Dual Defense: Adversarial, Traceable, and Invisible Robust Watermarking against Face Swapping

Yunming Zhang, Dengpan Ye, Caiyun Xie, Long Tang, Xin Liao, Ziyi Liu, Chuanxi Chen, Jiacheng Deng

***Abstract*—Malicious applications of deep face swapping technology pose security threats such as misinformation dissemination and identity fraud. Some research propose the utilization of robust watermarking methods to track the copyright of facial images, facilitating post-forgery identity attribution. However, these methods cannot fundamentally prevent or eliminate the adverse impacts of face swapping. To address this issue, we present Dual Defense, an innovative framework based on robust adversarial watermarking. It simultaneously tracks image copyrights and disrupts the face swapping model by one-time embedding the robust adversarial watermark. Specifically, we propose an Original-domain Feature Emulation Attack (OFEA) method, which makes the traceable watermark adversarial through specially designed original domain adversarial loss. Additionally, we conduct a wavelet domain image structural information compensation loss, combined with a channel attention mechanism, to jointly balance watermark invisibility, adversariality, and traceability. Furthermore, we design a more comprehensive and rational evaluation method to thoroughly assess the effectiveness of adversarial attacks against face swapping models. Extensive experiments demonstrate that Dual Defense exhibits exceptional cross-task generality and dataset generalization. It maintains impressive adversariality and traceability in both original and robust settings, surpassing current forgery defense methods that possess only one of these capabilities.**

***Index Terms*—Watermark, adversarial attack, face swap, active defense.**

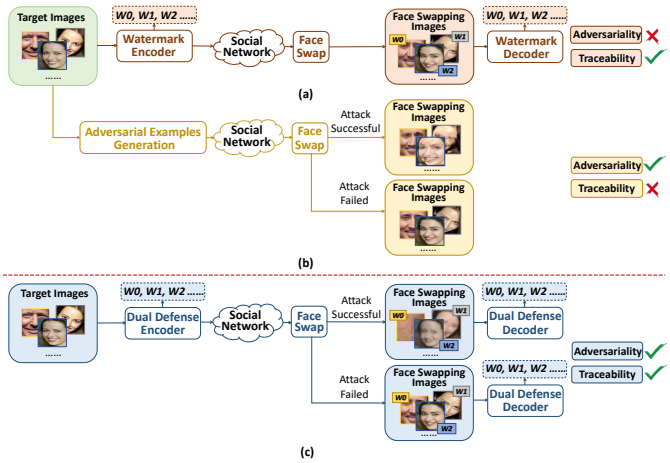


Fig. 1. Illustration of deep forgery active defense scenarios. (a) Active defense based on watermarking. Enables tracing the source of forged images but cannot prevent forgery and eliminate its adverse effects at the source. (b) Active defense based on adversarial examples. It can disrupt forgery generation but does not support traceability, offering no traceability basis upon attack failure. (c) Our Dual Defense active defense. While tracking the copyright of facial images, it can disrupt the FaceSwap model while ensuring watermark integrity. Additionally, it provides auxiliary traceability in case of attack failure.

I. Introduction

THE groundbreaking research in deep learning has driven the rapid advancement of deep forgery, enriching people's lives significantly but simultaneously presenting a substantial threat to multimedia information security [1]–[3]. Especially for face swapping techniques in facial images, which seamlessly replace the target face with a source face for identity forgery. Malicious forgers can exploit this technology to create high quality images and videos of political figures or celebrities, engaging in illegal political and commercial activities, identity fraud, and other unlawful behaviors [4], [5]. Therefore, there is an urgent need to implement effective measures to counter the potential threats introduced by face swapping technology.

Existing defenses against deep forgery mainly include passive detection and active defense. Passive detection involves training detectors to retrospectively validate forged images [4], [6]–[8], while active defense methods primarily include deep watermarking-based defense [9], [10] and adversarial examples-based defense [11]–[13], as shown in Fig. 1 (a). Deep watermarking algorithm provides active defense by pre-encoding a robust watermark representing identity information into the carrier image. After the image has been forged, users can extract the watermark for tracing its source. However, the robust watermark cannot disrupt the forgery model, thus failing to fundamentally eliminate the impact of malicious forgery. The typical active defense involves generating adversarial perturbations based on gradient iteration and directly overlaying them onto the original pixels, as shown in Fig. 1 (b). However, its adversariality is significantly reduced after undergoing image post-processing. The adversarial perturbation has a certain vulnerability and cannot be recovered after the carrier is forged, making the adversarial example

This work was supported by National Natural Science Foundation of China NSFC (No. 62072343); the Fundamental Research Funds for the Central Universities (No. 2042023kf0228); the National Key Research and Development Program of China (No. 2019QY(Y)0206); the National Natural Science Foundation of China (No. U22A2030); the Hunan Provincial Funds for Distinguished Young Scholars (No. 2024JJ2025). *(Corresponding author: Dengpan Ye.)*

Yunming Zhang, Dengpan Ye, Caiyun Xie, Long Tang, Ziyi Liu, Chuanxi Chen, Jiacheng Deng are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China. Xin Liao is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China.(e-mail: {zhangyunming, yedp, caiyunxie, l_tang, chencx, ziyi_liu, dengjiacheng} @whu.edu.cn, xinliao@hnu.edu.cn)

method non-traceable. In case the adversarial perturbation attack fails on the forgery model, users cannot protect their rights through image traceability. Therefore, there is an urgent need for a comprehensive active defense method in real-world social networks, capable of disrupting forgery models while tracing the forged results, and withstand most image post-processing operations.

As a robust copyright protection tool, robust watermarking can be transmitted alongside the carrier in complex channel environments. Therefore, image tracing can be employed in conjunction with adversarial attacks as a joint defense mechanism to ensure the effectiveness of defense methods in most complex scenario. However, the generation method of adversarial examples hinders their simultaneous end-to-end training with robust watermarking models. Additionally, the susceptibility of universal adversarial perturbations makes them unsuitable for direct overlay with watermarks, leading to significant degradation in both adversariality of the adversarial examples and the accuracy of watermark recovery. The characteristics of adversarial examples and robust watermarking have prompted a new consideration for active defense against face swapping. Is it possible to encode the deep watermark into the carrier image in a manner that deviates from the source face manifold, making it adversarial to the face swapping model while maintaining the robustness of the watermark? Building upon this, we focus on face swapping that are more susceptible to legal disputes and propose a novel dual-effect active defense method combating the classic face swapping model FaceSwap, named Dual Defense, which effectively combines adversariality and traceability.

Dual Defense provides the first validation of the adversariality of invisible robust watermark in the generative model. Our research demonstrates that the reconstruction process of the target facial watermarked image can serve as an imitable feature, effectively circumventing the gradient explosion caused by non-targeted attacks in multi-objective optimization. To achieve this, we propose an Original-domain Feature Emulation Attack (OFEA) method, embedding the watermark into the robust adversarial features of the carrier through a carefully-designed original domain adversarial loss. This process causes deviations in the output of the FaceSwap model from the source facial manifold. To alleviate the decline in image quality caused by adversarial optimization, we propose image quality loss with structure information compensation. This involves converting images to the low-frequency domain and computing structural similarity to compensate for the loss of semantic information within the pixel domain. Simultaneously, we train the watermark decoder to extract watermark information both before and after face swapping, ensuring the accurate extraction of watermarks at any stage of carrier transmission, catering to diverse tracing requirements.

As shown in Fig. 1 (c), compared with the existing active defense methods, Dual Defense can fundamentally protect the security of facial images in social networks. The embedded robust watermark can track the entire transmission process of carrier images within social networks. In the event of malicious face swapping during transmission, the watermarked image can promptly disrupt the output of the FaceSwap model while ensuring the integrity of the watermark. Even in complex network environments with multiple image post-processing stages, the watermark maintains a high level of adversarial robustness. If the attack fails, users can still extract the robust watermark for timely traceability, thereby breaking the transmission chain. Our research potentially opens up a novel research direction for applying deep watermarking in the comprehensive active defense against deep forgery.

Our contributions can be summarized as the following:

- We propose Dual Defense, a novel adversarial watermark network, which is the first dual-effect active defense method against face swapping models, combining both adversariality and traceability. It exhibits exceptional robustness, cross-task universality and dataset generalization ability.
- We innovatively propose the OFEA method, which makes the traceable watermark adversarial by embedding it into the robust adversarial features of the carrier. Simultaneously, we address optimization conflicts in multi-objective learning of watermarking by incorporating a specially designed wavelet domain structural information compensation loss.
- We specifically design a more reasonable and comprehensive evaluation method to fully assess the adversariality of Dual Defense against face swapping. Combining traditional evaluation metrics, we have demonstrated the dual effectiveness of Dual Defense in both source tracing and adversarial attacks across three large datasets.

