Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.059370

REVIEW

# Enhancing Deepfake Detection: Proactive Forensics Techniques Using Digital Watermarking

**Zhimao Lai[1,2], Saad Arif[3], Cong Feng[4], Guangjun Liao[5] and Chuntao Wang[6,*]**

[1]School of Automation, Guangdong University and Technology, Guangzhou, 510006, China

[2]School of Immigration Administration (Guangzhou), China People's Police University, Guangzhou, 510663, China

[3]Department of Mechanical Engineering, College of Engineering, King Faisal University, Al Ahsa, 31982, Saudi Arabia

[4]Cyber Police Division, Guangzhou Public Security Bureau, Guangzhou, 510030, China

[5]School of Forensic Science and Technology, Guangdong Police College, Guangzhou, 510320, China

[6]College of Mathematics and Informatics, South China Agricultural University, Guangzhou, 510642, China

*Corresponding Author: Chuntao Wang. Email: wangct@scau.edu.cn

Received: 05 October 2024 Accepted: 26 November 2024 Published: 03 January 2025

## ABSTRACT

With the rapid advancement of visual generative models such as Generative Adversarial Networks (GANs) and stable Diffusion, the creation of highly realistic Deepfake through automated forgery has significantly progressed. This paper examines the advancements in Deepfake detection and defense technologies, emphasizing the shift from passive detection methods to proactive digital watermarking techniques. Passive detection methods, which involve extracting features from images or videos to identify forgeries, encounter challenges such as poor performance against unknown manipulation techniques and susceptibility to counter-forensic tactics. In contrast, proactive digital watermarking techniques embed specific markers into images or videos, facilitating real-time detection and traceability, thereby providing a preemptive defense against Deepfake content. We offer a comprehensive analysis of digital watermarking-based forensic techniques, discussing their advantages over passive methods and highlighting four key benefits: real-time detection, embedded defense, resistance to tampering, and provision of legal evidence. Additionally, the paper identifies gaps in the literature concerning proactive forensic techniques and suggests future research directions, including cross-domain watermarking and adaptive watermarking strategies. By systematically classifying and comparing existing techniques, this review aims to contribute valuable insights for the development of more effective proactive defense strategies in Deepfake forensics.

## KEYWORDS

Deepfake; proactive forensics; digital watermarking; traceability; detection techniques

## Nomenclature

| | |
|---|---|
| GANs | Generative Adversarial Networks |
| AIGC | Artificial Intelligence Generated Content |
| VAEs | Variational Autoencoders |

| | |
|---|---|
| DCGAN | Deep Convolutional GAN |
| WGAN | Wasserstein GAN |
| PGGAN | Progressive Growing of GAN |
| STGAN | Style Transfer Generative Adversarial Network |
| STU | Selective Transfer Unit |
| DCT | Discrete Cosine Transform |
| DWT | Discrete Wavelet Transform |
| CNN | Convolution Neural Network |
| AF | Artificial Fingerprinting |
| EDA-AF | Enhanced Digital Artificial Fingerprinting |
| PSNR | Peak Signal-to-Noise Ratio |
| SSIM | Structural Similarity Index |
| LPIPS | Learned Perceptual Image Patch Similarity |
| MSE | Mean Squared Error |
| FID | Frechet Inception Distance |
| ACC | Accuracy |
| BER | Bit Error Rate |
| TP | True Positives |
| FP | False Positives |
| LFW | Labeled Faces in the Wild |
| CelebA | CelebFaces Attributes |
| FFHQ | Flickr Faces High Quality |
| FS | Face Swapping |
| FAE | Facial Attribute Editing |
| FR | Face Reenactment |
| FG | Face Generation |

## 1 Introduction

Artificial Intelligence Generated Content (AIGC) technologies have resulted in the widespread proliferation of Deepfake technology [1]. Deepfake is an image synthesis technique that creates convincing fake faces by transferring identity information between original and target faces or by altering attributes of target [2,3]. This technology was initially employed primarily in entertainment software and film production. However, it poses significant potential threats. For instance, Deepfake videos targeting national leaders can quickly spread across domestic and international online platforms, generating widespread public concern and potentially influencing public sentiment. Furthermore, Deepfake technology can be used to alter critical speeches of prominent figures, which may have far-reaching implications for a nation's political, diplomatic, and military dynamics. Beyond political and military ramifications, Deepfake-generated audio and video content severely undermines societal trust, disrupting the daily lives and work of individuals. In response, researchers have been actively developing Deepfake detection and defense technologies. They employ various technical approaches to reliably detect and label Deepfake content, aiming at preventing the spread of malicious information, and have achieved notable results [4,5].

Early defense technologies primarily rely on passive detection methods, which extract information or features from facial images or videos to identify the forged content. These methods are categorized into image-level [6–10] and video-level [11–15] detection techniques. Image-level detection focuses on identifying inconsistencies in spatial and frequency domains, analyzing both local and global cues

to detect alterations in facial images. Deepfake images often show noticeable discrepancies from their surrounding context, making these methods effective for spotting tampering. Conversely, video-level detection techniques assess temporal inconsistencies across different timescales, allowing for a more thorough analysis to determine if a Deepfake video has been manipulated. Despite their high accuracy, passive detection methods face several challenges. They perform poorly when confronted with unknown facial manipulation techniques. Additionally, these methods require large datasets, which results in low computational efficiency. They are also vulnerable to counter-forensic techniques that can eliminate forgery traces and to adversarial noises that can degrade detection accuracy. Furthermore, passive detection methods are reactive and cannot prevent the advanced spread of forged facial images or videos, and their results lack robust evidence to validate facial forgery [16–20].

Furthermore, Deepfake technology is advancing towards greater realism and more sophisticated adversarial capabilities. However, detection methods have not kept pace with these advancements, resulting in inadequate performance for practical applications. To tackle this problem, researchers have proposed proactive digital watermarking methods for Deepfake detection. The core idea is to embed specific digital watermarking into images or videos before they are shared online, which facilitates authenticity verification and traceability, thus providing a preemptive defense. Compared with the passive detection methods, proactive digital watermarking offers four key advantages [21–24]: (1) It allows information owners to detect the forged data in real time and respond promptly, providing a proactive defense, whereas passive detection usually identifies the forgery only after the information has been disseminated. (2) Digital watermarking functions as an embedded defensive measure to detect and track the unauthorized use of data. (3) The watermarking is designed to be challenging to detect and remove, making it difficult for attackers to alter or tamper with the data without detection. (4) Digital watermarking acts as a technical security measure that can provide substantial legal evidence in intellectual property disputes and data security compliance cases.

There has been no comprehensive analysis or summary of proactive forensic techniques for Deepfake based on digital watermarking in the literature. We seek to eliminate this gap by providing a detailed review. It begins with an overview of the research background, including common Deepfake generation techniques and digital watermarking methods. The paper then summarizes and categorizes existing proactive forensic techniques based on digital watermarking. It also discusses widely used datasets and evaluation methods. Finally, the paper addresses the challenges faced by these forensic techniques and suggests future research directions. This review offers valuable insights for developing more effective proactive defense strategies in Deepfake forensics. While existing studies have explored the application of digital watermarking techniques in deepfake detection, this paper provides a more comprehensive perspective by systematically classifying and comparing existing techniques. We not only evaluate the effectiveness of these techniques in practical applications but also deeply analyze their limitations. Furthermore, this paper proposes new perspectives for future research directions, especially in the areas of cross-domain watermarking techniques and adaptive watermarking strategies, which have not been fully explored in existing research.

## 2 Deepfake and Digital Watermarking

### *2.1 Deepfake Generation*

Deepfake generation technologies mainly include Variational Autoencoders (VAEs) [25,26], GANs [27,28], and diffusion models [29,30]. These methods have progressed substantially, achieving high-quality content. In the literature, there are four primary types of Deepfake technologies: face

generation [31], face swapping [32–34], face attribute editing [35–38], and face reenactment [39–41]. Face generation employs GANs to create realistic but non-existent human faces, encompassing features like facial structure, hair, and posture. Face swapping replaces one person's face with another in images or videos. Facial attribute editing alters specific features. Face reenactment transfers expressions from one face to another. Face swapping and face reenactment are currently the most prevalent. However, they also pose significant risks and ethical issues due to their potential for misuse.

### *2.1.1 Face Generation*

Face generation seeks to create images of non-existent individuals. Most underlying networks are GANs and their variants. By exploiting the adversarial interaction between the generator and discriminator, GANs can closely mimic the characteristics of real samples. GANs are inspired by the "zero-sum game" concept in game theory, a method that learns data distributions through adversarial interaction. The generative model generates samples from a given random variable, while the discriminative model predicts whether the data samples belong to the real training set. Both models improve their capabilities through adversarial training, with the ideal state being that the generative model can produce samples indistinguishable from real. The training process of GANs is illustrated in Fig. 1. Popular GANs used for generating synthetic faces include Deep Convolutional GAN (DCGAN) [42], Wasserstein GAN (WGAN) [43], Progressive Growing of GAN (PGGAN) [44], StyleGAN [31], and StyleGAN2 [28]. Examples of face synthesis results from different GAN techniques are shown in Fig. 2, demonstrating highly realistic and detailed images.

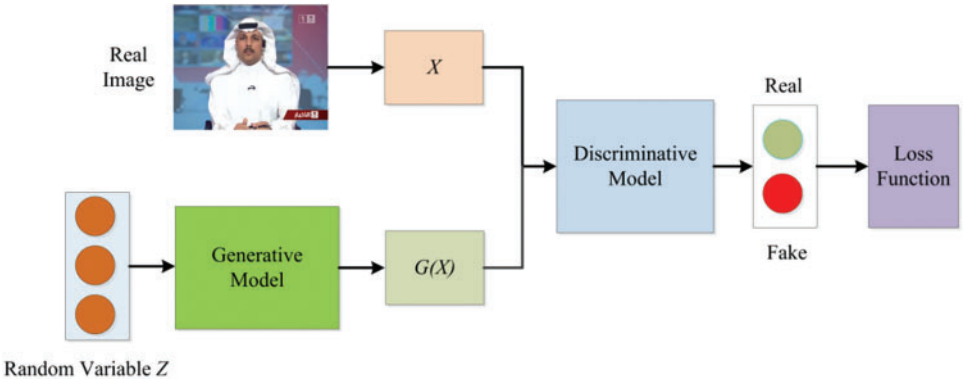


**Figure 1:** Training process of GAN

### *2.1.2 Face Swapping*

Face swapping involves exchanging facial features between two individuals, allowing for an identity switch. This process uses the original person's image to replace the entire head of the target person while preserving other attributes. This technique enables the original individual to appear in different contexts. The face swapping workflow based on deep generative models is illustrated in Figs. 3 and 4. During model training, images or videos of person A and person B are collected. Preprocessing steps such as face extraction, cropping, and alignment are performed. Two weight-sharing encoders, Encoder A and Encoder B, are then trained to extract shared facial attributes from faces A and B. Independent decoders, Decoder A and Decoder B, are subsequently trained to reconstruct each person's unique facial information. In the identity exchange phase, Encoder B encodes the attributes of Person B, and Decoder A reconstructs the image of person B, showing person A's appearance with

person B's identity. Post-processing, including image fusion and video composition, is then applied. Fig. 5 shows the flowchart of the FaceSwap operation.

