarse-grained background and contours, as well as finer details such as hair color and facial expressions, we opted for a separable watermark design [Wu *et al.*, 2023] to decode two watermarks at different robustness levels. We extract a robust watermark and a semi-fragile watermark. The decoding process is formulated as

$$\begin{cases} \widetilde{S}_{ctrl}^{robust} = \mathbf{Dec_{robust}}\left(\mathbf{Noise}\left(\mathbf{DF}\left(I_w\right)\right)\right) \\ \widetilde{S}_{ctrl}^{semi} = \mathbf{Dec_{semi}}\left(\mathbf{Noise}\left(I_w\right)\right) \end{cases}, \quad (2)$$

where $\widetilde{S}_{ctrl}^{robust}$ and $\widetilde{S}_{ctrl}^{semi}$ denotes the decoded robust and semi-fragile attribute watermarks. $\mathbf{Noise}(\cdot)$ denotes a differential noise layer containing JPEG compression, resize, crop, and random noise distortions. $\mathbf{DF}(\cdot)$ denotes surrogate Deepfake models for the robust watermark, including SimSwap [Chen *et al.*, 2020] and StarGAN [Choi *et al.*, 2018].

Identity Watermark Extracting: During decoding, it is necessary to detect whether the corresponding sequence exists in the identity features of the watermarked image (further details on the verification method are presented in Section 3.3). We extracted the reconstructed identity vector as $\widetilde{V}_{id} = \mathbf{Enc_{id}}\left(I_w\right)$, where $\widetilde{V}_{id}$ denotes the identity of the watermarked image.

3.2 Training Loss

Message Loss. For the robust and semi-fragile attribute watermarks, we use the L_2 loss to constrain the message before and after decoding. The formula is as follows:

$$\mathcal{L}_{wm}^{ctrl} = L_2\left(S_{ctrl}, \widetilde{S}_{ctrl}^{robust}\right) + L_2\left(S_{ctrl}, \widetilde{S}_{ctrl}^{semi}\right). \quad (3)$$

For the identity watermark, we do not decode specific watermark values. Instead, we constrain the cosine similarity between the identity vector in the watermarked image and the original identity watermark vector. The process is as follows:

$$\mathcal{L}_{wm}^{pn} = 1 - \mathbf{cos}\left(\widetilde{V}_{id} - \widehat{V}_{id}\right). \quad (4)$$

Reconstruction Loss. To maintain the high fidelity of the reconstructed image to the original and minimize conflicts with watermark injection at the pixel level, we utilize a perceptual similarity loss (LPIPS) [Zhang *et al.*, 2018] instead of the pixel-level reconstruction loss:

$$\mathcal{L}_{rec}^{lpips} = \left\| \mathbf{LPIPS}\left(I, I_w\right) \right\|_2. \tag{5}$$

Additionally, to ensure the training stability of the face reconstruction model, we also constrain the reconstruction loss for attributes as follows:

$$\mathcal{L}_{rec}^{att} = \frac{1}{2}\sum_{i=1}^{n} \left\| \widehat{F}_{att}^{i} - \widetilde{F}_{att}^{i} \right\|_2^2, \tag{6}$$

where $\widetilde{F}_{att}^{i}$ is the attribute features extracted from the watermarked image I_w.

Adversary Loss. To enhance the realism of the reconstructed image, we utilize a multi-scale discriminator $\mathbf{Dis}(\cdot)$ from [Isola *et al.*, 2017] with hinge loss functions, training our model in an adversarial manner:

$$\mathcal{L}_{adv} = \mathbf{log}\left(\mathbf{Dis}\left(I\right)\right) + \mathbf{log}\left(1 - \mathbf{Dis}\left(I_w\right)\right). \tag{7}$$

Overall Loss. Our framework is trained using a weighted sum of the aforementioned losses, as defined by:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{wm}^{ctrl} + \lambda_2 \mathcal{L}_{wm}^{pn} + \lambda_3 \mathcal{L}_{rec}^{lpips} + \lambda_4 \mathcal{L}_{rec}^{att} + \lambda_5 \mathcal{L}_{adv}, \tag{8}$$

where $\lambda_1 \sim \lambda_5$ are hyper-parameters to weight the corresponding loss.