The remainder of this paper is organized as follows. Section II introduces the related works. Section III presents the problem statement of the proposed method. Section IV provides the specific implementation of the proposed method. The experiments and performance evaluation are described in Section V. Finally, we provide a brief conclusion in Section VI.

II. Related Works

A. Face Swapping Models

As a representative technique in facial image forgery, face swapping involves changes to personal identity information, making them more prone to legal issues related to facial privacy and image copyrights [14]–[17]. Early face swapping employed classical graphics techniques [18], involving modeling, rendering, and appropriate color-corrected post-processing to replace faces between different identities. With the development of deep learning technology, CNNs have propelled various computer vision tasks, with their layered architecture enabling the model to extract high level visual patterns for describing the visual properties of images [19]. FaceSwap [20] is a representative algorithm in the field of deep forgery technology, which uses an autoencoder to achieve face swapping between two individuals. GAN-based face swapping methods, such as FSGAN [21] and FaceShifter [22], improve the ability to represent facial features and generate controllable results, leading to higher quality generations. The AOT algorithm [23] addresses the chromatic attribute differences in the replacement process from the perspective of optimal

transport. Among these methods, the autoencoder structure of FaceSwap significantly improves the operability of face swapping while also lowering the technical threshold, making it the most widely used face swapping algorithm to date. In summary, it is imperative to conduct defensive research on FaceSwap, which emerges as the most menacing deepfake method.

B. Robust Watermarking-based Active Defense

Robust watermarking technology plays a vital role in the field of information hiding [24]–[26]. Recent research has focused on using invisible watermarks for deep forgery active defense. For facial image tracing, Wang et al. [9] utilize the deep watermarking method to encode watermarks into carrier face images, enabling effective tracing of manipulated facial images and protecting users' facial image copyrights. Yu et al. [10] introduce Artificial Fingerprints, validating their transferability from training data to generative models, thereby enabling direct labeling of the origins of generated forged images, preventing misuse of forged models. Wu et al. [27] propose SepMark, applying deep watermarking for active forensics on facial images. SepMark employs two decoders to extract the embedded watermark with different levels of robustness. The robust decoder resists against numerous distortions for image tracking, while the semi-robust decoder is selectively sensitive to malicious distortions, used for the detection of deepfake images. However, the aforementioned active defense methods based on robust watermarking have no adversariality. In cases of sudden malicious forgery on the carrier, the watermarked image is unable to disrupt the forgery model and mitigate its adverse impact. In this paper, we investigate the adversariality of robust watermarks against face swapping models, enabling them to actively disrupt the generation model while tracking carrier copyrights, thus fundamentally preventing malicious events.

C. Adversarial Examples-based Active Defense

The DeepFake disruption is the promising countermeasures for fighting against DeepFakes in a active manner [28], [29]. To achieve cross-model image-agnostic adversarial perturbations, Huang et al. [30] propose CMUA-Watermark, which solves the problem of mutual cancellation of anti-noise between different images and different models. Li et al. [31] propose UnGANable, which attacks GAN-inversion based face manipulation by searching for alternative images around the original images in image space. But the above methods are only effective in attribute editing and face reenactment since these two types of manipulation model share a similar pipeline [12]. The face swapping model modifies the high-dimensional semantic features of the images, which is more challenging to defend than the attribute editing model. Wang et al. [12] introduce Anti-Forgery, an active defense method against deepfakes using robust adversarial perturbations. This method is applicable to various image transformation scenarios. However, it requires multiple iterations for each image, which results in noticeable degradation of perceptual quality. Yang et al. [32] proposed Adversarial Faces, using PGD attacks to poison the training data of FaceSwap, resulting in face swapping images with easily detectable artifacts. However, due to the inability of real network users to access the training process of forgers, it is not suitable for practical applications. Adversarial attacks face vulnerability issues, with the adversariality significantly reduced when images undergo post-processing. In cases where attacks on forgery models fail, users are unable to protect copyrights. Therefore, in this paper, we investigate the robust adversariality of traceable watermarks in complex environments while providing additional identity tracing measures for worst-case attack failures.

TABLE I
COMPARISON OF THE RELEVANT ACTIVE DEFENSE METHODS.

Method	Type	Adversariality		Traceability	
		Original	Robust	Original	Robust
CMUA-Watermark [30]	Perturbation	✓	✘	✘	✘
Distorting Attack [29]	Perturbation	✓	✘	✘	✘
Anti-Forgery [12]	Perturbation	✓	✓	✘	✘
Adversarial Faces [32]	Perturbation	✓	✓	✘	✘
Artificial Fingerprints [10]	Watermark	✘	✘	✓	✓
FakeTagger [9]	Watermark	✘	✘	✓	✓
SpeMark [27]	Watermark	✘	✘	✓	✓
Dual Defense (ours)	Watermark	✓	✓	✓	✓

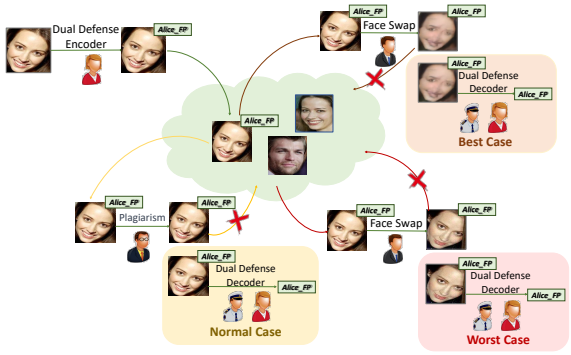


Fig. 2. Explanatory example of Dual Defense for active defense. Defensive measures are presented for three different threat cases. **Normal case**: Standard copyright tracking scenario to prevent plagiarism and theft. **Best case**: The most robust defense attack scenario where we can extract the watermark for tracing even while defending against the face swapping model. **Worst case**: Defense scenario when adversarial attacks fail due to excessive image post-processing, and we can prevent further propagation by extracting the watermark for tracing.

We present a comparative analysis of the proposed method with several existing representative active defense methods in Table I. The data embedded in the carrier can be categorized into two types: irrecoverable perturbations, generated through iterative adversarial training, and user-customizable watermarks, These watermarks contain copyright information and can be retrieved using a decoder for carrier tracing.

III. PROBLEM STATEMENT

In this section, we present the problem statement of our Dual Defense. Given the target user L_t, who possesses a personal image collection $\{x_i\}_{i=1}^{N} \in X_t$ containing N facial images x_i. In the Dual Defense protection mechanism, users can pre-embed watermark into images using the encoder $En(\cdot)$. This watermark is user-defined and can include, but is not limited to, IDs, timestamps, or scene markers. Serving

as a form of generative AI fingerprint, the watermark is rooted within the deep image features to trace image copyrights:

$$X_t^{W_{ID}} = En(X_t, W_{ID}), \tag{1}$$

where $X_t^{W_{ID}}$ represents the obtained watermarked image set $\{x_i^{W_{ID}}\}_{i=1}^{N} \in X_t^{W_{ID}}$. Users can then upload the watermarked images to social networks. Based on the context of a user's facial image transmission on social networks, as shown in Fig. 2, Dual Defense offers active defense strategies from three key cases, encompassing a wide range of potential scenarios:

Normal case: Firstly, as a robust watermarking method, Dual Defense possesses the fundamental capability for copyright tracking. The custom watermark can trace the entire dissemination process of the watermarked image. In cases where there are no sudden instances of malicious face swapping, the watermarked image may still be vulnerable to illegal activities such as plagiarism and unauthorized distribution.

When watermarked images face copyright disputes, users can employ the watermark decoder $De(\cdot)$ to extract the pre-embedded watermark representing copyright ownership for tracing the source:

$$W_{norm} = De(SN(X_t^{W_{ID}})), \tag{2}$$

where $SN(\cdot)$ represents social networks, and W_{norm} represents the extracted watermark information in this scenario.