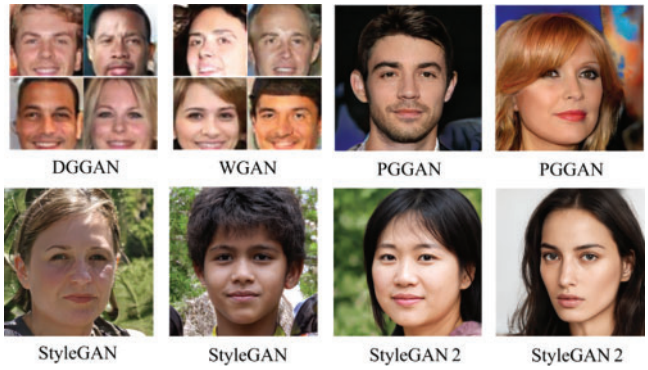


**Figure 2:** Generated images from different face synthesis techniques

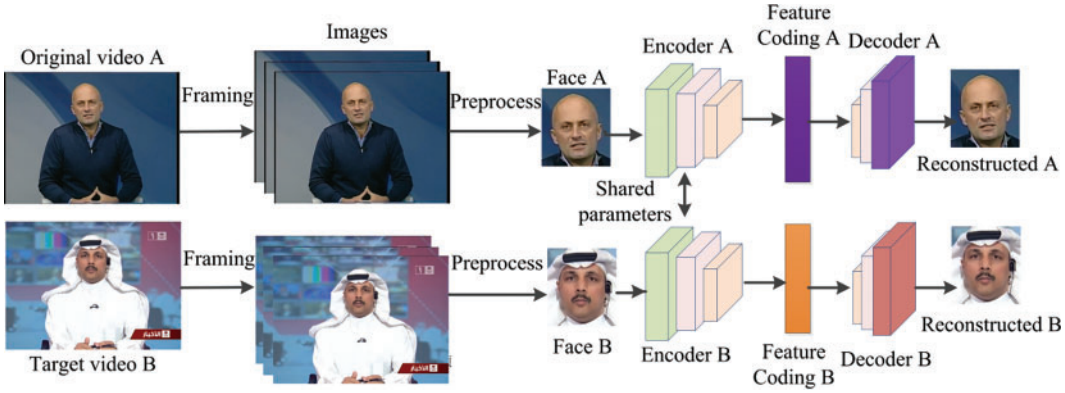


**Figure 3:** Training phase of face swapping using deep generative models

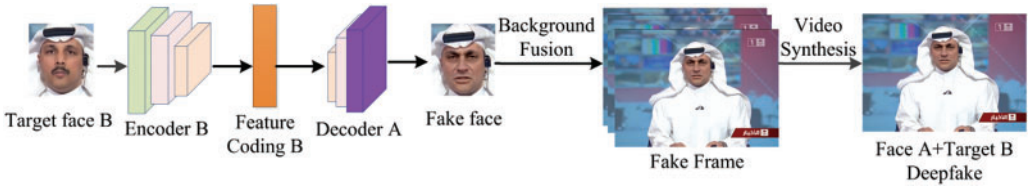


**Figure 4:** Inference stage of face swapping using deep generative models

### *2.1.3 Face Attribute Editing*

Face attribute editing entails altering specific facial features, such as hair color, skin color, gender, age, or adding glasses. Face attribute editing refers to generating new faces with the desired attributes. Arbitrary attribute editing is primarily achieved through a combination of an encoder, a decoder, and a GAN. Liu et al. [45] proposed a Style Transfer Generative Adversarial Network (STGAN) model. This model focuses on the differences between attribute vectors rather than treating the entire attribute vector as a class label. The STGAN employs a U-Net-like network structure for its generator, which is instrumental in enhancing image quality and attribute manipulation capabilities. A key innovation

in this architecture is the use of Selective Transfer Unit (STU) operations during the skip connections. These STU operations adaptively select and modify encoder features before concatenating them with decoder features, thereby optimizing the transfer of relevant information and improving the overall performance of the network. This approach effectively adds a hard decoupling effect to generator, as illustrated in Fig. 6.

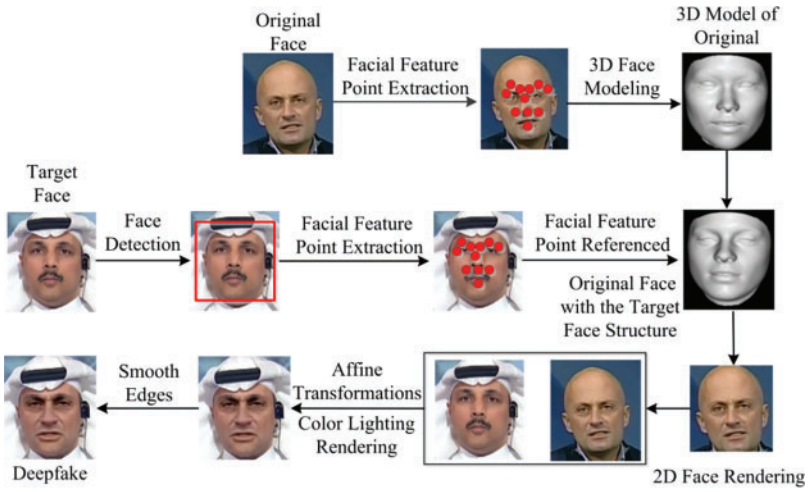


**Figure 5:** Flowchart of the FaceSwap operation

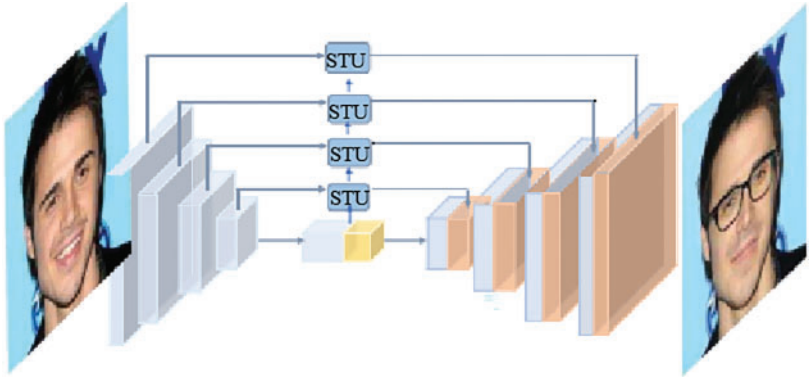


**Figure 6:** Overall structure of STGAN

### *2.1.4 Face Attribute Editing*

Face reenactment transfers one person's facial expressions and postures to another, preserving their original identity. This process enables the modified face to maintain the appearance of the target individual but display the expressions and postures of another person. When combined with specific audio content, it allows the target individual to exhibit expressions and perform actions that do not reflect their true emotions. The facial reenactment workflow based on deep generative models is illustrated in Fig. 7. First, key facial points of the target individual A are extracted and input into an image translation model to reconstruct A's expressions and movements. Next, facial points from the original individual B are used as input to drive the model's expression transfer, resulting in a face that combines A's features with B's expressions. Finally, image blending and video synthesis are

performed as necessary during post-processing. Fig. 8 shows the method of Face2Face. Face2Face is a pioneering real-time face reenactment method that allows for the dynamic transfer of facial expressions from one person to another in videos. This technology leverages deep learning to capture and analyze facial movements and expressions with high accuracy. In the context of deepfake detection, Face2Face represents a significant advancement in the field, as it can be used to both generate and detect manipulated facial expressions in real-time. By understanding the underlying mechanisms of Face2Face, researchers and developers can devise more effective strategies for identifying deepfakes and preserving the authenticity of visual content.

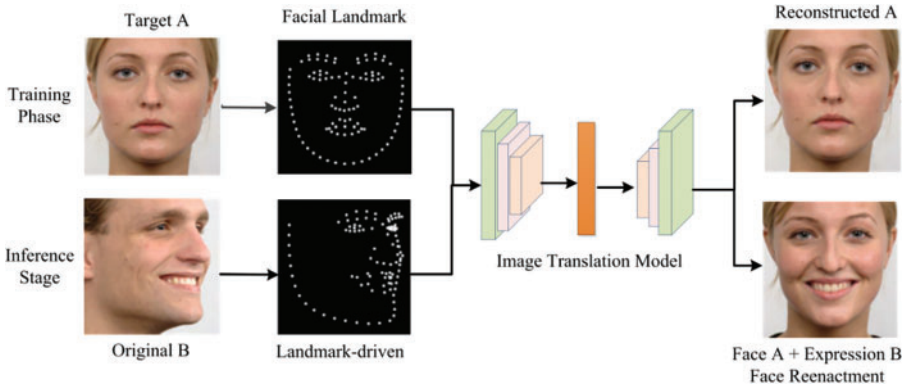


**Figure 7:** Basic principles of face reenactment using deep generative models

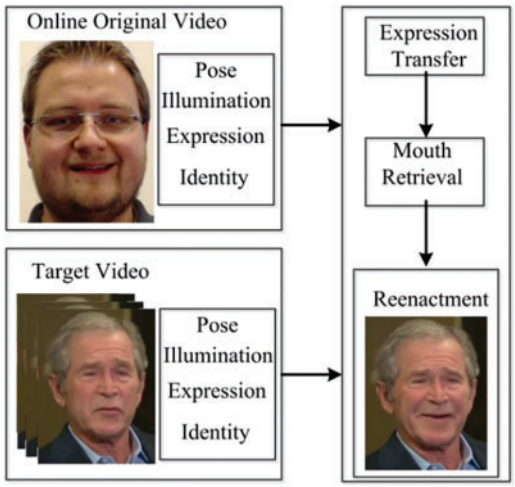


**Figure 8:** Method of Face2Face

### *2.2 Digital Watermarking*

The Digital Watermarking technique covertly embeds information into digital media to authenticate and protect intellectual property and content integrity. In the context of images, it comes in several forms: spatial watermarking, which adjusts pixel values in the spatial domain [46,47]; frequency watermarking, which modifies information in the frequency domain [48,49]; and hybrid watermarking, which integrates both spatial and frequency methods [50]. Watermarking approaches

can be categorized into three types based on their resistance to attacks: robust, fragile, and semi-fragile. The robust watermarking [51,52] is engineered to withstand various attacks and common signal processing operations, maintaining the original content's integrity even after redistribution or modification. In contrast, the fragile watermarking [53] is highly sensitive, and even minor changes can compromise the authentication system, making them ideal for detecting tampering. The semi-fragile watermarking [54,55] combines the advantages of both robust and fragile types. This watermarking can endure normal signal processing during network transmission but reveal tampering if the image is maliciously altered. Fig. 9 presents the general flowchart of watermarking.