3.3 Watermark Verification

In this section, we will elaborate on how to use attribute and identity watermarks for source tracing and Deepfake detection. Source tracing requires the secret message from a robust watermark and cross-verification with the original message. Thus, we leverage the robust attribute watermark $\widetilde{S}_{ctrl}^{robust}$ to accomplish it. As for Deepfake detection, we aim to detect the presence of the PN sequence S_{pn} in the identity feature of the watermarked image. The existence of S_{pn} will determine whether the image is Deepfaked or not. Given the watermarked image is not modified by Deepfakes, we extract its identity vector $\widetilde{V}_{id}$, and it can be expanded as $\widetilde{V}_{id} = V_{id} + \gamma S_{pn}$, where V_{id} is the clean identity vector, S_{pn} is the PN sequence. Then, we compute its correlation with the PN sequence decoded from the semi-fragile watermark $\widetilde{S}_{ctrl}^{semi}$ to generate a corresponding PN sequence $\widetilde{S}_{pn}^{semi}$ through the LFSR. Then verify whether the watermark exists in the input image. The correlation process is defined as:

$$\begin{aligned} Corr[l] &= \sum_{i=0}^{N-1} \widetilde{S}_{pn}^{semi}[i] * \widetilde{V}_{id}[j] \\ &= \sum_{i=0}^{N-1} \left(\widetilde{S}_{pn}^{semi}[i] * \gamma S_{pn}[j] + \widetilde{S}_{pn}^{semi}[i] * V_{id}[j] \right), \end{aligned} \tag{9}$$

where $j = (N - 1 + i - l)$, N is the sequence length and l is ranged from $\{-N+1, ..., 0, ...N-1\}$. The first term in Equation (9) is equivalent to an impulse function and the second term is a bounded constant. Thus, according to Equation (1), $Corr[l]$ will have a distinct peak value at 0 index if the watermarked image is not manipulated. Thus we can determine whether the watermarked image has been Deepfaked by comparing the Peak to Average Power Ratio (PAPR) with threshold τ.

4 Experiments

4.1 Implementation Details

Experiments Settings. Our experiments are conducted on the CelebA-HQ dataset [Karras *et al.*, 2017] following the official split for training, validation, and testing. We further conduct experiments on the Flickr-Faces-HQ (FFHQ) [Karras *et al.*, 2019] dataset to provide more comprehensive results. Unless stated otherwise, all images in the experiment have been aligned and cropped to the size of 256×256. We set $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = \lambda_4 = 10$, $\lambda_5 = 0.1$, scale factor $\gamma = 0.1$, and threshold $\tau = 4$. As for the choice of PN sequence type, we select the widely used M Sequence.

Evaluation Metrics. For Deepfake detection, we compute image-level Accuracy (Acc) and F1-Score. For source tracing, we evaluate the robust watermarks with the average bit error rate (BER) metric. For visual quality, we report the average PSNR, SSIM, and LPIPS for watermarked images through the testing set.

Compared Methods. We mainly compared our work with SepMark [Wu *et al.*, 2023], which is the first work of unified source tracing and Deepfake detection. Additionally, most existing works only embed watermarks with a single functionality. Therefore, we adopt robust watermarking methods like HiDDeN [Zhu *et al.*, 2018], MBRS [Jia *et al.*, 2021], and semi-fragile watermarking methods such as FaceSign [Neekhara *et al.*, 2022], IDMark [Zhao *et al.*, 2023] as our baselines. Note that SepMark only reports the BER of robust and semi-fragile watermarks, thus we compare the delta of two watermarks' BER with a certain threshold to determine whether the image is Deepfaked (See details in supplementary materials). We also compare the proposed proactive defense strategy with SOTA detectors such as ICT [Dong *et al.*, 2022], SBI [Shiohara and Yamasaki, 2022], and UIA-ViT [Zhuang *et al.*, 2022].

4.2 Visual Quality

We evaluate the visual quality of the generated watermark images, as shown in Table 1. Our method achieves competitive results against the SepMark method (33.43 of PNSR, 0.9446 of SSIM, and 0.0021 of LPIPS). Additionally, as illustrated in Figure 3, no artifacts can be observed by the naked eye in the generated watermark images, demonstrating the high fidelity of our approach. Specifically, from the residual between the watermark image and the original image, it can be observed that in regions associated with facial identity, the distribution of watermarks is similar to white noises. Conversely, in regions related to the background and certain facial attributes, the watermark is distributed differently in three RGB channels. This demonstrates the effectiveness of our approach to separately embedding watermarks on attribute and identity information.