Best case: During the distribution of watermarked images, copyright is persistently at risk, particularly when unauthorized individuals exploit them for malicious face swapping. Traditional robust watermarking methods can merely track and provide evidence post-incident, falling short of genuinely disrupting such misconduct. However, Dual Defense introduces strong adversariality in these scenarios. It can disrupt the FaceSwap model while preserving the integrity of robust watermarks, thus guaranteeing end-to-end copyright tracking.

When watermarked images are subjected to malicious face swapping, the malicious forger can use FaceSwap model $FS(\cdot)$ to perform face swapping with the source facial images $\{x_m\}_{m=1}^{M} \in X_s$ that belongs to the source user L_s. The adversariality of the watermark can disrupt the output of FaceSwap model, resulting in visually distorted image dataset $X_{adv}^{W_{ID}}$ where identity is unrecognizable as follows:

$$X_{adv}^{W_{ID}} = FS(X_t^{W_{ID}},\ 'L_s'). \tag{3}$$

However, this image may still retain background information related to the original identity, making it susceptible to secondary manipulation for the dissemination of fake news, which poses a significant threat, especially to public figures. In such cases, the original user can trace the source and provide evidence of image ownership by extracting the pre-embedded watermark information representing a timestamp or original scene tag using Dual Defense decoder:

$$W_{best} = De(X_{adv}^{W_{ID}}), \tag{4}$$

where W_{best} represents the extracted watermark information in this best case.

Algorithm 1 Dual Defense Training Framework

Input: X_t (training original target images), W_{ID} (watermarks), $En(\cdot)$ (watermark encoder), $De(\cdot)$ (watermark decoder), $FS(\cdot)$ (FaceSwap model), $Dis(\cdot)$ (discriminator), $maxiter$ (number of iterations), $deiter$ (number of iterations when the decoder starts to update) .

Output: best model parameters.

```
Initialization;
for i ∈ [0, maxiter] do
    X_t^{W_ID} ⇐ En(X_t, W_ID);
    X_t^{(t)} ⇐ FS(X_t^{W_ID}, 'L_t');
    X_adv^{W_ID} ⇐ FS(X_t^{W_ID}, 'L_s');
    0 or 1⇐ Dis(X_t^{W_ID}, X_t);
    Compute disciminator loss L_D with Eq.(8);
    Update disciminator;
    Compute image loss L_img with Eq.(20);
    if i > deiter then
        W_en, W_adv ⇐ De(X_t^{W_ID}, X_adv^{W_ID});
        Compute message loss L_wm with Eq.(21)
        L_total = αL_img + βL_wm;
        Update watermark encoder and decoder;
    else
        L_total = L_img;
        Update watermark encoder;
    end if
end for
return best model parameters.
```

Worst case: Social networks often employ various lossy post-processing operations on uploaded images, including compression, resizing, and filtering, to ensure conformity with standardized transmission or storage protocols [33]. complex lossy operations alter the data distribution, thereby invalidating the images adversariality. While Dual Defense maintains robust adversarial capabilities and can resist most image processing operations, there are still rare extreme cases that can cause attacks to fail, enabling FaceSwap models to successfully generate swapped images. In such case, network administrators or the original users can detect the true ownership of the images by extracting the watermark using the Dual Defense decoder, promptly halting the dissemination of forged images, and preventing the spread of detrimental consequences.

$$X_{s(t)}^{W_{ID}} = FS(X_t^{W_{ID}},\ 'L_s'), \tag{5}$$

$$W_{worst} = De(X_{s(t)}^{W_{ID}}), \tag{6}$$

where $X_{s(t)}^{W_{ID}}$ represents face swapped watermarked images, and W_{worst} represents the extracted watermark.

IV. Methodology

In this section, we present a detailed explanation of the proposed Dual Defense pipeline. The pipeline achieves end-to-end optimization through an adversarial watermark network and comprises four key components: watermark encoder,

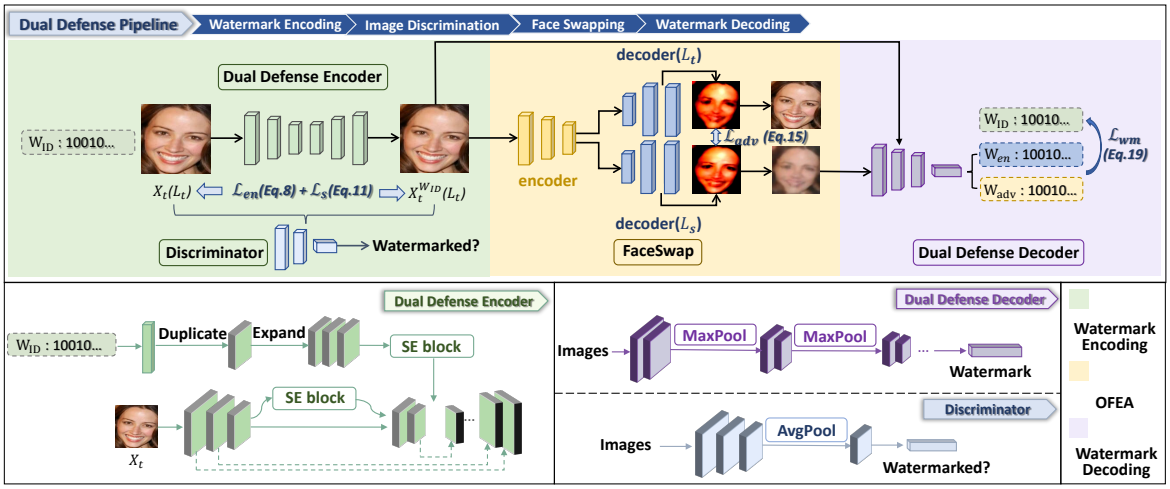


Fig. 3. The whole pipeline of Dual Defense. Dual Defense optimizes watermarking model through end-to-end training. The process begins by inputting target images X_t, along with the user-defined watermark W_{ID}, into the encoder to generate watermarked images. Subsequently, the watermarked images undergo FaceSwap to perform Original-domain Feature Emulation Attack (OFEA) and calculate the original domain adversarial loss. Both disturbed images and watermarked images then pass through the watermark decoder for decoder optimization.

discriminator, FaceSwap model, and watermark decoder, as shown in Fig. 3. The multiple optimization objectives of watermark invisibility, adversariality, and traceability pose a difficult trade-off. One of the main challenges that Dual Defense aims to address is how to encode watermarks into the original target images in a way that is more perceptually aligned with human vision while ensuring adversariality and traceability, balancing the optimization conflicts among different watermark performance aspects.

A. Model Architecture

1) Watermark Encoder: The encoder perceptually encodes watermark information into the original target image. Constructed based on the U-Net architecture, the encoder initially extends carrier channel dimensions through Conv-ReLU blocks composed of convolution ($kernel\,size = 3$, $stride = 1$) and ReLU layers, and downsamples it by Conv-BN-ReLU blocks consisting of convolution, batch normalization, and ReLU layers to extract the carrier features. Additionally, SENet is introduced during downsampling to adaptively recalibrate channel feature responses by explicitly modeling inter-channel dependencies, aiding in determining the optimal watermark embedding strength. Simultaneously, watermark information introduces redundant information to match the dimensions of carrier feature maps, weighted across channels by SENet. Then the downsampled encoded features and the watermark diffusion information are concatenated. The concatenated features layer-wise upsampling through the up-sampling and convolution blocks to get the carrier watermarked feature map. Finally, this feature map is concatenated to the original image and fed into a convolution layer ($kernel\,size = 1$) to yield the three-channel watermarked image.

2) Watermark Decoder: The decoder is employed at the receiving end to extract watermark information from the watermarked image. It comprises five Conv-BN-ReLU blocks with $kernel\,size = 3$ and $stride = 1$, facilitating successive downsampling of the watermarked image. Additionally, we introduce a max-pooling layer and a residual structure after each Conv-BN-ReLU block, enhancing the decoder's capability to learn information mapping within the deep feature space of the carrier. Ultimately, a linear layer is applied to map the extracted features back to the original watermark information length.

3) Discriminator: We further constrain the quality degradation of watermarked images using a discriminator. The discriminator structure primarily consists of Conv-BN-RELU blocks consisting of convolution ($kernel\,size = 3$, $stride = 1$), batch normalization, and ReLU layers. After three layers of convolution, the discriminator reduces feature dimensions through average pooling. Finally the discriminative probability of whether each image is a watermarked image is obtained through the sigmoid layer.