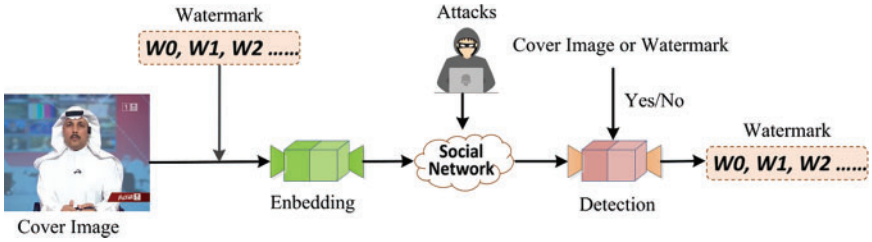


**Figure 9:** General flowchart of watermarking

Many digital watermarking methods have been proposed, embedding watermark information into both the spatial and frequency domains of images and videos. In spatial domain watermarking, watermarks are embedded directly into pixel values using techniques like block-based embedding or least significant bit modification. In frequency domain watermarking, watermarks are embedded by altering coefficients from transformations such as Discrete Cosine Transform (DCT) or Discrete Wavelet Transform (DWT). However, traditional methods face notable limitations in terms of robustness *vs.* fragility, invisibility, and the balance between capacity and quality, necessitating ongoing improvements. Furthermore, current watermark embedding and extraction techniques largely rely on manual design.

Recently, CNN-based neural network watermarking has emerged as an end-to-end solution, replacing manually designed embedding processes with neural network-based encoding. In this method, an encoder inputs an image and a watermark to produce a watermarked image, while a decoder extracts the watermark from it. These encoders and decoders are trained on a dataset of images. For instance, frameworks like StegaStamp [56] and HiDDeN [51] have demonstrated effective watermarking that can robustly conceal and transmit data, ensuring that the embedded information remains recoverable despite various physical and digital distortions. Furthermore, specialized watermarking techniques [57,58] have been developed for large-scale image generation models, such as stable diffusion models.

## 3 Proactive Forensics Techniques Based on Digital Watermarking

### *3.1 Classification*

The concept of proactive forensic techniques for Deepfake detection was initially introduced by Yu et al. [59], who utilized image steganography to embed Artificial Fingerprinting (AF). This approach ensures that the fingerprints can be transferred from the training data to the generated model, aiding in the identification and tracking of forged outputs. Following this, Yu et al. [60] developed a model fingerprinting method using multiple generators with distinct fingerprints to detect and trace generated samples. Kim et al. [61] further proposed a model tracing technique involving a user-specific model retrained with a parameterized secret key, which helps trace the forged content

back to the user's model. However, these methods often struggle with ensuring the robustness of the embedded fingerprints, which are susceptible to distortions from image processing. To enhance robustness, Liao et al. [62] introduced adversarial learning strategies Enhanced Digital Artificial Fingerprinting (EDA-AF) to simulate various distortion conditions, improving the resilience of both artificial and model fingerprints. Fig. 10 illustrates the methods of AF and EDA-AF. AF is a technique where unique digital fingerprints are embedded into the training data of generative models, allowing the generated outputs to be traced back to their source model. EDA-AF is an advanced version of AF that incorporates adversarial learning to enhance the robustness of the embedded fingerprints against various image distortions and manipulations, ensuring the fingerprints remain detectable even in the presence of aggressive attacks aimed at obfuscating the watermarks. Together, these proactive forensic techniques lay a strong foundation for future research.

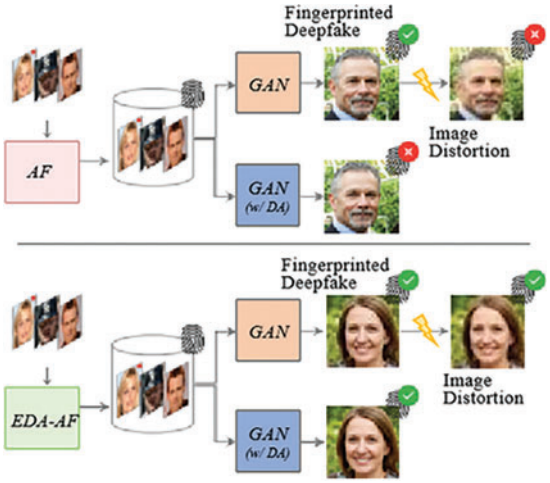


**Figure 10:** Illustrations of AF and EDA-AF method

Digital watermarking offers distinct advantages for proactive evidence collection against Deepfake due to its characteristics of invisibility, robustness, and traceability. As a result, many researchers have focused on developing digital watermarking techniques for Deepfake detection, leading to significant advancements. As illustrated in Fig. 11, this process involves embedding watermark information into images or videos with faces before they are uploaded to online platforms. This preemptive measure enables effective tracing and authenticity verification, even if users create convincing Deepfake. Watermark information can be categorized by strength into robust, fragile, and semi-fragile types. The watermarking process can target different image regions: the face region, the background, or the entire image. Techniques for embedding and extracting watermarking include traditional methods and neural network-based approaches. Digital watermarking for proactive Deepfake detection can be categorized into three defense strategies: robust watermarking techniques, semi-fragile watermarking techniques, and dual-watermarking techniques. Each category will be briefly reviewed.

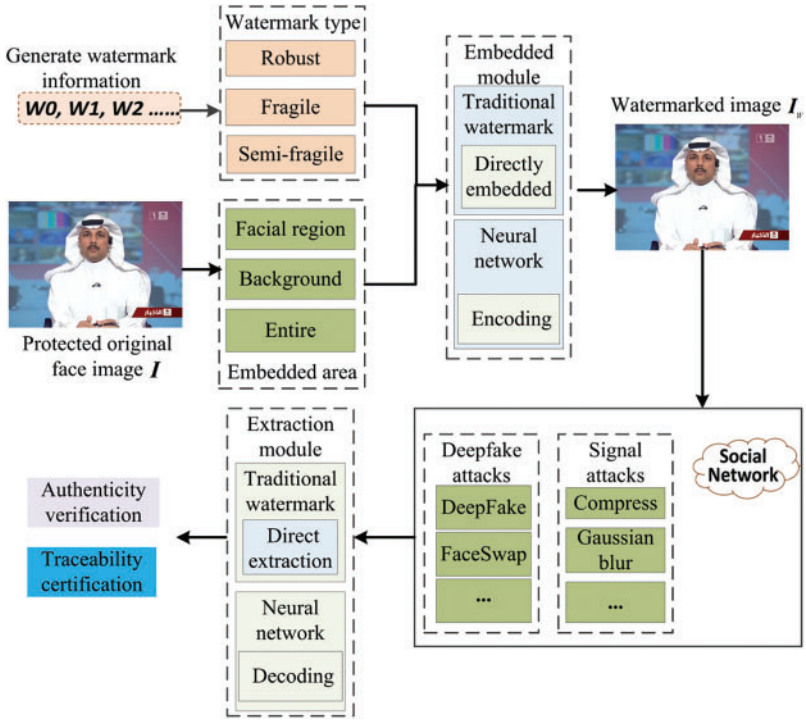


**Figure 11:** Schematic of the proactive forensic framework for Deepfake based on digital watermarking

### *3.2 Proactive Forensics Techniques Based on Robust Watermarking*

Proactive forensics technology using robust watermarking embeds watermark information into the original image, ensuring any modifications cause a mismatch between the embedded data and the visible content. Various methods exist within this framework. For instance, Wang et al. [63] developed FakeTagger, which uses the embedded message content to enhance facial security and privacy, as illustrated in Fig. 12. This method embeds the information within the victim's image and recovers it after deepfaked media are generated, allowing detection of manipulations by GANs. Wang et al. also identified three main challenges: (1) generalizing across different forgery types and recovering embedded information from unseen forgeries; (2) maintaining robustness against conventional image operations; (3) concealing the embedded information. While FakeTagger is effective in embedding and recovering messages, it is limited to the inference phase and does not influence the forgery model itself, which may lead to the misclassification of genuine images subjected to post-processing. To overcome these limitations, Sun et al. [64] introduced FakeTracer, a model that embeds sustainable and erasable traces into facial images using autoencoders. Before uploading images or videos to social networks, users can embed specific traces into their content. Deepfake models trained on such content will incorporate sustainable traces while disregarding the erasable ones. The presence of these traces can be effectively detected, thereby providing a robust mechanism for identifying forged content and offering substantial protection against Deepfake manipulations.

To ensure semantic-level protection for facial images and prevent impersonation through identity manipulation, Zhao et al. [65] introduced a watermarking mechanism. This approach embeds a watermark as an anti-counterfeiting label within facial identity features through two main steps: watermark injection and verification, as shown in Fig. 13. The watermark is intricately integrated

with facial identity features, making it sensitive to manipulation while robust against common image modifications such as resizing and compression. However, this technique requires a high-intensity watermark to ensure effective detection. To further protect facial image and video owners' rights, Lin et al. [66] developed an end-to-end self-encoding method known as SIDT (Source-ID-Tracker), inspired by data hiding principles. This technique enables implicit embedding of the original facial image into Deepfake results without noticeable visual changes. It integrates the entire original facial image into the target face and uses a distortion simulation layer to mimic social media's lossy channels, meeting the high perceptual quality and embedding capacity requirements for traceability and forensics. Similarly, Shen et al. [67] proposed FHnet, a proactive facial hiding network designed to protect facial features. This network extracts and encodes facial components using information hiding techniques, embedding them within the remaining background. This allows for full recovery of the original face, unaffected by variations from different face-swapping algorithms, thereby providing better generalization.

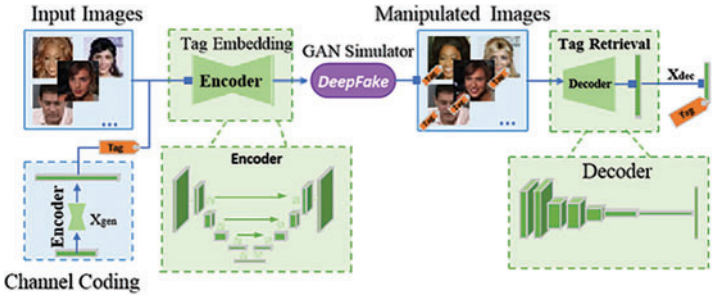


**Figure 12:** Illustrations of FakeTagger

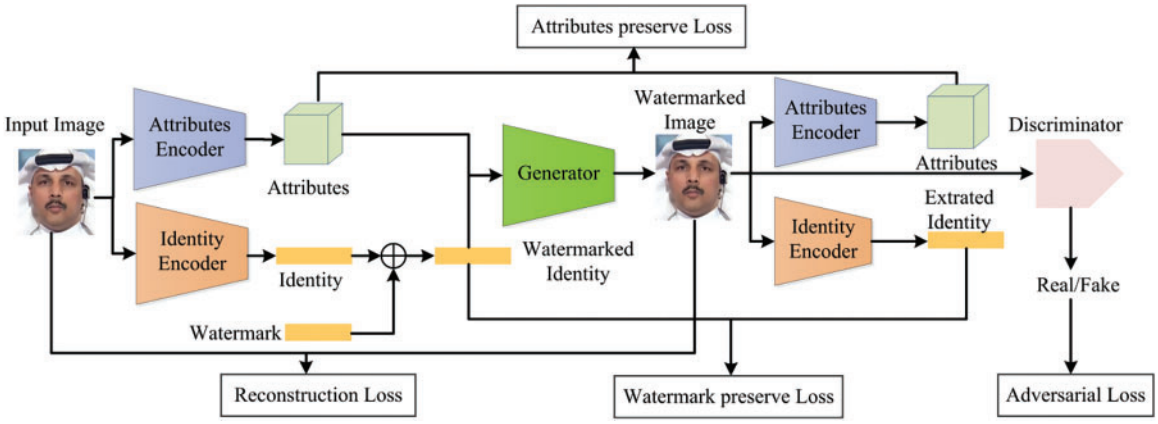


**Figure 13:** Illustrates of the proactive forensics method based on identity watermarking

Furthermore, Wang et al. [68] developed the first robust identity-aware watermarking framework. This framework is designed to detect Deepfake facial swaps while also tracking their sources. It integrates identifying semantics related to the image content into the watermark and employs an unpredictable, irreversible chaotic encryption system to ensure the confidentiality of the watermark. However, this approach has limitations in detecting facial reproduction operations that alter expressions and poses while preserving the facial identity. To address these limitations, Zhang et al. [69]

introduced the dual defense, a novel proactive defense method based on robust adversarial watermarking. This technique embeds a robust adversarial watermark into the facial image in a single process, disrupting Deepfake models and enabling copyright tracking. As face-swapping technology evolves, models based on GANs and diffusion processes are becoming more common, making it essential not just to verify authenticity but also to understand image origins. Wu et al. [70] explored the interplay between proactive watermark injection and passive detection, noting that pre-injected watermarking can interfere with detection processes, leading to an increase in false negatives where watermarked fake images are misclassified as real. Then, they proposed AdvMark, a harmless proactive forensic scheme that uses adversarial fine-tuning to convert robust watermarking into adversarial watermarking. This method aims at improving both the traceability of watermarked images and their detectability.