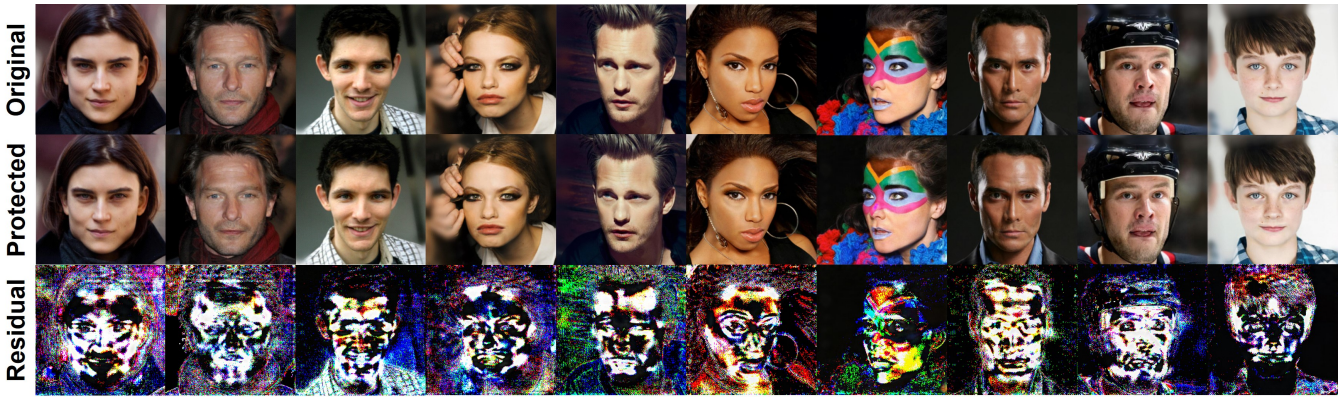


Figure 3: Visual Quality of watermarked images. From top to bottom are the original image, watermarked image, and normalized residual between the original and watermarked images.

Method	CelebA-HQ			FFHQ		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
HiDDeN	33.26	0.8880	0.0210	31.13	0.8670	0.0170
FaceSign	32.33	0.9211	0.0260	32.99	0.8890	0.0150
MBRS	33.05	0.8106	0.0141	33.00	0.7750	0.0241
IDMark	33.32	0.9400	0.0030	33.50	0.9500	0.0042
SepMark	**38.56**	0.9328	0.0080	**37.60**	0.8793	0.0569
Ours	33.43	**0.9446**	**0.0021**	33.75	**0.9533**	**0.0023**

Table 1: Visual quality of the watermarked image. We achieve the best fidelity on SSIM and LPIPS and the second-best on PSNR.

Distortion	Robust Watermark BER(%) ↓			
	FaceSign	MBRS	SepMark	Ours
Identity	0.0136	0.0000	0.0000	0.0000
JPEG	0.8258	0.2597	0.2136	0.0782
Resize	1.0726	0.0000	0.0059	0.0000
GaussianBlur	0.1671	0.0000	0.0024	0.0010
MedianBlur	0.0977	0.0000	0.0012	0.0100
Brightness	10.8196	0.0000	0.0059	0.0079
Contrast	0.0334	0.0000	0.0012	0.0011
Saturation	0.7116	0.0000	0.0000	0.0000
Hue	8.3780	0.0000	0.0000	0.0000
SaltPepper	12.3238	0.0000	0.0413	0.0177
GaussianNoise	7.0697	0.0000	0.7460	0.3120
Avg. Common	3.7738	**0.0236**	0.0848	0.0389
SimSwap	0.0537	19.3744	13.8255	0.2231
StarGAN	0.5311	18.1358	0.1268	0.1300
Avg. Malicious	0.2924	18.7551	6.9762	**0.1766**

Table 2: Quantitative comparison between SepMark, IDMark, and ours on CelebA-HQ regarding the bit-wise error rate under common and malicious Deepfake manipulations.

4.3 Source Tracing

The performance of sourcing tracing primarily depends on the robust attribute watermark $\widetilde{S}_{ctrl}^{robust}$, which should have a low BER under both common and Deepfake distortions. As shown in Table 4, the average BER of $\widetilde{S}_{ctrl}^{robust}$ in our method achieves 0.0389% under common distortions and 0.1766% under Deepfake distortions, demonstrating stronger robustness compared to the baselines. This verifies the advantage of our hierarchical fine-grained attribute watermark embedding and separable design. The $\widetilde{S}_{ctrl}^{robust}$, which is mainly distributed in the background region, can better handle strong distortions, especially background-independent Deepfake distortions. Notably, due to the learning-based attribute watermarking paradigm, a portion of $\widetilde{S}_{ctrl}^{robust}$ is inevitably dispersed across facial-relevant attributes. Consequently, when encountering face-swapping techniques such as SimSwap, there may be a marginal increase in BER.