B. Perceptual Adversarial Watermark Encoding

In the optimization process of the watermark encoder, we propose a perceptual adversarial encoding strategy based on OFEA. This strategy embeds the watermark into the robust feature map of the carrier in a manner that deviates from the source facial manifold, and compensates for the image quality degradation caused by the adversariality of the watermark through carefully-designed invisibility loss. The watermark encoder receives the target user's facial image along with a customized watermark representing the user's identity and performs perceptual adversarial encoding. To prevent overfitting during training, we randomly generate watermark information for encoding in each batch:

$$x_i^{W_{ID}} = En(x_i, W_{ID}), \tag{7}$$

The watermark encoder simultaneously optimizes the invisibility and adversarial aspects of the watermark. We integrate a CNN-based discriminator $Dis(\cdot)$ into our framework. The discriminator is trained to distinguish between watermark and

carrier images. Through adversarial optimization with the watermark encoder, the discriminator guides the watermark image distribution to closely resemble real images, thereby constraining the image distortion caused by adversarial optimization. We initially employ Binary Cross Entropy (BCE) $\mathcal{L}_D$ to update the discriminator:

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}_{x_i \sim X_t} - log(Dis(x_i)) \\ &+ \mathbb{E}_{x_i \sim X_t} - log[1 - Dis(x_i^{W_{ID}}))].\end{aligned} \tag{8}$$

Then, by combining the original-domain adversarial loss with the image invisibility loss, we jointly optimize the watermark encoder.

1) Original-domain Feature Emulation Attack (OFEA): FaceSwap achieves face swapping by exchanging the decoders corresponding to different individuals. For target facial images, FaceSwap employs the encoder $FS_E(\cdot)$ to extract the target face features and utilizes the source face decoder $FS_D^s(\cdot)$ for face swapping:

$$X_s^{(t)} = FS_D^s(FS_E(X_t)). \tag{9}$$

Dual Defense uses the original-domain facial features of watermarked images as the optimization target, allowing the face swapping process to learn the attribute features during the watermark image reconstruction, thus attacking the FaceSwap decoder. We first reconstruct the encoded features of the target watermarked image using the corresponding target face decoder $FS_D^t(\cdot)$, extract the output feature map $I_{i(t)}^{W_{ID}}$ of the hidden layer of the decoder as the original-domain attack object. We then use the source face decoder to decode the target feature map, obtaining the face swapped feature map $I_{i(s)}^{W_{ID}}$:

$$I_{i(t)}^{W_{ID}} = FS_D^t(FS_E(x_i^{W_{ID}}))_{k-1}, \tag{10}$$

$$I_{i(s)}^{W_{ID}} = FS_D^s(FS_E(x_i^{W_{ID}}))_{k-1}, \tag{11}$$

where k represents the last layer in the FaceSwap decoder, which is the sigmoid layer. The sigmoid function maps the feature map data to the $[0, 1]$ space, narrowing the search space for gradient optimization, so we perform targeted attack by extracting the input feature map of the sigmoid. The original-domain adversariality loss function is defined as follows:

$$\mathcal{L}_{adv} = \frac{1}{N}\sum_{i=1}^{N}(I_{i(t)}^{W_{ID}} - I_{i(s)}^{W_{ID}})^2. \tag{12}$$

Dual Defense enhances various performances through multi-objective optimization. In the initial iteration, the watermark encoder introduces significant distortion to the watermarked image, leading to severe distortion in the FaceSwap output. In multi-objective optimization, untargeted attacks inevitably lead to gradient explosions. This can be attributed to the fact that untargeted attacks result in different gradient update directions for multiple optimization tasks [34]. Therefore, Dual Defense employs target face reconstruction features for targeted attacks, avoiding gradient explosion while preserving the identity characteristics of the target face simultaneously. Additionally, the attack targets the reconstructed image of the watermarked image instead of the original image. This gradual optimization process allows the encoding learning of the watermark to progress alongside image quality optimization, preventing failures in watermark extraction caused by large disparities in image quality during the initial iterations.

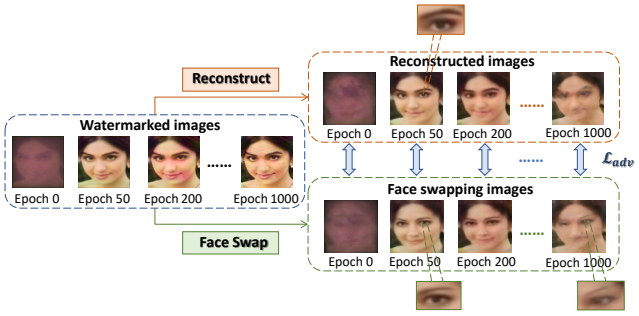


Fig. 4. Optimization process of OFEA. *Reconstruction* refers to using the target face decoder for reconstruction. *Face Swap* indicates using the source face decoder for face swapping.

Fig. 4 shows the watermarked images, their reconstructed images, and the face swapping images at different epochs. It is evident that face swapping images gradually align with reconstructed images of watermarked images in each epoch. At about 50 epochs, a relatively successful face swapping outcome is achieved. However, in subsequent epochs, interference from the target facial features further disrupts the face swapping output, revealing certain features of the original target face, such as eyebrows.

2) Invisibility Loss with Structural Information Compensation: Adversarial encoding method comes at the cost of significantly degrading the perceptual quality of the watermarked image. Low-frequency information in an image encompasses its brightness and structural semantic features, whereas high-frequency information comprises numerous texture details [35]. Traditional image reconstruction objective functions, like the MSE function, have typically prioritized the restoration of global brightness and color while overlooking structural details such as edges and semantic information. We aim to minimize the modification of the carrier image's semantic information and preserve its structural features during the watermark embedding learning process. To achieve this, we introduce image structural information compensation to effectively mitigate the decline in carrier image quality caused by adversarial processing. The invisibility loss of the watermarked image consists of three parts: discriminator loss $\mathcal{L}_D$, image quality loss $\mathcal{L}_{en}$ and structural information compensation loss $\mathcal{L}_s$.

The image quality degradation is controlled at the pixel level through Mean Squared Error (MSE) loss $\mathcal{L}_{en}$. Additionally, adversarial training is conducted using the adversarial loss against discriminator. The formal expressions are as follows:

$$\mathcal{L}_{en} = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i^{W_{ID}})^2 + \mathbb{E}_{x_i \sim X_t} log(1 - Di(x_i^{W_{ID}})). \tag{13}$$

Due to the limitations of the MSE loss, we further enhance structural information compensation in the low-frequency part

of the image. We convert both the original and watermarked images to the YCbCr color space, where the Y channel contains most of the texture information of the images. Therefore, we perform Discrete Wavelet Transform (DWT) on the Y channel of the image and extract its low-frequency sub-band:

$$LL, LH, HL, HH = DWT(Y_{x_i}), \tag{14}$$

$$LL', LH', HL', HH' = DWT(Y_{En(x_i)}), \tag{15}$$

where Y_{x_i} represents the Y channel of the original carrier image, $Y_{En(x_i)}$ represents the Y channel of the watermarked image, LL and LL' respectively represent the low-frequency subbands of the original image and watermarked image, while LH, HL, HH and LH', HL', HH' represent the high-frequency subbands in different directions of the original image and watermarked image. We employ the Structural Similarity (SSIM) loss [36] $\mathcal{L}_s$ to measure the structural information difference between the original carrier and the watermarked image in the low frequency subband, serving as compensation for invisibility loss:

$$\mathcal{L}_s = \mathcal{L}_{ssim} = 1 - SSIM(LL, LL_{en}). \tag{16}$$

$$SSIM(O, W) = \frac{(2\mu_O\mu_W + c_1)(2\sigma_{OW} + c_2)}{(\mu_O^2 + \mu_W^2 + c_1)(\sigma_O^2 + \sigma_W^2 + c_2)}, \tag{17}$$

where O and W represent the images before and after encoding respectively. μ_* and σ_* represent the mean and variance of images. σ_{OW} is the covariance of O and W. $c_1 = k_1L^2$, $c_2 = k_2L^2$ are two variables to stabilize the division with weak denominator. L is the dynamic range of the pixel-values, $k_1 = 0.01$ and $k_2 = 0.03$ by default.

C. Adversarial Watermark Decoding

The watermark decoder is trained end-to-end and co-optimized with the encoder to search for the global optimal solution. The decoder is tasked with decoding the watermark from both the watermarked image and the disrupted image after face swapping, this is done to avoid overfitting to a specific type of attack and to meet the traceability requirements in various real-world scenarios. Dual Defense incorporates the adversarial unit during training, introducing a certain degree of modification to the carrier data distribution. The decoder enhances its robustness against common image noise by learning to recover the watermark from this diverse data distribution.