While these methods enhance forensic capabilities for detecting Deepfake content, they have notable limitations under complex attack scenarios. A key limitation is their dependence on robust assumptions about Deepfake generation, which often requires high-intensity watermarking for effective detection. Additionally, proactive forensic methods based on robust watermarking may not generalize well to unseen datasets and can struggle with new forgery techniques, limiting their overall effectiveness.

### *3.3 Proactive Forensic Techniques Based on Semi-Fragile Watermarking*

The proactive forensics technique using semi-fragile watermarking is robust against conventional signal attacks like compression and minor adjustments. However, it is susceptible to malicious manipulations like Deepfake, which can compromise the watermark and jeopardize image integrity. Yang et al. [71] introduced the FaceGuard framework, employing a neural network-based semi-fragile watermarking method. This framework employs an encoder to merge a facial image with a watermark, creating a watermarked image. A decoder then processes this image to extract a binary vector representing the watermark. Although FaceGuard is resilient to common image post-processing techniques, it struggles with Deepfake operations, limiting its effectiveness in detecting forged images. Notably, it can only authenticate images with detectable traces, failing to differentiate between authentic and synthetic images. In contrast, Neekhara et al. [72] developed the FaceSigns framework, which embeds semi-fragile watermarking specifically in facial regions. This method renders the embedded information irretrievable in Deepfake without needing a Deepfake-specific training dataset. However, it only confirms absence of a watermark and cannot verify if the video has been altered. Beuve et al. [73] introduced WaterLo, an end-to-end model featuring a local semi-fragile watermark, as shown in Fig. 14. WaterLo embeds the watermark across the entire image, allowing for the detection of modified areas where the watermark has been removed from the facial region, while remaining visible elsewhere. It also includes a compression module to enhance robustness against compression-related attacks.

Semi-fragile watermarking can indicate whether an image or video is authentic or forged, but it tends to be unstable when subjected to common image processing operations. Furthermore, semi-fragile watermarking cannot trace the source of the target image, complicating the creation of a complete evidence chain in forensic investigations of Deepfake-related cybercrimes. This issue is critical because it is essential not only to address the forgery but also to obtain information about the source material. Thus, tracking the original target image and performing detection tasks are crucial for maintaining the integrity of the evidence.

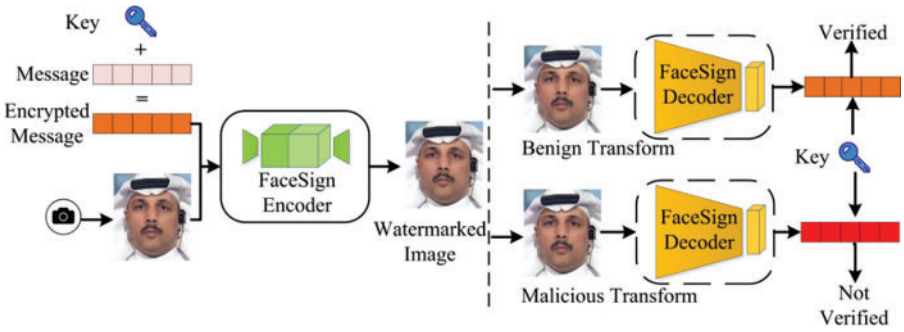


**Figure 14:** Authentication based on semi-fragile watermarking

### *3.4 Proactive Forensic Techniques Based on the Dual Watermarking*

Single-function deep watermarking methods fall short of meeting the needs of proactive forensics. The robust watermarking is so resilient that it can be accurately extracted from both the original and Deepfake images. Furthermore, even when an image is manipulated, the source information remains intact. Conversely, fragile watermarking can reveal if the content has been maliciously altered but cannot provide information about the original image's source after manipulation, as the watermark cannot be reliably extracted post forgery. This makes it challenging to differentiate between forged images and those without embedded watermarking. To address both challenges, an intuitive approach is to embed robust and fragile watermarking. This method might require additional segmentation models, which could introduce prior knowledge and make the watermarking more detectable. To tackle these challenges, Liu et al. [74] presented the BiFPro, which combines fragile and robust watermarking techniques for comprehensive facial data protection in Deepfake applications. Wu et al. [75] proposed a deep separable watermarking method called SepMark, as illustrated in Fig. 15. This framework introduces a new paradigm where a single embedding process through an encoder enables the extraction of watermarking with varying robustness levels using robust and fragile decoders. Zhang et al. [76] developed EditGuard, a multifunctional proactive forensics method. They framed tampering localization and copyright protection as a joint image bit-steganography problem, using a serial encoding and parallel decoding structure.

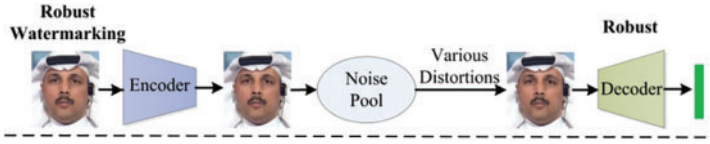


**Figure 15:** (Continued)

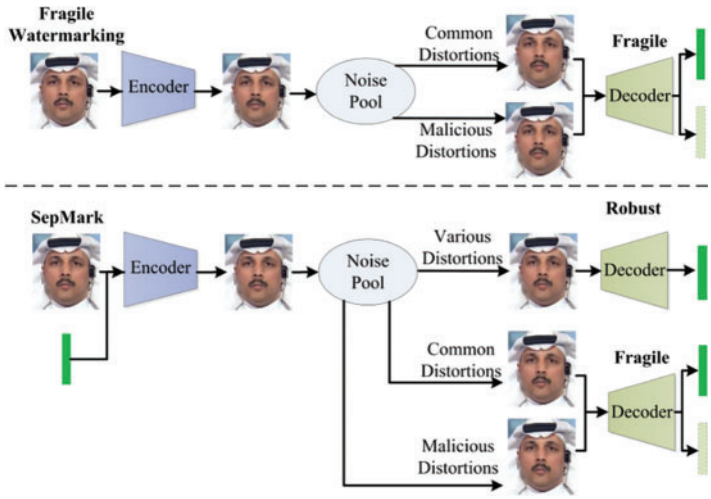


**Figure 15:** Deepfake detection and traceability based on depth-separable watermarking

### *3.5 Summary of this Section*

This section reviews prevalent proactive forensic methods based on various criteria, including algorithm types, publication years, embedding regions, information formats, watermarking functions, and key characteristics, as outlined in Table 1. Methods that use robust watermarking can enhance the detection of Deepfake content to a certain degree. Nonetheless, these methods face significant challenges in complex attack scenarios.

**Table 1:** Proactive forensic techniques for Deepfake based on digital watermarking

| Task type | Mode | Year | Embedded area | Form of information | Stage of action | Application |
|---|---|---|---|---|---|---|
| Proactive forensics based on robust watermarking | AF [59] | 2021 | Entire image | Fingerprints | Training | Detection, Traceability |
| | FakeTagger | 2021 | Facial region | Binary sequence watermark | Inference | Detection, Traceability |
| | EDA-AF | 2022 | Entire image | Fingerprints | Training | Detection, Traceability |
| | FakeTracer | 2022 | Facial region | Binary sequence watermark | Training | Detection |
| | SIDT | 2022 | Facial region | Image watermark | Inference | Traceability |
| | PDIW | 2023 | Facial region | Binary sequence watermark | Inference | Detection |
| | FHnet | 2024 | Background | Image watermark | Training | Traceability |

(Continued)

**Table 1 (continued)**

| Task type | Mode | Year | Embedded area | Form of information | Stage of action | Application |
|---|---|---|---|---|---|---|
| | RIPW | 2024 | Facial region | Image watermark | Training | Detection, Traceability |
| | Dual defense | 2024 | Facial region | Binary sequence watermark | Training | Detection, Traceability |
| | AdvMark | 2024 | Facial region | Image watermark | Inference | Detection, Traceability |
| Proactive forensics based on semi-fragile watermarking | FaceGuard | 2021 | Facial region | Binary sequence watermark | Inference | Detection |
| | FaceSigns | 2023 | Facial region | Binary sequence watermark | Training | Detection |
| | WaterLo | 2023 | Entire image | Image watermark | Inference | Detection |
| Proactive forensics based on dual watermarking | BiFPro | 2023 | Facial region | Image watermark | Inference | Detection, Traceability |
| | SepMark | 2023 | Facial region | Binary sequence watermark | Inference | Detection, Traceability |
| | EditGuard | 2024 | Entire image | Binary sequence watermark | Training | Detection, Traceability |

They rely on robust assumptions about Deepfake generation, necessitating high-intensity watermarking for effective detection. Furthermore, these methods often lack the ability to generalize when confronted with unknown forgery techniques, particularly with unfamiliar datasets. On the other hand, proactive forensic methods that employ semi-fragile watermarking are useful for determining whether an image or video is authentic or forged based on the watermark's presence. However, these methods may become unstable when exposed to common image processing operations. Additionally, the semi-fragile watermarking does not facilitate the tracing of the source of the victim's images. Methods utilizing dual watermarking offer the advantage of performing both source tracing and Deepfake detection. However, the interaction between the two watermarkings can negatively impact each other, potentially leading to a decrease in visual quality.

## 4 Model Evaluation Results

### *4.1 Evaluation Metrics*

As outlined in previous sections, proactive forensic methods using digital watermarking aim at achieving two key objectives: the proactive detection of Deepfake attacks and the authentication of the provenance of facial images. Consequently, evaluating these forensic models typically involves assessing the visual quality of watermarked facial images, the efficacy of Deepfake detection, and the robustness of provenance verification.