4.4 Deepfake Detection

For Deepfake detection, both SOTA detector-based and watermark-based methods are selected for comparison with our method. We evaluate these methods on four different Deepfake models including face-swap methods like SimSwap and DiffFace, as well as attribute-edit methods like StarGAN and HFGI. As shown in Table 3, our method achieves satisfactory Deepfake detection accuracy across all the models with an average of 97.36%. It is worth mentioning that SepMark performs slightly better than our method in StarGAN. This is because SepMark incorporates StarGAN and SimSwap during the training process, whereas for our method, they are considered as black-box scenarios. Conversely, on the unseen Deepfake methods like HFGI and DiffFace, the detection accuracy of SepMark significantly drops to nearly 50%, while our method maintains high accuracy. The results demonstrate the superiority of our method in Deepfake-agnostic detection by leveraging the inherent fragility of both the identify and attribute information.

Meanwhile, we evaluate the performance of our method in Deepfake detection under common distortions, as shown in Table 4. Our method maintains a high detection accuracy under different distortions, verifying the robustness of the semi-fragile watermarks embedded in identity and attribute information against common distortions.

Method	CelebA-HQ (Acc(%) ↑ / F1-Score(%) ↑)				FFHQ (Acc(%) ↑ / F1-Score(%) ↑)			
	SimSwap	StarGAN	HFGI	DiffFace	SimSwap	StarGAN	HFGI	DiffFace
ICT	76.70/76.63	77.90/78.56	69.72/68.19	76.80/78.08	76.95/76.49	78.80/78.82	68.15/67.74	78.45/78.51
SBI	76.65/77.08	70.90/76.34	70.48/75.91	42.96/32.34	71.83/76.46	74.53/79.18	71.53/76.15	44.86/37.49
UIA-ViT	48.65/19.07	62.21/50.94	71.33/66.71	48.95/19.92	48.75/65.25	69.21/66.36	69.01/66.07	45.90/26.07
IDMark	94.05/94.02	78.95/81.72	89.35/89.87	95.00/94.92	92.80/92.58	80.00/81.63	90.70/88.07	91.83/91.57
SepMark	89.67/90.63†	**99.78/99.78**†	50.00/66.67	50.00/66.67	87.60/88.97†	**98.73/98.73**†	50.00/66.67	50.00/66.67
Ours	**97.42/97.48**	97.50/96.97	**97.28/97.23**	**97.25/97.19**	**98.90/98.89**	96.90/96.94	**98.67/98.66**	**98.07/98.07**

Table 3: Performance of Deepfake detection Accuracy ↑ and F1-score ↑. We compare the proposed proactive defense strategy with SOTA detectors, as well as watermarking baselines. †: White-box evaluation.

Distortion	Acc(%) ↑			
	SimSwap	StarGAN	HFGI	DiffFace
Identity	97.42	97.50	97.28	97.25
JPEG	93.63	76.75	87.95	77.42
Resize	97.15	97.00	97.20	97.58
GaussianBlur	90.52	89.95	89.80	97.95
MedianBlur	93.23	92.80	93.75	97.65
Brightness	89.97	87.45	76.45	86.67
Contrast	94.52	94.05	94.95	96.27
Saturation	97.13	96.70	97.45	97.93
Hue	96.85	96.85	97.55	97.20
SaltPepper	83.53	72.75	74.05	71.37
GaussianNoise	92.37	77.90	79.35	73.83
Avg.	93.30	89.06	89.62	90.10

Table 4: Robustness of Deepfake detection on CelebA-HQ dataset. Our method maintains a high accuracy under various distortions.

γ	Visual Quality		Acc(%) ↑			
	PSNR	SSIM	SimSwap	StarGAN	HFGI	DiffFace
0.01	34.75	0.9501	53.17	52.18	52.83	52.67
0.05	34.11	0.9480	85.67	86.15	83.00	84.73
0.10	33.43	0.9446	97.42	97.50	97.28	97.25
0.25	32.08	0.9400	97.67	97.83	94.67	94.83
0.50	30.63	0.9340	89.83	89.33	83.17	86.63
1.00	26.87	0.9150	53.17	50.83	51.83	50.13

Table 5: Ablations of scale factor γ in identity watermark. We find that $\gamma = 0.1$ is the best trade-off point.

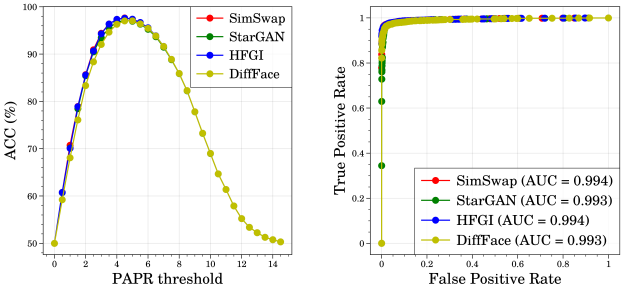
4.5 Ablation Study

Impact of scale factor γ in identity watermark. We investigate the impact of different scale factors on identity watermarks. The ideal injection strength needs to be strong enough for detection while maintaining high fidelity. As illustrated in Table 5, we find that $\gamma = 0.1$ for the best trade-off point.