The traceability loss consists of watermarked image message loss $\mathcal{L}_{wm\_en}$ and disrupted image message loss $\mathcal{L}_{wm\_adv}$. The distance between extracted watermark message and original watermark message W_{ID} is measured by the BCE loss $\mathcal{L}_{BCE}$:

$$\mathcal{L}_{wm\_en} = \mathcal{L}_{BCE}(W_{ID}, W_{en}), \tag{18}$$

$$\mathcal{L}_{wm\_adv} = \mathcal{L}_{BCE}(W_{ID}, W_{adv}), \tag{19}$$

where W_{en} denotes the watermark information extracted from the original watermarked image, while W_{adv} represents the watermark information extracted from the dirupted image.

D. Total Training Loss

Dual Defense simultaneously trains the model end-to-end in the three aspects of watermark invisibility, watermarked image adversariality, and watermark traceability.

The image loss $\mathcal{L}_{img}$ is employed to optimize the watermark encoder and consists of three components: image quality loss, structural information compensation loss, and original domain attack loss.

$$\mathcal{L}_{img} = \lambda_{en}\mathcal{L}_{en} + \lambda_s\mathcal{L}_s + \lambda_{adv}\mathcal{L}_{adv}, \tag{20}$$

where λ_{en}, λ_s and λ_{adv} are positive hyperparameters. The traceability loss is used to optimize the watermark decoder and is composed of the watermark message loss from both before and after the face swapping stages:

$$\mathcal{L}_{wm} = \mathcal{L}_{wm\_en} + \mathcal{L}_{wm\_adv}. \tag{21}$$

The total optimization objective $\mathcal{L}_{total}$ is as follows:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{img} + \beta\mathcal{L}_{wm}, \tag{22}$$

where α and β are positive hyperparameters. We present the experiments and analysis used to determine the hyperparameters in Section V-G. The training process details of Dual Defense can be referred to as Algorithm 1.

V. Experiments

A. Experimental Settings

1) Datasets: We use three large-scale face recognition datasets for model training and verification, VGGFace2 [38] ,CASIA-WebFace [39] and LFW [40]. The image size is 160×160, the watermark message is 30 bits long, representing over 1 billion different AI fingerprints, which is sufficient for practical use. We train the FaceSwap model and the Dual Defense watermarking model on pairs of person image sets. The training set, verification set, and test set of each character are divided according to the ratio of $0.6 : 0.2 : 0.2$.

2) Training Details: Our Dual Defense is implemented by PyTorch and executed on NVIDIA RTX 3090. Due to computational resource constraints, all images are resized to 160×160. The entire training process spans 2500 epochs with a batch size of 16. We empirically adjust the Adam optimizer [41] with an initial learning rate of 0.00005 for stable training. For hyperparameter settings, λ_{en}, λ_s and λ_{adv} are set to 0.8, 0.1, 0.1 respectively, α and β are empirically set to 0.5, 2 respectively. Additionally, since watermarked images undergo watermark extraction through the FaceSwap model, which introduces high-intensity image distortion, there is no need to add an extra noise pool for robust training throughout the process. To ensure that the watermark decoder learns from a carrier that retains most of the image features, we set $deiter$ to 30. In other words, for the first 30 epochs, only the encoder is trained. Starting from the 31st epoch, the decoder is introduced for end-to-end training.

TABLE II
QUANTITATIVE RESULTS OF DUAL DEFENSE IN ORIGINAL SETTINGS WITHOUT IMAGE POST-PROCESSING. * REPRESENTS ACROSS DIFFERENT FACESWAP TASKS WITHIN THE SAME DATASET.

Train	Test	Invisibility		Adversariality						Traceability	
		PSNR↑	SSIM↑	PSNR↓	SSIM↓	L_1↑	LPIPS↑	FN_{acc} ↓	SR_{mask} ↑	Acc_{org}↑	Acc_{adv}↑
VGG-Face2	VGGFace2	31.782	0.917	20.368	0.631	0.078	0.312	0.003	0.772	0.996	0.923
	VGGFace2*	30.121	0.878	23.703	0.833	0.057	0.254	0.197	0.612	0.957	0.868
	CASIA-WebFace	30.021	0.849	22.348	0.791	0.055	0.222	0.027	0.582	0.974	0.843
	LFW	32.318	0.918	24.870	0.878	0.042	0.218	0.233	0.568	0.962	0.854
CASIA-WebFace	CASIA-WebFace	31.830	0.925	22.363	0.764	0.062	0.273	0.009	0.683	0.998	0.986
	CASIA-WebFace*	31.539	0.910	24.102	0.832	0.049	0.178	0.064	0.295	0.992	0.937
	VGGFace2	31.322	0.898	23.949	0.831	0.043	0.162	0.216	0.648	0.962	0.841
	LFW	29.602	0.874	24.675	0.846	0.040	0.165	0.366	0.846	0.982	0.843
LFW	LFW	30.381	0.922	22.558	0.806	0.061	0.274	0.122	0.525	0.987	0.942
	LFW*	29.706	0.948	23.687	0.832	0.045	0.216	0.225	0.661	0.984	0.902
	VGGFace2	28.620	0.920	24.108	0.840	0.048	0.166	0.337	0.442	0.975	0.889
	CASIA-WebFace	28.721	0.943	24.201	0.846	0.043	0.168	0.428	0.582	0.969	0.874

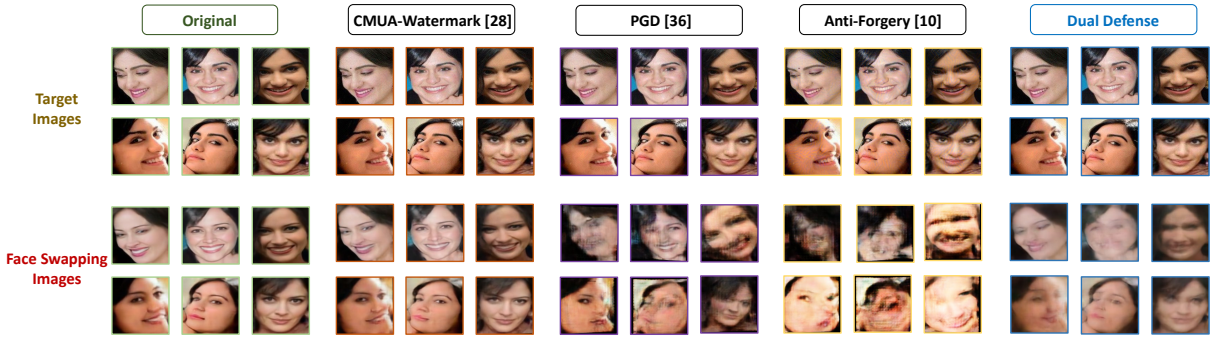


Fig. 5. Visualization results of four methods on the VGGFace2 dataset. *Target Images* in different methods indicate the generated adversarial examples (CMUA-Watermark [30], PGD [37] and Anti-Forgery [12]) or watermarked images (Dual Defense).

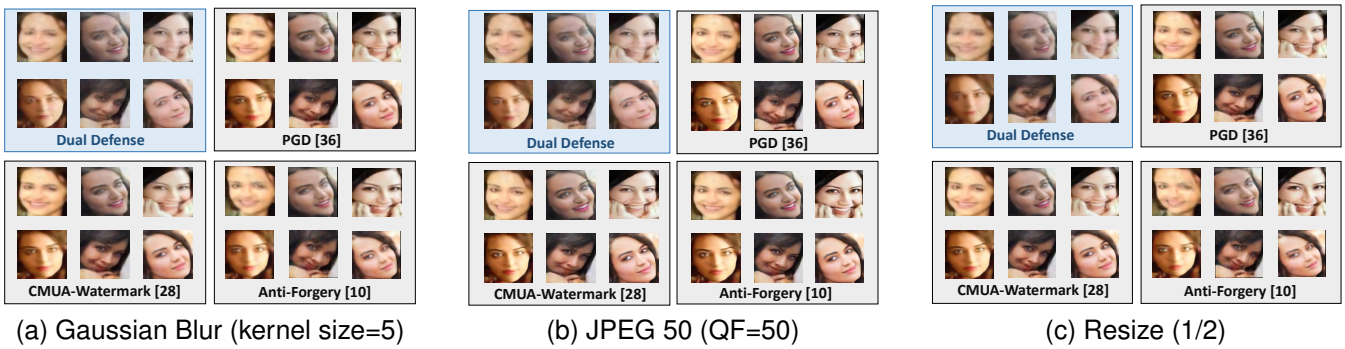


Fig. 6. Visualization results of four methods under robust scenarios.