#### 4.1.1 Visual Quality Assessment Metrics

To evaluate the fidelity of watermarked facial images, four primary metrics were employed. These metrics include two conventional measures: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [77]. Additionally, two metrics based on features extracted by deep learning models were utilized: Learned Perceptual Image Patch Similarity (LPIPS) [78] and Frechet Inception Distance (FID) [79]. They are described as follows:

(1) PSNR is a widely utilized metric for evaluating image quality, with higher values indicating lower distortion and superior quality. PSNR is measured in decibels (dB), and specific ranges of values correspond to different levels of image quality. A PSNR value above 40 dB suggests that the image quality is nearly visually identical to the original image. Values between 30 and 40 dB indicate acceptable levels of distortion, while values between 20 and 30 dB reflect poor quality. Values below 20 dB signify severe distortions. Despite its widespread use, PSNR has limitations and is typically employed to assess image quality relative to the maximum signal and background noise. For an original image $I$ of size $m \times n$ and a noisy image $K$, the mean squared error (MSE) is defined as:

$$MSE = \frac{1}{mn} \sum_{i=o}^{m-1} \sum_{j=o}^{n-1} [I(i,j) - K(i,j)]^2 \tag{1}$$

Building on this, the formula for calculating PSNR is defined as follows:

$$PSNR = 10 \log_{10} \frac{{MAX_I}^2}{MSE} \tag{2}$$

where $MAX_I$ represents the value pixel value in the original image $I$.

(2) SSIM quantifies the similarity of structural information in images as perceived by human eyes, providing a more accurate representation of visual perception. SSIM evaluates images based on three primary components: luminance, contrast, and structure. For a reference image $I$ of size $m \times n$ and a noisy image $K$, SSIM is defined as follows:

$$SSIM = \frac{(2\mu_I\mu_K + C_1)(2\sigma_{IK} + C_2)}{(\mu_I^2 + \mu_K^2 + C_1)(\sigma_I^2 + \sigma_K^2 + C_2)} \tag{3}$$

The term $\mu_I$ and $\mu_K$ represents the average pixel values of images $I$ and $K$, calculated using the following formula:

$$\mu_I = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j)] \tag{4}$$

The variance of the pixel values of images $I$ and $K$ is represented by $\sigma_I^2$ and $\sigma_K^2$, and its calculation formula is as follows:

$$\sigma_I^2 = \frac{1}{mn-1} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - \mu_i]^2 \tag{5}$$

The covariance of $\sigma_{IK}$ represent the pixels in images $I$ and $K$:

$$\sigma_{IK} = \frac{1}{mn-1} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - \mu_I][K(i,j) - \mu_K] \tag{6}$$

Additionally, $C_1$ and $C_2$ are two parameters calculated as $C_1 = (D_1L)^2$ and $C_2 = (D_2L)^2$, where $C_1$ and $C_2$ are set to 0.001 and 0.003. The value of $L$ is related to the image type, which is set to 255 as the images used in the experiment are of type uint8.

(3) LPIPS, known as perceptual loss, quantifies perceptual similarity between image patches. This metric evaluates differences between two images by guiding the generator to learn the inverse mapping from generated images to the ground truth, thereby emphasizing perceptual similarity. Compared to conventional metrics such as PSNR and SSIM, LPIPS better reflects human visual perception. A lower LPIPS value indicates higher similarity between images, while a higher value denotes greater dissimilarity. The calculation method is as follows:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_i \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \right\|_2 \tag{7}$$

In this context, the output from the original image $x$ is referred to as the forged output, whereas the output from the watermarked image $x_0$ is termed the forged output after watermark embedding. The distance between $x$ and $x_0$ is denoted $d$, $l$ corresponds to a specific layer in the feature stack extracted by the neural network. The features produced by this layer are represented as $\hat{y}^l$ and $\hat{y}^l_0$, where $H_l$ and $W_l$ denote the height and width of the feature map, respectively. The vector $w_l$ scales the number of activated channels, and $\odot$ stands for the element-wise product.

(4) FID is a metric used to evaluate the distance between two sets of samples in feature space. Initially, features are extracted using the Inception network. These features are then modeled using a Gaussian distribution. The distance between these Gaussian distributions is subsequently computed to determine the FID value. A lower FID value indicates a smaller distance between the sample sets, which in turn reflects higher sample quality. The calculation method is as follows:

$$d^2((\mu, \varepsilon), (\mu_w, \varepsilon_w)) = |\mu - \mu_w|^2 + tr(\varepsilon + \varepsilon_w - 2(\varepsilon\varepsilon_w)^{1/2}) \tag{8}$$

In this context, $\mu$ and $\mu_w$ represent the means of the two samples, while $\varepsilon$ and $\varepsilon_w$ denote theirrespective covariances.

### *4.1.2 Performance Evaluation Metrics*

To evaluate the performance of the model in detecting Deepfake, one key metrics are utilized: accuracy (ACC). ACC is used to evaluate the accuracy of predictions, represents the percentage of samples that are correctly predicted out of the total number of samples. The calculation method for accuracy is as follows:

$$ACC = \frac{TP}{TP + FP} \tag{9}$$

Here, $TP$ (True Positives) refers to the number of samples that are correctly predicted as the positive class. $FP$ (False Positives) refers to the number of samples that are incorrectly predicted as the positive class.

### *4.1.3 Performance Evaluation Metrics for Traceability*

To evaluate the traceability performance of the model, the primary evaluation metric employed is the average bit error rate (BER). This metric quantifies the percentage of altered bits caused by noise and interference during transmission, relative to the total number of bits in the received data

stream. Specifically, the BER is determined by comparing the binary watermark sequence extracted from the watermarked image with the originally embedded watermark information. A smaller value signifies fewer transmission errors and greater robustness. Suppose that the embedded watermark is $\omega \in \{0,1\}^{B\times L}$ and the extracted watermark is $\tilde{w} \in \{0,1\}^{B\times L}$. Then $BER$ is computed as follows:

$$BER(\omega,\tilde{w}) = \frac{1}{R}\times\frac{1}{L}\times\sum_{i=1}^{B}\sum_{j=1}^{L}\left|\omega^{i\times j}-\tilde{w}^{i\times j}\right|\times 100\% \tag{10}$$

And the bitwise accuracy is defined as:

$$Bitwiseaccuray(\omega,\tilde{w}) = 1-\frac{1}{R}\times\frac{1}{L}\times\sum_{i=1}^{B}\sum_{j=1}^{L}\left|\omega^{i\times j}-\tilde{w}^{i\times j}\right| \tag{11}$$

### *4.2 Datasets*

In the study of proactive forensics techniques for Deepfake, researchers primarily use standard facial image and video datasets for training, testing, and evaluating Deepfake forensic algorithms. Table 2 summarizes the commonly used datasets.

**Table 2:** Public datasets for proactive forensics

| Dataset | Type | Year | Scale | Link |
|---|---|---|---|---|
| LFW | Image | 2008 | 13,233 | http://vis-www.cs.umass.edu/lfw/ (accessed on 10 September 2024) |
| CASIA-WebFace | Image | 2014 | 494,414 | http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html (accessed on 10 September 2024) |
| CelebA | Image | 2015 | 202,599 | http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 10 September 2024) |
| CelebA-HQ | Image | 2017 | 30,000 | https://github.com/tkarras/progressive_growing_of_gans (accessed on 10 September 2024) |
| VGGFace2 | Image | 2018 | 33,000,000 | https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/ (accessed on 10 September 2024) |
| FFHQ | Image | 2019 | 70,000 | https://github.com/NVlabs/ffhq-dataset (accessed on 10 September 2024) |
| FaceForensics++ | Video | 2019 | Real 1000 Fake 4000 | https://github.com/ondyari/FaceForensics (accessed on 10 September 2024) |
| CelebAMask-HQ | Image | 2020 | 30,000 | https://github.com/switchablenorms/CelebAMask-HQ (accessed on 10 September 2024) |

(Continued)

**Table 2 (continued)**

| Dataset | Type | Year | Scale | Link |
|---|---|---|---|---|
| Celeb-DF | Video | 2020 | Real 590<br>Fake 5639 | https://github.com/yuezunli/celeb-deepfakeforensics (accessed on 10 September 2024) |

(1) Labeled Faces in the Wild (LFW) dataset [80] is a benchmark for face recognition, consisting of 13,233 images from natural scenes. It presents challenges due to variability in pose, lighting, expression, age, and occlusion. Some images include multiple faces; only the central face is used as the target, while others are considered background noise. The dataset covers 5749 individuals, with most having a single image. Images are mostly in color, sized at 250 × 250 pixels, with a few in black and white.

(2) The CASIA-WebFace dataset [81], developed by the Institute of Automation at the Chinese Academy of Sciences, comprises 494,414 images representing 10,575 identities. It features diverse faces across ages, genders, ethnicities, and expressions, with challenging lighting and pose conditions. The high-quality images make it suitable for algorithm training and evaluation.

(3) The CelebA (CelebFaces Attributes) dataset [82], provided by The Chinese University of Hong Kong, includes 202,599 images of 10,177 identities. Each image provides a face bounding box, coordinates for five facial landmarks, and labels for 40 attributes, useful for attribute editing.

(4) The CelebA-HQ dataset [83] is an enhanced version of CelebA, containing 30,000 high-resolution face images in resolutions of 256 × 256, 512 × 512, and 1024 × 1024 pixels.

(5) The VGGFace2 dataset [84] includes 3.31 million images of 9131 identities, each represented by multiple images. It covers a wide range of poses, ages, and ethnicities, minimizing noise.

(6) The FFHQ (Flickr-Faces-High-Quality) dataset [31] contains 70,000 high-quality PNG (Portable Network Graphics) images of faces, each at a resolution of 1024 × 1024 pixels. It features diverse age, ethnicities, and rich facial attributes including gender, skin tone, expression, hairstyle, and accessories.

(7) The CelebAMask-HQ dataset [85] contains over 30,000 images at 512 × 512 resolution, annotated with 19 attribute features covering facial components and decorative items.

(8) The Faceforensics++ dataset [86] comprises 1000 real videos sourced from YouTube, featuring frontal faces with a resolution of 480p or higher. Additionally, it includes 4000 Deepfake videos produced using four different techniques: Face2Face, FaceSwap, Deepfakes, and NeuralTextures. These Deepfake videos are available in three different compression levels: uncompressed (C0), compression rate 23 (C23), and compression rate 40 (C40).

(9) The Celeb-DF dataset [87], created for Deepfake detection, includes 590 celebrity videos from YouTube and 5639 high quality Deepfake videos, featuring a range of ages, ethnicities, and genders, with face resolutions of 256 × 256 pixels.

### *4.3 Model Performance Analysis*

To illustrate the key designs and performance of various proactive forensic algorithms tailored to different requirements and scenarios, this study selects several representative algorithms for replication

and comparative experiments. The proposed method is implemented using the PyTorch deep learning framework with a batch size of 50 on an NVIDIA GTX 3090 GPU platform.