Impact of different identity watermark sequences. According to Equation (9), the correlation property of the injected identity watermark sequence plays a significant role in the verification. To analyze the impact of the different sequence types in the identity watermarking process, we select two representative PN sequences: M Sequence and Gold Code. Along with two common random sequences: Bernoulli and Gaussian Sequences to compare the detection performance. As illustrated in Table 6, two PN sequences achieve competitive detection accuracy, while the random sequences failed to detect Deepfakes, demonstrating the necessity of choosing the PN sequences for identity watermarking.

ID Watermark Sequence Type		Acc(%) ↑			
		SimSwap	StarGAN	HFGI	DiffFace
Random	Bernoulli	50.00	52.10	53.21	51.12
	Guassain	51.20	53.25	52.42	52.50
Pseudo Random	Glod Code	96.56	**97.72**	96.31	97.00
	M Sequence	**97.42**	97.50	**97.28**	**97.25**

Table 6: Different sequences' detection results. Demonstrating the necessity of choosing the PN sequences for identity watermarking.



(a) Acc (b) ROC

Figure 4: Ablations of PAPR threshold τ, we compute the Acc-PAPR curve and the ROC curve.

Impact of different PAPR threshold τ. Figure 4a illustrates the impact of different PAPR thresholds τ on the accuracy of Deepfake detection. It can be observed that the optimal threshold for various Deepfake models is approximately equal to 4, which aligns with the value chosen in our experiments. With the variation of threshold values, we also plot the ROC curves and compute corresponding AUC as shown in Figure 4b. For all four Deepfake models, the AUC values are as high as 0.994. These results indicate the excellent performance of our method in Deepfake detection.

5 Conclusion

In this paper, we proposed a novel watermarking paradigm for protecting face images from malicious Deepfakes. By em-

bedding pairs of self-correlated watermarks into the facial attributes and identity, we fully leverage the inherent fragility of high-level facial semantic information, accomplishing unified robust source tracing and Deepfake-agnostic detection. Extensive experiments demonstrate our method's superior performance on widely used datasets, establishing a strong baseline for future research.

References

[Amerini *et al.*, 2019] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[Asnani *et al.*, 2022] Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, and Xiaoming Liu. Proactive image manipulation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15386–15395, 2022.

[Chen *et al.*, 2020] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[Cox, 2007] IJ Cox. Digital watermarking and steganography. *Morgan Kaufmann google schola*, 2:893–914, 2007.

[Dong *et al.*, 2022] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer, 2022.

[Dong *et al.*, 2023] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization, 2023.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Haliassos *et al.*, 2021] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021.

[Hou *et al.*, 2021] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[Jia *et al.*, 2021] Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pages 41–49, 2021.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[Li *et al.*, 2019] L Li, J Bao, H Yang, D Chen, and F Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. arxiv 2019. *arXiv preprint arXiv:1912.13457*, 2019.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[Luo *et al.*, 2020] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13548–13557, 2020.

[Neekhara *et al.*, 2022] Paarth Neekhara, Shehzeen Hussain, Xinqiao Zhang, Ke Huang, Julian McAuley, and Farinaz Koushanfar. Facesigns: semi-fragile neural watermarks for media authentication and countering deepfakes. *arXiv preprint arXiv:2204.01960*, 2022.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[Thies *et al.*, 2016] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[Thies *et al.*, 2019] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[Wang *et al.*, 2020] Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Lei Ma, Yang Liu, and Lina Wang. Deeptag: Robust image tagging for deepfake provenance. *arXiv preprint arXiv:2009.09869*, 3, 2020.

[Wang *et al.*, 2021] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3546–3555, 2021.

[Wu *et al.*, 2023] Xiaoshuai Wu, Xin Liao, and Bo Ou. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. *arXiv preprint arXiv:2305.06321*, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhao *et al.*, 2021] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[Zhao *et al.*, 2023] Yuan Zhao, Bo Liu, Ming Ding, Baoping Liu, Tianqing Zhu, and Xin Yu. Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4602–4611, 2023.

[Zheng *et al.*, 2021] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021.

[Zhu *et al.*, 2018] 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.

[Zhuang *et al.*, 2022] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision*, pages 391–407. Springer, 2022.