3) Comparisons: Since Dual Defense is the first active defense method that combines both adversariality and traceability, we compare it with adversarial attack methods and deep watermarking methods separately: *(1) active defense methods based on adversarial attacks:* CMUA-Watermark [30] , Anti-Forgery [12]; *(2) classical adversarial attack methods:* PGD [37]; *(3) active defense method based on deep watermarking:* FakeTagger [9]. We strictly adhere to the experimental setups outlined in the papers of the comparison methods and conduct white-box training on FaceSwap. FakeTagger and Dual Defense follow the same dataset partition. We generate the universal perturbation for CMUA on 64 training images and evaluate it on the same test set. For PGD and Antiforgery, as they do not involve a training process but instead iteratively generate corresponding adversarial examples for individual images, we directly generate their adversarial examples on the same test set. We compared their adversariality and traceability separately. Notably, the aforementioned methods only possess either adversariality or traceability, while Dual Defense possesses both of these capabilities simultaneously.

4) Metrics: We use PSNR, SSIM [36] to evaluate the quality of the watermarked image after encoding the water-

TABLE III
RESULTS OF DUAL DEFENSE IN ROBUST SETTINGS WITH VARIOUS DIFFERENT IMAGE POST-PROCESSING OPERATIONS ON VGGFACE2 DATASET.

Processing	Parameters	$FN_{acc}\downarrow$	SR_{mask}	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$
JPEG	30	0.298	0.315	0.927	0.818
	50	0.162	0.408	0.952	0.846
	70	0.091	0.517	0.965	0.857
	90	0.022	0.686	0.977	0.882
Gaussian Noise	0.001	0.082	0.558	0.964	0.876
	0.002	0.058	0.622	0.932	0.850
	0.003	0.027	0.725	0.907	0.830
	0.004	0.007	0.819	0.880	0.806
Resize	3/4	0.018	0.619	0.981	0.902
	5/4	0.004	0.739	0.983	0.920
	3/2	0.067	0.756	0.984	0.920
	7/4	0.045	0.786	0.983	0.920
Salt&pepper Noise	0.001	0.025	0.744	0.956	0.864
	0.002	0.001	0.700	0.976	0.893
	0.003	0.011	0.678	0.992	0.967
	0.004	0.014	0.692	0.961	0.878

TABLE IV
RESULTS OF DUAL DEFENSE IN ROBUST SETTINGS WITH VARIOUS DIFFERENT IMAGE POST-PROCESSING OPERATIONS ON CASIAWEBFACE DATASET.

Processing	Parameters	$FN_{acc}\downarrow$	SR_{mask}	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$
JPEG	30	0.280	0.477	0.9849	0.948
	50	0.132	0.518	0.993	0.968
	70	0.062	0.536	0.994	0.974
	90	0.048	0.590	0.997	0.981
Gaussian Noise	0.001	0.002	0.647	0.993	0.963
	0.002	0.008	0.659	0.993	0.987
	0.003	0.010	0.660	0.992	0.979
	0.004	0.015	0.676	0.989	0.966
Resize	3/4	0.161	0.462	0.996	0.976
	5/4	0.036	0.523	0.998	0.984
	3/2	0.062	0.574	0.997	0.983
	7/4	0.035	0.528	0.999	0.988
Salt&pepper Noise	0.001	0.008	0.647	0.996	0.980
	0.002	0.012	0.659	0.995	0.976
	0.003	0.015	0.678	0.993	0.969
	0.004	0.017	0.672	0.993	0.962

mark, use the accuracy rate Acc_{org} and Acc_{adv} to evaluate the watermark recovery accuracy of the watermarked image and the disrupted image respectively, which are calculated as follows:

$$Acc_* = 1 - \frac{1}{L}\sum_{l=1}^{L}\left|W_{ID} - W_*\right|, \tag{23}$$

where Acc_* denotes Acc_{org} or Acc_{adv}, W_* denotes W_{en} or W_{adv}, l is used to index the bits of the watermark information, and L denotes the total length of the watermark. L_1, LPIPS [42], PSNR, SSIM and SR_{mask} [30] are used to measure the quality of disrupted images to reflect the adversariality of methods, but they mainly measure the color, texture, and structural differences in the image, they are traditional metrics used for evaluating natural images. Dual Defense mainly aims to disrupt the replacement of target face identity information by FaceSwap, focusing on the protection of facial features. Therefore, we propose to use the face recognition model FaceNet [43] to evaluate the adversariality. If FaceNet cannot recognize face swapping images as the source face identity, the disruption is considered successful. In order to exclude the effect of the FaceSwap on the FaceNet recognition accuracy, we propose the FaceNet recognition accuracy index:

$$FN_{acc} = \frac{FN_{acc}(X_{adv}^{W_{ID}} \rightarrow source)}{FN_{acc}(X_s^{(t)} \rightarrow source)}, \tag{24}$$

where $FN_{acc}(X_{adv}^{W_{ID}} \rightarrow source)$ represents the recognition accuracy of FaceNet to recognize the disrupted image as the source face, $FN_{acc}(X_s^{(t)} \rightarrow source)$ represents the accuracy of FaceNet recognizing the original face swapping image as the source person. A smaller FN_{acc} indicates that the face swapping images deviate more from the source face, indicating that Dual Defense has better adversariality.

B. Results of Dual Defense in Original Settings

In this subsection, we present the results of Dual Defense on the original settings without image post-processing. Table II shows the results of Dual Defense on three datasets, covering white-box, cross-task, and cross-dataset scenarios. In the white-box scenarios, Dual Defense consistently achieves the FN_{acc} below 0.01 on the VGGFace2 and CASIAWebFace datasets, while on the LFW dataset, FN_{acc} is slightly higher but remains below 0.13. This is attributed to the lower resolution and fewer images per identity in the LFW dataset, resulting in limited training data. Regarding traceability, watermark recovery accuracy is consistently maintained above 0.9. Cross-task refers to testing Dual Defense on the FaceSwap model of another pair of unknown characters in the same dataset. Dual Defense maintains excellent performance in both cross-task and cross-dataset scenarios. The results validate the excellent cross-task universality and dataset generalization ability of Dual Defense, demonstrating that a single-trained Dual Defense is not limited to defending against face swapping task for a single identity but can be universally applied across different datasets and identities involved in face swapping tasks.

C. Results of Dual Defense in Robust Settings

In real channels and social networks, images often undergo various post-processing operations. Therefore, we evaluate the adversariality and traceability of Dual Defense against FaceSwap under four common image post-processing operations: JPEG compression, Gaussian noise, resizing and salt&pepper noise. As shown in Table III and Table IV, Dual Defense maintains excellent adversariality and traceability. Although high-intensity JPEG compression weakens the overall performance of Dual Defense, even at a JPEG compression factor of 30, the FN_{acc} remains below 0.3, and the watermark recovery accuracy stays above 0.8. Considering all scenarios, when the watermarked image is subjected to various processing operations, the proposed Dual Defense consistently maintains excellent performance, thereby validating the feasibility of our method in practical scenarios.

D. Comparison with Other Active Defense Methods

We report the experimental results comparing Dual Defense to the other four methods in Table V and visualize their performance in original and robust scenarios in Fig. 5vand Fig. 6.

TABLE V
COMPARISON OF DUAL DEFENSE AND OTHER ACTIVE DEFENSE METHODS. THE BEST RESULTS ARE EMPHASIZED IN BOLD. **N/A** MEANS THAT THE METHOD LACKS THIS FUNCTION AND THE DATA CANNOT BE OBTAINED.