#### *4.3.1 Evaluation of Deepfake Detection*

To evaluate the effectiveness of the digital watermarking-based proactive forensics method for Deepfake detection, we initially selected 50,000 authentic facial images from the CelebA and CelebA-HQ datasets. Among these, 35,000 images were used for training, 10,000 for validation, and 5000 for testing. We then generated 1000 counterfeit facial images with each of the four Deepfake methods described in Section 3, resulting in a total of 4000 counterfeit images for the test set. The Deepfake methods used were Face Swapping (FS), Facial Attribute Editing (FAE), Face Reenactment (FR), and Face Generation (FG). All images were resized to 256 × 256 pixels. Four representative algorithms were selected for comparative analysis, ensuring a consistent experimental setup for training and testing. The results of Deepfake detection performance across different algorithms are shown in Table 3. The findings reveal that detection performance was generally better on the CelebA-HQ dataset compared to the CelebA dataset. Notably, the SepMark detection method achieved an average detection rate of 97.69% on the CelebA dataset and an average accuracy of 98.38% on the CelebA-HQ dataset. Unlike conventional passive detection methods, the proactive forensics approach converts the challenge of detecting Deepfake in an open world into a digital watermark detection and matching problem. This method consistently demonstrated superior performance, with an average detection rate exceeding 90% across all scenarios.

**Table 3:** Detection accuracy of different models for Deepfake detection (%)

| Method | CelebA | | | | | CelebA-HQ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | FS | FAE | FR | FG | Average | FS | FAE | FR | FG | Average |
| AF | 85.22 | 88.88 | 90.56 | 100 | 91.16 | 88.79 | 90.23 | 92.32 | 100 | 92.83 |
| FakeTagger | 94.86 | 91.25 | 95.32 | 92.81 | 93.56 | 96.80 | 93.12 | 96.22 | 93.85 | 94.99 |
| PDIW | 97.50 | 88.00 | 90.23 | 91.56 | 91.82 | 99.02 | 89.50 | 92.34 | 93.18 | 93.51 |
| SepMark | 98.22 | 94.50 | 98.46 | 99.60 | 97.69 | 99.00 | 95.56 | 99.15 | 99.80 | 98.38 |

#### *4.3.2 Evaluation of Traceability Forensics*

The traceability performance of the digital watermark-based proactive forensics method was evaluated using an experimental setup akin to that described in Section 4.3.1. This evaluation focused on comparing watermark information extracted from forged faces with that embedded in original faces. Detailed traceability results for various algorithm models are presented in Table 4. The results indicate that the five representative algorithms exhibit superior traceability performance on the CelebA-HQ dataset compared with their detection performance on the CelebA dataset. The RIPW (Robust identity perceptual watermark) method shows an average traceability error rate of 3.50% on the CelebA dataset and 2.42% on the CelebA-HQ dataset. Overall, the proactive forensics approach demonstrates strong traceability performance, with average error rates generally below 10%.

In our experiments, we aimed to compare the performance of different watermarking techniques, which included both image watermark information (SIDT, FHnet, RIPW) and binary sequence watermark information (FakeTagger, SepMark). To make these methods comparable, we standardized

the evaluation metrics and the experimental conditions for all techniques. Specifically, for the image watermark methods, we converted the embedded watermarks into a binary format for consistency in analysis. This involved a thresholding process where the watermarked images were binarized to create a binary sequence that could be compared against the binary watermarks embedded by FakeTagger and SepMark. To ensure fairness in the comparison, we also normalized the watermark detection sensitivity and robustness across all methods. This was achieved by adjusting the watermark strength and the detection algorithms to ensure that the binary sequences and image watermarks were detectable under the same set of conditions. Additionally, we ensured that the watermark extraction process was blind to the type of watermark, meaning that the same extraction algorithm was applied to both types of watermarks, allowing for a direct comparison of detection accuracy and robustness. By standardizing the evaluation process in this manner, we were able to compare the performance of the different watermarking techniques on an equal footing, providing a comprehensive analysis of their effectiveness in detecting and tracing deepfakes.

**Table 4:** Traceability performance of different models (%)

| Method | CelebA | | | | | CelebA-HQ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | FS | FAE | FR | FG | Average | FS | FAE | FR | FG | Average |
| FakeTagger | 5.45 | 12.11 | 5.14 | 7.12 | 7.45 | 3.20 | 11.00 | 3.80 | 6.20 | 6.05 |
| SIDT | 8.68 | 6.79 | 7.45 | 9.77 | 8.17 | 7.48 | 5.88 | 6.34 | 7.42 | 6.78 |
| SepMark | 13.82 | 10.25 | 9.78 | 8.44 | 10.57 | 11.44 | 9.84 | 9.12 | 7.58 | 9.50 |
| FHnet | 4.20 | 3.78 | 4.22 | 5.44 | 4.41 | 3.50 | 2.88 | 4.02 | 4.56 | 3.74 |
| RIPW | 3.14 | 2.44 | 3.66 | 4.58 | 3.50 | 1.98 | 1.88 | 2.68 | 3.14 | 2.42 |

#### *4.3.3 Evaluation of Visual Quality*

In practical applications, proactive watermarking methods should minimize their impact on image quality after watermark embedding. To evaluate the effects of various proactive watermarking methods on visual quality, we selected 5000 real face images from both the CelebA and CelebA-HQ datasets. watermarking was embedded into the original images using different algorithm models. For each pair of images (original and watermarked), we measured the PSNR, SSIM, and LPIPS. Results, detailed in Table 5, show that the watermarked images from existing methods retain high visual quality, with PSNR values exceeding 30 dB, SSIM values above 0.9, and most algorithms achieving LPIPS scores below 0.1.

**Table 5:** Visual quality performance of the model across different datasets

| Method | CelebA | | | CelebA-HQ | | |
|---|---|---|---|---|---|---|
| | PSNR (dB)↑ | SSIM↑ | LPIPS↓ | PSNR (dB)↑ | SSIM↑ | LPIPS↓ |
| AF | 36.93 | 0.965 | 0.0526 | 30.69 | 0.915 | 0.0556 |
| EDA-AF | 35.57 | 0.960 | 0.1022 | 36.56 | 0.922 | 0.1206 |
| SIDT | 33.01 | 0.968 | 0.0301 | 38.81 | 0.972 | 0.0202 |
| PDIW | 33.32 | 0.945 | 0.0802 | 34.25 | 0.956 | 0.0318 |

(Continued)

**Table 5 (continued)**

| Method | CelebA | | | CelebA-HQ | | |
|---|---|---|---|---|---|---|
| | PSNR (dB)↑ | SSIM↑ | LPIPS↓ | PSNR (dB)↑ | SSIM↑ | LPIPS↓ |
| FHnet | 39.56 | 0.958 | 0.0116 | 40.97 | 0.984 | 0.0158 |
| RIPW | 40.25 | 0.982 | 0.0331 | 45.38 | 0.994 | 0.0362 |
| Dual defense | 32.31 | 0.918 | 0.2101 | 36.20 | 0.922 | 0.2550 |
| FaceSigns | 35.43 | 0.962 | 0.0882 | 39.99 | 0.889 | 0.0838 |
| BiFPro | 39.56 | 0.985 | 0.0203 | 40.86 | 0.986 | 0.0108 |
| SepMark | 38.51 | 0.958 | 0.0028 | 38.56 | 0.933 | 0.0080 |

We observed that the performance on the CelebA-HQ dataset is generally better than that on the CelebA dataset. This can be attributed to several factors. Firstly, the CelebA-HQ dataset consists of higher resolution images, which provide more detailed facial features. The increased clarity allows our digital watermarking techniques to embed more robust and distinct watermarks that are less susceptible to being obscured by compression or other forms of image degradation. Secondly, the CelebA-HQ dataset, with its larger and more diverse collection of faces, offers a more comprehensive training ground for our models. This diversity helps in enhancing the generalizability of our techniques, leading to improved performance in detecting deepfakes. Lastly, the CelebA-HQ dataset benefits from a more sophisticated annotation process, which aligns well with the proactive forensic techniques that rely on precise facial attribute manipulation and authentication. The higher quality annotations facilitate more accurate watermark embedding and extraction, contributing to the better performance observed.

#### *4.3.4 Evaluation of Robustness*

To evaluate the robustness of various methods and verify the effectiveness of proactive forensic techniques in detecting digital watermarking under different image and model perturbations, we simulated four types of image disturbances: additive Gaussian noises, Gaussian blurring, JPEG (Joint Photographic Experts Group) compression, and central cropping. For additive Gaussian noise, the standard deviation ranged from 0.0 to 0.5 with a step size of 0.1055. The Gaussian blurring kernel size varied from 0.0 to 15 with a step size of 1.6. For JPEG compression, the quality factors were from 100 to 10 with a decrement of 10. For resizing and cropping, images were initially cropped to dimensions ranging from 64% to 128% of the original size, with a step size of 10%, and then resized to 128 $\times$ 128 pixels. Each sequence of watermarked images underwent these operations to generate the corresponding processed images. Different proactive forensic methods were then employed to detect the watermarking in these processed images. Detection accuracy was assessed by comparing the decoded binary sequence with the original input sequence, measuring bit accuracy. Fig. 16 summarizes the experimental results under these conditions. The results indicate that the accuracy of digital watermark detection decreases monotonically with increasing levels of perturbation. For smaller disturbances, the decline in detection accuracy is relatively gradual.