Dataset	Method	Original			JPEG ($QF = 50$)			Gaussian Noise ($\sigma = 0.005$)			Resize ($1/2$)		
		$FN_{acc}\downarrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$	$FN_{acc}\downarrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$	$FN_{acc}\downarrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$	$FN_{acc}\downarrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$
VGG-Face2	PGD [37]	0.009	N/A	N/A	0.396	N/A	N/A	0.024	N/A	N/A	0.648	N/A	N/A
	CMUA-Watermark [30]	0.188	N/A	N/A	0.612	N/A	N/A	0.036	N/A	N/A	0.668	N/A	N/A
	Anti-Forgery [12]	**0.000**	N/A	N/A	0.703	N/A	N/A	0.065	N/A	N/A	0.245	N/A	N/A
	FakeTagger [9]	1.000	0.9871	0.8295	0.964	**0.956**	0.802	0.748	**0.968**	0.803	0.696	0.980	0.810
	Dual Defense (ours)	0.003	**0.996**	**0.923**	**0.014**	0.962	**0.873**	**0.004**	0.967	**0.854**	**0.084**	**0.987**	**0.885**
CASIA-WebFace	PGD [37]	0.075	N/A	N/A	0.461	N/A	N/A	0.107	N/A	N/A	0.584	N/A	N/A
	CMUA-Watermark [30]	0.461	N/A	N/A	0.497	N/A	N/A	0.153	N/A	N/A	0.636	N/A	N/A
	Anti-Forgery [12]	**0.003**	N/A	N/A	0.830	N/A	N/A	0.182	N/A	N/A	0.555	N/A	N/A
	FakeTagger [9]	0.784	0.997	0.962	0.874	0.942	0.7875	0.408	0.958	0.733	0.815	0.580	0.571
	Dual Defense (ours)	0.009	**0.998**	**0.986**	**0.113**	**0.994**	**0.974**	**0.009**	**0.968**	**0.893**	**0.328**	**0.997**	**0.966**
LFW	PGD [37]	0.381	N/A	N/A	0.461	N/A	N/A	0.107	N/A	N/A	0.584	N/A	N/A
	CMUA-Watermark [30]	0.819	N/A	N/A	0.497	N/A	N/A	0.153	N/A	N/A	0.636	N/A	N/A
	Anti-Forgery [12]		N/A	N/A		N/A	N/A		N/A	N/A		N/A	N/A
	FakeTagger [9]	0.966	0.954	0.922	0.958	0.911	0.884	0.926	0.819	0.812	0.962	0.784	0.761
	Dual Defense (ours)	**0.122**	**0.987**	**0.942**	**0.228**	**0.985**	**0.936**	**0.173**	**0.968**	**0.896**	**0.114**	**0.987**	**0.939**

PGD, CMUA-Watermark, and Anti-Forgery are gradient-based adversarial attack methods without traceability, so traceability metrics are unavailable, denoted as N/A. We only compare their adversarial performance and contrast the traceability with FakeTagger. As shown in Table V, under the original setting, Antiforgery demonstrates the strongest adversarial performance on the VGGFace2 and CASIA-WebFace datasets due to its longer iteration time. However, its adversarial performance decreases on the low-resolution LFW dataset, indicating its limitations across multiple data types. In contrast, Dual Defense exhibits significant adversariality across all three datasets. In robust scenarios, image processing operations notably reduce the adversarial performance of the other three methods, whereas Dual Defense maintains excellent adversarial robustness, particularly after JPEG compression, where its FN_{acc} decreases by over 70% compared to Antiforgery. These observations imply that iterative-based adversarial attack methods often lack robustness against FaceSwap attacks, thereby limiting their practical applicability, while demonstrating the effectiveness of Dual Defense under conditions of limited examples and low-resolution data. Moreover, Dual Defense achieves consistently higher watermark recovery accuracy compared to FakeTagger while ensuring adversariality. Notably, on the CASIA dataset, when images are resized, Dual Defense's Acc_{org} and Acc_{adv} are approximately 40% higher than those of FakeTagger.

E. Visual comparison and analysis

In Fig. 5, we present the visual performance comparison between Dual Defense and the other three adversarial example methods on VGGFace2 datasets under the original setting. Notably, we exclude the visualization results of FakeTagger due to its lack of adversariality. As shown in Fig. 5, the adversarial examples generated by CMUA-Watermark, PGD, and Anti-Forgery exhibit clear and uniform noise patterns, especially Anti-Forgery, which contains noticeable noise. This is due to the extended iteration cycles for a single image, resulting in noticeable noise being introduced into the carrier, contributing to its high attack effectiveness. However, this characteristic is easily detectable and can be circumvented by malicious forgers. The watermarked images obtained by Dual Defense also show some quality degradation, but compared to other methods, they exhibit a more natural slight blurring rather than regular noise, which avoids raising suspicion from attackers. For disrupted images of CMUA-Watermark, minor changes in facial feature positions are observed while preserving the overall facial structure. The disrupted images produced by Anti-Forgery are the most severely damaged, but this is due to a significant reduction in the quality of its adversarial examples. In contrast, the watermarked images obtained by Dual Defense not only ensure the invisibility of the watermark but also completely blur the facial features, making them visually unrecognizable, thereby ensuring facial anonymity.

Fig. 6 displays the disrupted images obtained from the four methods under robust scenarios. Except for Dual Defense, the other three methods exhibit a noticeable decrease in adversariality after undergoing image processing operations, while Dual Defense still maintains excellent adversariality, blurs facial features, and conceals identity information. This indicates that Dual Defense possesses not only traceability robustness but also promising adversarial robustness.

F. Comprehensive Defense Performance Evaluation

In correspondence to the complex scenarios that facial images may encounter, as proposed in Section III, we comprehensively assess the overall defense success rate of active defense methods from multiple perspectives. As shown in Table VI, adversarial example methods experience a significant reduction in adversarial performance after undergoing several common image post-processing steps, with the resulting perturbed images still displaying distinct source facial identity features, enabling correct recognition by facial recognition models. The introduced Dual Defense method offers traceability mechanisms in cases of weakened adversarial effectiveness, enhancing adversarial robustness while enabling image tracing. Therefore, in this section, we trace images recognized correctly by the FaceNet model in cases of adversarial attack failure to evaluate the comprehensive defense performance of Dual Defense. As shown in Table VI, across various

TABLE VI
COMPREHENSIVE DEFENSE PERFORMANCE EVALUATION OF DUAL DEFENSE IN ROBUST SCENARIOS. THE TEST IMAGES ARE ALL FACE-SWAPPING IMAGES THAT CAN BE SUCCESSFULLY IDENTIFIED AS THE SOURCE FACE.

Processing	Parameters	VGGFace2		CASIA-WebFace	
		$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$
JPEG	30	0.945	0.932	0.972	0.975
	50	0.952	0.913	0.993	0.983
	70	0.978	0.919	0.994	0.989
	90	0.982	0.932	0.997	0.991
Gaussian Noise	0.001	0.972	0.916	0.995	0.976
	0.002	0.961	0.925	0.993	0.983
	0.003	0.957	0.907	0.992	0.979
	0.004	0.928	0.914	0.995	0.986
Resize	3/4	0.987	0.932	0.998	0.966
	5/4	0.987	0.946	0.998	0.958
	3/2	0.988	0.948	0.997	0.973
	7/4	0.988	0.946	0.999	0.963
Salt&pepper Noise	0.001	0.968	0.914	0.997	0.987
	0.002	0.982	0.935	0.998	0.978
	0.003	0.992	0.967	0.998	0.976
	0.004	0.954	0.918	0.997	0.974

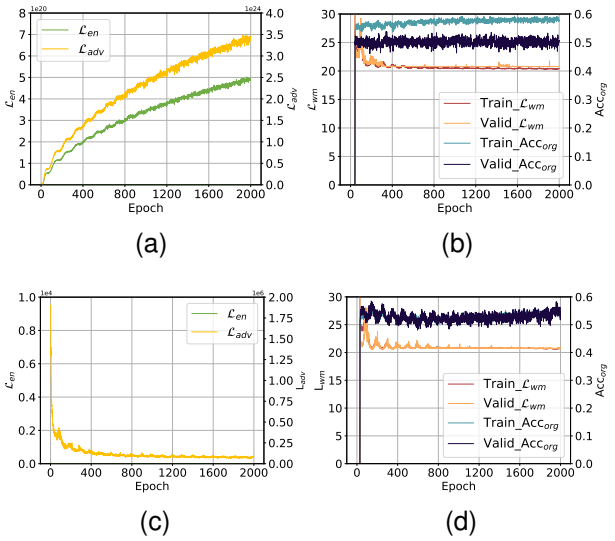


Fig. 7. Image loss during training and traceability loss, as well as watermark recovery accuracy during training and validation. (a-b) Untargeted attack. (c-d) Targeted attack using original carrier Images.

complex scenarios, for images where attacks fail, Dual Defense consistently maintains a traceability accuracy of 0.9 or above. This demonstrates that in situations of poor adversarial effectiveness, Dual Defense can effectively provide auxiliary means to assist network administrators or the original target users in tracing the source.