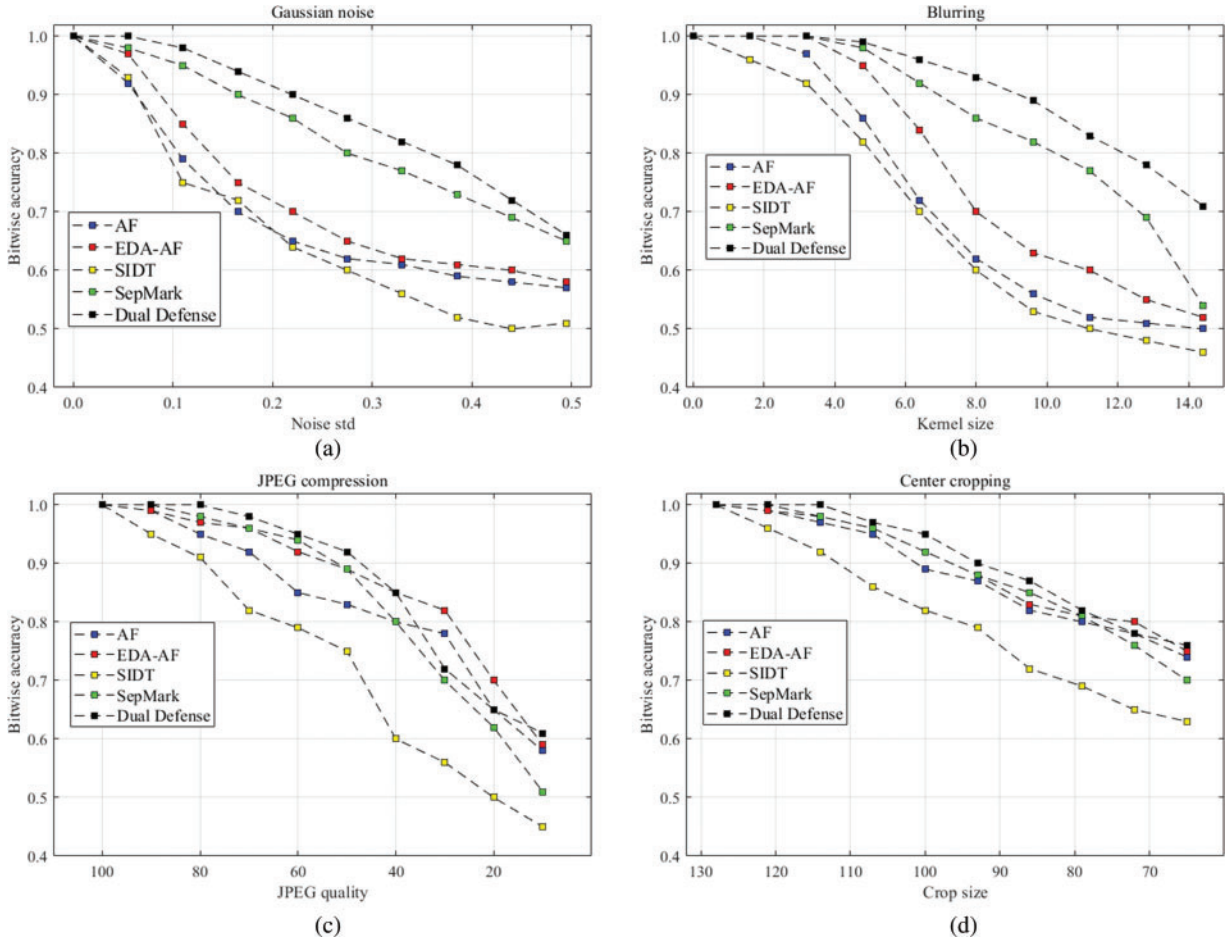


**Figure 16:** Robustness evaluation with Gaussian noises, Blurring, JPEG compression and Central cropping

## 5 Challenges

(1) Assessment criteria for proactive forensics. In response to the risks posed by deepfake content, researchers have proposed various proactive forensic methods from different perspectives. However, the absence of a standardized benchmark for fair comparison of these methods has led to potentially misleading results, which presents a significant challenge. In particular, inconsistencies in data processing modules lead to varied data inputs for forensic detection models. Additionally, there are notable differences among experimental setups and a lack of standardization in evaluation strategies and metrics. The absence of publicly available source code for many methods further hinders the reproducibility and comparability of their reported results.

(2) The issue of robustness in proactive forensics. Images are often utilized in deepfake creation, but the watermark information embedded in these images can be lost during online distribution and is particularly vulnerable to removal during the deepfake generation process. Hence, watermarks for proactive evidence collection in deepfakes must not only retain traditional robustness against noise and compression but also possess enhanced resilience against deep learning-based generation techniques.

Although current watermarking methods demonstrate some robustness, their effectiveness can be limited when confronted with advanced generation technologies and novel attack methods.

(3) The practical issues of proactive forensics. Proactive forensic techniques typically depend on sophisticated deep learning models, which require substantial computational resources for both training and inference. For instance, developing an end-to-end deep digital watermark embedding and extraction network involves processing large datasets during the training phase, leading to significant time consumption and resource inefficiency. In scenarios demanding real-time or near-real-time feedback, such as on social media platforms, the detection methods must operate swiftly. However, complex forensic models often struggle to meet these time constraints, resulting in delays that can negatively impact user experience. Moreover, while existing proactive forensic models may be effective in detecting specific types of forgeries, their performance can vary across different types, limiting their applicability in diverse contexts. As deepfake technology continues to advance, new forgery techniques regularly emerge, necessitating frequent updates to existing models to keep pace with these evolving threats.

## 6 Future Prospects

(1) Establish a comprehensive benchmark for proactive forensics. To enhance the calibration of existing proactive forensic techniques for Deepfake and to foster future innovative developments, establishing a comprehensive benchmark is essential for creating a unified platform for proactive Deepfake forensics. Firstly, implementing a standardized data processing module will ensure consistency across input data, thereby minimizing the time required for data processing and evaluation. Secondly, developing a modular training and testing framework will allow for direct comparisons among different proactive forensic algorithms, facilitating a clearer understanding of their relative performance. Lastly, introducing a unified evaluation scheme will improve the transparency and reproducibility of performance assessments, providing a more reliable basis for evaluating and advancing forensic techniques.

(2) Design an adaptive and cross-domain collaborative proactive forensics model. To enhance the robustness of proactive evidence collection for Deepfake, future research should prioritize the development of an image watermarking model utilizing GANs. This model should be capable of autonomously identifying optimal locations and strengths for watermark embedding while preserving image quality. Additionally, it should integrate noise layers to simulate scenarios such as compression and Deepfake generation. These improvements will enhance the extractor's accuracy in handling both the conventional lossy processing and Deepfake scenarios, thereby bolstering the algorithm's robustness.

(3) Optimize the model structure and design a lightweight model. Current researches on proactive forensics for Deepfake detection primarily focuses on the effectiveness of the methods, with less emphasis on the efficiency of forensic models. To enhance the efficiency of these models, research should be directed towards several key fields: designing lightweight models, applying model compression techniques, utilizing multi-scale data processing, and enabling feature reuse. Implementing an end-to-end optimization framework can significantly boost model performance, making it better suited for real-time applications. Specifically, optimizing digital watermark embedding and extraction techniques will result in more efficient algorithms for watermark insertion and decoding, ensuring that these processes do not substantially increase computational complexity. This approach will help balance the effectiveness of proactive forensic methods with their operational efficiency, facilitating their practical application in dynamic environments.

(4) Ethical Considerations. The application of digital watermarking for deepfake detection raises ethical questions, particularly concerning privacy and the potential for misuse of such technology. It is crucial to ensure that watermarking techniques are designed and implemented in a way that respects user privacy and does not infringe upon civil liberties. This section discusses the measures taken to address these concerns and the ethical framework guiding the development and deployment of our digital watermarking methods. Moving forward, it is imperative that the development of digital watermarking techniques for deepfake detection continues to prioritize ethical considerations. Future work will focus on enhancing privacy measures and developing international standards for the ethical use of such technology.

(5) In addition to advancing digital watermarking techniques, the future of deepfake detection lies in the development of hybrid approaches that leverage the strengths of both passive detection methods and proactive watermarking. Such hybrid methods could provide a more robust defense mechanism by combining the real-time detection capabilities of passive methods with the traceability and authenticity verification offered by watermarking. One promising direction is the integration of passive detection algorithms that analyze inconsistencies in facial images or videos with watermarking techniques that embed unique identifiers into the media. This dual approach could enhance the accuracy of deepfake detection while also providing a means to trace the origin of the manipulated content.

## 7 Conclusions

This survey provides a thorough review of proactive forensic techniques for Deepfake, with a particular emphasis on digital watermarking. With the advancement of Deepfake technology, significant societal challenges have emerged, and existing research predominantly addresses passive detection of counterfeit content. However, these approaches often encounter issues with generalizability and robustness in practical applications. In contrast, digital watermarking presents a promising alternative. The paper systematically categorizes proactive forensic techniques into three primary types: robust watermarking, semi-fragile watermarking, and dual watermarking. It provides a thorough discussion of the technical principles, implementation methods, and the strengths and limitations of various algorithms, complemented by experimental results demonstrating model performance. This paper examines the challenges of proactive forensics in dealing with Deepfake and suggests future research directions. It highlights the need for improvements in generalization, robustness, and the development of multimodal detection technologies to effectively address the dynamic challenges posed by evolving Deepfake technology. Based on an indepth analysis and evaluation of existing technologies, we propose several new directions for future research. Firstly, the development of cross-domain watermarking techniques will allow us to detect deepfakes more effectively across different media types and application scenarios. Secondly, adaptive watermarking strategies can dynamically adjust the strength of watermarks based on the importance and sensitivity of the content, achieving better performance and security. These research directions not only address the shortcomings of existing technologies but also have the potential to advance the field of deepfake detection.

**Acknowledgement:** The authors would like to thank the editors and reviewers for their detailed review and insightful advice.

**Funding Statement:** This work was financially supported by the National Fund Cultivation Project from China People's Police University (Grant Number: JJPY202402), and National Natural Science Foundation of China (Grant Number: 62172165).

**Author Contributions:** The authors confirm contribution to the paper as follows: study conception and design: Zhimao Lai; data collection: Saad Arif; analysis and interpretation of results: Cong Feng; draft manuscript preparation: Zhimao Lai, Guangjun Liao, Chuntao Wang. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data are contained within the article. All databases are publicly available.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

[1] S. Lyu, "Deepfake the menace: Mitigating the negative impacts of AI-generated content," *Organ. Cybersecur. J.: Pract., Process People*, vol. 4, no. 1, pp. 1–18, 2024. doi: 10.1108/OCJ-08-2022-0014.

[2] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *Int. J. Comput. Vis.*, vol. 130, no. 7, pp. 1678–1734, 2022. doi: 10.1007/s11263-022-01606-8.

[3] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, vol. 53, no. 4, pp. 3974–4026, 2023. doi: 10.1007/s10489-022-03766-z.

[4] J. Deng *et al.*, "Towards benchmarking and evaluating deepfake detection," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 6, pp. 5112–5127, 2024. doi: 10.1109/TDSC.2024.3369711.

[5] Q. Abbas *et al.*, "Reducing dataset specificity for deepfakes using ensemble learning," *Comput. Mater. Contin.*, vol. 74, no. 2, pp. 4261–4276, 2023. doi: 10.32604/cmc.2023.034482.

[6] W. Qin, T. Lu, L. Zhang, S. Peng, and D. Wan, "Multi-branch deepfake detection algorithm based on fine-grained features," *Comput. Mater. Contin.*, vol. 77, no. 1, pp. 467–490, 2023. doi: 10.32604/cmc.2023.042417.

[7] B. Xu, J. Liu, J. Liang, W. Lu, and Y. Zhang, "Deepfake videos detection based on texture features," *Comput. Mater. Contin.*, vol. 68, no. 1, pp. 1375–1388, 2021. doi: 10.32604/cmc.2021.016760.

[8] Y. Ni, W. Zeng, P. Xia, G. S. Yang, and R. Tan, "A deepfake detection algorithm based on fourier transform of biological signal," *Comput. Mater. Contin.*, vol. 79, no. 3, pp. 5295–5312, 2024. doi: 10.32604/cmc.2024.049911.

[9] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu and N. Yu, "F2Trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1039–1051, 2023. doi: 10.1109/TIFS.2022.3233774.

[10] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, "FInfer: Frame inference-based deepfake detection for high-visual-quality videos," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, pp. 951–959, 2022. doi: 10.1609/aaai.v36i1.19978.

[11] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2022, pp. 14950–14962.

[12] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7236–7251, 2023. doi: 10.1109/TCSVT.2023.3278310.

[13] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "MRE-Net: Multi-rate excitation network for deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3663–3676, 2023. doi: 10.1109/TCSVT.2023.3239607.

[14] Y. Yu *et al.*, "MSVT: Multiple spatiotemporal views transformer for deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4462–4471, 2023. doi: 10.1109/TCSVT.2023.3281448.

[15] L. Tan *et al.*, "Deepfake video detection via facial action dependencies estimation," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, AAAI Press, 2023, vol. 37, no. 4, pp. 5276–5284.
[16] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Vancouver, BC, Canada, IEEE, 2023, pp. 10491–10503.
[17] Q. Yin, W. Lu, B. Li, and J. Huang, "Dynamic difference learning with spatio-temporal correlation for deepfake video detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4046–4058, 2023. doi: 10.1109/TIFS.2023.3290752.
[18] G. -L. Chen and C. -C. Hsu, "Jointly defending deepfake manipulation and adversarial attack using decoy mechanism," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9922–9931, 2023. doi: 10.1109/TPAMI.2023.3253390.
[19] K. Yao, J. Wang, B. Diao, and C. Li, "Towards understanding the generalization of deepfake detectors from a game-theoretical view," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2031–2041.
[20] J. Dong, Y. Wang, J. Lai, and X. Xie, "Restricted black-box adversarial attack against deepfake face swapping," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 2596–2608, 2023. doi: 10.1109/TIFS.2023.3266702.
[21] G. Pei *et al.*, "Deepfake generation and detection: A benchmark and survey," 2024, *arXiv:2403.17881*.
[22] A. Qureshi, D. Megías, and M. Kuribayashi, "Detecting deepfake videos using digital watermarking," in *2021 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, IEEE, 2021, pp. 1786–1793.
[23] A. Alattar, R. Sharma, and J. Scriven, "A system for mitigating the problem of deepfake news videos using watermarking," *Electron. Imaging*, vol. 32, pp. 1–10, 2020. doi: 10.2352/ISSN.2470-1173.2020.4.MWSF-117.
[24] L. -Y. Hsu, "AI-assisted deepfake detection using adaptive blind image watermarking," *J. Vis. Commun. Image Represent.*, vol. 100, 2024, Art. no. 104094. doi: 10.1016/j.jvcir.2024.104094.
[25] D. P. Kingma and M. Welling, "Auto-Encoding VariationalBayes," in *Conf. Proc.: Papers Accepted Int. Conf. Learn. Representations (ICLR)*, 2014.
[26] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst., Ser. NIPS'17*, Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 6309–6318.
[27] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020. doi: 10.1145/3422622.
[28] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and improving the image quality of stylegan," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, 2020, pp. 8107–8116.
[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, Jun. 2022, pp. 10674–10685.
[30] E. M. Aaron Mir and E. Alonso, "DiT-Head: High resolution talking head synthesis using diffusion transformers," in *ICAART*, 2024, pp. 24–26.
[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.
[32] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimed., Ser. MM '20*, New York, NY, USA, Association for Computing Machinery, 2020, pp. 2003–2011.
[33] K. Shiohara, X. Yang, and T. Taketomi, "BlendFace: Re-designing identity encoders for face-swapping," in *2023 IEEE/CVF Int. Conf. Comput. Vis.(ICCV)*, Los Alamitos, CA, USA, IEEE Computer Society, Oct. 2023, pp. 7600–7610.
[34] 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 *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, 2023, pp. 8568–8577.

[35] W. Huang, S. Tu, and L. Xu, "IA-FaceS: A bidirectional method for semantic face editing," *Neural Netw.*, vol. 158, pp. 272–292, Jan. 2023. doi: 10.1016/j.neunet.2022.11.016.
[36] Y. Pang *et al.*, "DPE: Disentanglement of pose and expression for general video portrait editing," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, 2023, pp. 427–436.
[37] Y. Xu *et al.*, "TransEditor: Transformer-based dual-space gan for highly controllable facial editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7683–7692.
[38] J. Sun *et al.*, "FENeRF: Face editing in neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7672–7682.
[39] G. Hsu, C. Tsai, and H. Wu, "Dual-generator face reenactment," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, 2022, pp. 632–640. Accessed: Nov. 25, 2024. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00072.
[40] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 183–1, 2015. doi: 10.1145/2816795.2818056.
[41] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of rgb videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.
[42] A. Radford, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[43] J. Adler and S. Lunz, "Wasserstein gan," in *Adv. Neural Inf. Process. Syst.*, 2018, vol. 31
[44] K. Shin, *et al.*, "An image Turing test on realistic gastroscopy images generated by using the progressive growing of generative adversarial networks," *J. Digital Imaging*, vol. 36, no. 4, pp. 1760–1769, 2023. doi: 10.1007/s10278-023-00803-2.
[45] M. Liu *et al.*, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3673–3682.
[46] N. F. Soliman, F. E. Fadl-Allah, W. El-Shafai, M. I. Aly, M. Alabdulhafith and F. E. A. El-Samie, "A hybrid cybersecurity algorithm for digital image transmission over advanced communication channel models," *Comput. Mater. Contin.*, vol. 79, no. 1, pp. 201–241, 2024. doi: 10.32604/cmc.2024.046757.
[47] R. Taj *et al.*, "Reversible watermarking method with low distortion for the secure transmission of medical images," *Comput. Model. Eng. Sci.*, vol. 130, no. 3, pp. 1309–1324, 2022. doi: 10.32604/cmes.2022.017650.
[48] S. Xiang, H. J. Kim, and J. Huang, "Invariant image watermarking based on statistical features in the low-frequency domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 6, pp. 777–790, 2008. doi: 10.1109/TCSVT.2008.918843.
[49] J. Lin and Q. Chen, "An intelligent sensor data preprocessing method for oct fundus image watermarking using an rcnn," *Comput. Model. Eng. Sci.*, vol. 138, no. 2, pp. 1549–1561, 2024. doi: 10.32604/cmes.2023.029631.
[50] F. Y. Shih and S. Y. Wu, "Combinational image watermarking in the spatial and frequency domains," *Pattern Recognit.*, vol. 36, no. 4, pp. 969–975, 2003. doi: 10.1016/S0031-3203(02)00122-X.
[51] J. Zhu, R. Kaplan, J. Johnson, and F. Li, "HiDDeN: Hiding data with deep networks," 2018, *arXiv:1807.09937*.
[52] 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*, New York, NY, USA, Association for Computing Machinery, 2021, pp. 41–49.
[53] F. Di Martino and S. Sessa, "Fragile watermarking tamper detection via bilinear fuzzy relation equations," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 2041–2061, 2019. doi: 10.1007/s12652-018-0806-3.
[54] Y. Zhao, B. Liu, T. Zhu, M. Ding, X. Yu and W. Zhou, "Proactive image manipulation detection via deep semi-fragile watermark," *Neurocomputing*, vol. 585, 2024, Art. no. 127593. doi: 10.1016/j.neucom.2024.127593.

[55] P. Lefèvre, P. Carré, C. Fontaine, P. Gaborit, and J. Huang, "Efficient image tampering localization using semi-fragile watermarking and error control codes," *Signal Process.*, vol. 190, 2022, Art. no. 108342.
[56] M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, 2020, pp. 2114–2123.
[57] P. Fernandez, G. Couairon, H. Jegou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *2023 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Los Alamitos, CA, USA, 2023, pp. 22409–22420.
[58] Z. Jiang, J. Zhang, and N. Z. Gong, "Evading watermark based detection of AI-generated content," in *Proc. 2023 ACM SIGSAC Conf. Comput. Commun. Secur.*, New York, NY, USA, 2023, pp. 1168–1181.
[59] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Los Alamitos, CA, USA, 2021, pp. 14428–14437.
[60] N. Yu, V. Skripniuk, D. Chen, L. Davis, and M. Fritz, "Responsible disclosure of generative models using scalable fingerprinting," in *Int. Conf. Learn. Representat.*, 2021.
[61] C. Kim, Y. Ren, and Y. Yang, "Decentralized attribution of generative models," 2021, *arXiv:2010.13974*.
[62] C. -Y. Liao, C. -H. Huang, J. -C. Chen, and J. -L. Wu, "Enhancing the robustness of deep learning based fingerprinting to improve deepfake attribution," in *Proc. 4th ACM Int. Conf. Multimed. Asia*, New York, NY, USA, 2022.
[63] 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. Multimed.*, New York, NY, USA, 2021, pp. 3546–3555.
[64] P. Sun, Y. Li, H. Qi, and S. Lyu, "Faketracer: Exposing deepfakes with training data contamination," in *2022 IEEE Int. Conf. Image Process. (ICIP)*, 2022, pp. 1161–1165.
[65] Y. Zhao, B. Liu, M. Ding, B. Liu, T. Zhu and X. Yu, "Proactive deepfake defence via identity watermarking," in *2023 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 4591–4600.
[66] Y. Lin, H. Chen, E. Maiorana, P. Campisi, and B. Li, "Source-ID-Tracker: Source face identity protection in face swapping," in *2022 IEEE Int. Conf. Multimed. Expo (ICME)*, Los Alamitos, CA, USA, 2022, pp. 1–6.
[67] X. Shen, H. Yao, S. Tan, and C. Qin, "Hiding face into background: A proactive countermeasure against malicious face swapping," *IEEE Trans. Ind. Inform.*, vol. 20, no. 8, pp. 10613–10623, 2024. doi: 10.1109/TII.2024.3396268.
[68] T. Wang, M. Huang, H. Cheng, B. Ma, and Y. Wang, "Robust identity perceptual watermark against deepfake face swapping," 2024, *arXiv:2311.01357*.
[69] 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. doi: 10.1109/TIFS.2024.3383648.
[70] X. Wu, X. Liao, B. Ou, Y. Liu, and Z. Qin, "Are watermarks bugs for deepfake detectors? rethinking proactive forensics," in *Proc. Thirty-Third Int. Joint Conf. Artif. Intell., IJCAI-24*, Jeju, Republic of Korea, 2024, pp. 6089–6097.
[71] Y. Yang, C. Liang, H. Hè, X. Cao, and N. Z. Gong, "FaceGuard: Proactive deepfake detection," 2021, *arXiv:2109.05673*.
[72] P. Neekhara, S. Hussain, X. Zhang, K. Huang, J. McAuley and F. Koushanfar, "FaceSigns: Semi-fragile watermarks for media authentication," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 11, 2024, Art. no. 337. doi: 10.1145/3640466.
[73] N. Beuve, W. Hamidouche, and O. Déforges, "WaterLo: Protect images from deepfakes using localized semi-fragile watermark," in *Proc. IEEE/CVF Int.l Conf. Comput. Vis. (ICCV) Workshops*, Paris, France, Oct. 2023, pp. 393–402.
[74] H. Liu *et al.*, "BiFPro: A bidirectional facial-data protection framework against deepfake," in *Proc. 31st ACM Int. Conf. Multimed.*, New York, NY, USA, 2023, pp. 7075–7084.
[75] X. Wu, X. Liao, and B. Ou, "SepMark: Deep separable watermarking for unified source tracing and deepfake detection," in *Proc. 31st ACM Int. Conf. Multimed.*, New York, NY, USA, 2023, pp. 1190–1201.

[76] X. Zhang, R. Li, J. Yu, Y. Xu, W. Li and J. Zhang, "EditGuard: Versatile image watermarking for tamper localization and copyright protection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2024, pp. 11964–11974.

[77] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. doi: 10.1109/TIP.2003.819861.

[78] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, 2018, pp. 586–595.

[79] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2017, pp. 6629–6640.

[80] G. B. Huang, M. A. Mattar, T. L. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on Faces in "Real-Life" Images: Detection, Alignment, and Recognition*, 2008.

[81] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[82] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Los Alamitos, CA, USA, 2015, pp. 3730–3738.

[83] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Int. Conf. Learn. Representati.*, 2018.

[84] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *2018 13th IEEE Int. Conf. Automatic Face Gesture Recognit. (FG 2018)*, Los Alamitos, CA, USA, 2018, pp. 67–74.

[85] C. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, 2020, pp. 5548–5557.

[86] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Los Alamitos, CA, USA, Nov. 2019, pp. 1–11.

[87] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jun. 2020, pp. 3204–3213.