G. Ablation Study

In this section, we first investigate the significance of the proposed original-domain facial feature attack method. Then, we discuss the effectiveness of the various contributions proposed.

We introduce a perceptual adversarial watermark encoding strategy based on OFEA, using the watermarked image of the target face as the imitation target for targeted attacks. We analyze the drawbacks of two other typical attack methods and

TABLE VII
ABLATION STUDY ON ADVERSARIAL OPTIMIZATION STRATEGY. *SENet* REPRESENTS THE CHANNEL ATTENTION MECHANISM, *SIC* REPRESENTS THE STRUCTURAL INFORMATION COMPENSATION LOSS. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

Dataset	Method	Invisibility		Adversariality		Traceability	
		PSNR ↑	SSIM↑	$FN_{acc}\downarrow$	$SR_{mask}\uparrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$
VGG-Face2	w/o SIC	27.981	0.855	**0.000**	0.742	0.839	0.802
	w/o SENet	31.361	0.914	0.012	0.714	0.978	0.912
	w/o SIC+SENet	27.412	0.809	0.196	0.523	0.838	0.812
	w/o Adv	**42.112**	**0.996**	0.990	0.132	0.989	**0.988**
	Dual Defense	31.782	0.917	0.003	**0.772**	**0.996**	0.923
CASIA-WebFace	w/o SIC	30.213	0.859	**0.006**	0.457	0.978	0.968
	w/o SENet	31.824	0.936	0.0136	0.537	0.985	0.976
	w/o SIC+SENet	27.983	0.789	0.363	0.359	0.962	0.911
	w/o Adv	**41.036**	**0.995**	0.997	0.218	0.993	0.986
	Dual Defense	31.830	0.925	0.009	**0.683**	**0.998**	**0.986**

TABLE VIII
QUANTITATIVE RESULTS OF DUAL DEFENSE UNDER DIFFERENT WATERMARK LENGTHS. THE TRAINING AND TEST DATASET IS CASIA-WEBFACE. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

Length	Invisibility		Adversariality		Traceability	
	PSNR↑	SSIM↑	$FN_{acc}\downarrow$	$SR_{mask}\uparrow$	$Acc_{org}\uparrow$	$Acc_{adv}\uparrow$
15 bits	31.217	0.898	0.011	**0.671**	**0.999**	0.974
30 bits	**31.830**	**0.925**	**0.009**	0.683	0.998	**0.986**
45 bits	30.736	0.898	0.019	0.386	0.998	**0.985**

explain the rationale behind choosing the in-domain feature emulation attack approach.

Untargeted attack can be achieved by maximizing the distance between the faceswapped image of the watermarked image and the faceswapped image of the original carrier image (Eq. 25). As depicted in Fig. 7 (a-b), untargeted attack result in a loss of watermark image quality and an increase in adversarial loss. This is because the watermarked image obtained during the initial iterations loses almost all semantic and color features. Further disrupting with the faceswapped image prevents the optimization of watermarked image quality, leading to gradient explosion and the inability to extract the watermark message.

$$\mathcal{L} = 1 - \frac{1}{N}\sum_{i=1}^{N}(FS(X_t^{W_{ID}},\ 'L_s') - FS(X_t,' L_s'))^2. \quad (25)$$

Targeted attack based on the original carrier image can be achieved by minimizing the reconstruction distance between the face swapped image of the watermark image and the reconstructed image of the original carrier (Eq. 26). As shown in Fig. 7 (c-d), while the quality of the watermark image can continuously improve, there remains a significant difference in quality between the reconstructed watermark image and the original image reconstruction. This is still attributed to the quality degradation of the watermark image in the initial iterative steps, preventing progressive emulation attacks and hindering watermark extraction.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(FS(X_t^{W_{ID}},\ 'L_s') - FS(X_t,' L_t'))^2. \quad (26)$$

The original-domain facial feature attack of the watermarked image ensures a consistent gradient direction during

TABLE IX
QUANTITATIVE RESULTS OF DUAL DEFENSE UNDER DIFFERENT WEIGHTS. THE RESULTS OF α: β WERE OBTAINED WITH λ_{en}: λ_G: λ_{adv} SET TO 0.8: 0.1: 0.1, AND THE RESULTS OF λ_{en}: λ_G: λ_{adv} WERE OBTAINED WITH α: β SET TO 0.5: 2. THE TRAINING AND TEST DATASET IS CASIA-WEBFACE. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

Weight	Ratio	Invisibility		Adversariality						Traceability	
		PSNR↑	SSIM↑	PSNR↓	SSIM↓	L_1↑	LPIPS↑	FN_{acc} ↓	SR_{mask} ↑	Acc_{org}↑	Acc_{adv}↑
α: β (0.8: 0.1: 0.1)	0.5: 1	32.404	0.911	23.341	0.819	0.055	0.212	0.011	0.411	0.993	0.974
	0.5: 2	31.830	0.925	**22.363**	**0.764**	**0.062**	**0.273**	0.009	**0.683**	**0.998**	0.986
	0.5: 3	31.925	0.908	23.193	0.813	0.055	0.219	0.012	0.492	0.998	**0.992**
λ_{en}: λ_G: λ_{adv} (0.5: 2)	0.9: 0.1: 0.1	32.295	0.903	23.304	0.819	0.054	0.216	0.010	0.470	0.996	0.985
	0.7: 0.1: 0.1	31.354	0.891	22.782	0.806	0.058	0.232	**0.002**	0.455	0.998	0.985
	0.8: 0.01: 0.1	31.841	0.895	22.972	0.811	0.057	0.222	0.006	0.463	0.997	0.986
	0.8: 0.2: 0.1	31.714	0.894	22.749	0.803	0.058	0.230	0.004	0.494	0.998	0.987
	0.8: 0.1: 0.01	**36.195**	**0.956**	24.585	0.844	0.037	0.099	0.755	0.424	0.997	0.952
	0.8: 0.1: 0.2	29.160	0.849	23.378	0.818	0.050	0.187	0.175	0.408	0.995	0.922

the multi-task learning process. It reduces the quality difference between the watermarked image and the image to be imitated from the early iterations of the attack. This guarantees the watermarked image can effectively disrupt FaceSwap while optimizing the watermark decoder efficiently.

In addition, the watermark encoding network of Dual Defense incorporates SENet to guide watermark embedding and utilizes structural information compensation to enhance the quality of watermarked image reconstruction. We conducted ablation experiments to observe their impact on image invisibility, adversariality, and traceability. As shown in Table VII, SENet enhances the adversariality and traceability of Dual Defense but reduces the watermark invisibility. Image structure information compensation (represented by SIC) effectively enhances image quality but does not contribute to other aspects of performance. Dual Defense combines these two mechanisms to achieve effective trade-off in watermark performance. Additionally, Table VII indicates that the adversarial optimization method provides fundamental adversariality for Dual Defense.

We investigate the impact of watermark length on the performance of Dual Defense. As shown in Table VIII, Dual Defense maintains outstanding performance across various watermark requirements. Since the watermark length of 30 bits is sufficient to meet the needs of real social networks, and has better adversariality, we choose 30 bits watermark for model training in our scheme.

We investigate the impact of different weight configurations in the loss function on the performance of our method. We utilize repeated cross-tests to identify the optimal weight configuration that achieves a balanced trade-off between various performance aspects. The experimental results are presented in Table IX. When setting β to 1, we observe that the invisibility of the watermark reached its highest level. However, this came at the cost of compromised adversariality and traceability. On the other hand, as we increased the value of β, the accuracy of watermark recovery improved, but the adversariality also decreased. To strike a balance between adversariality and traceability, we selected a ratio of 0.5 : 2 for α to β. Subsequently, we determined the optimal internal image loss weights individually. Notably, when λ_{en}: λ_G: λ_{adv} was set to 0.8 : 0.1 : 0.1, the overall performance of watermarking reached its optimum.

VI. DISCUSSION

Dual Defense proactively defends against FaceSwap models and demonstrates competitive results in terms of traceability, adversariality, and robustness in complex environments. As the first attempt at dual-purpose defense against the face swapping model, Dual Defense also exhibits some limitations. It disrupts the feature reconstruction process of face swapping models, currently showing significant adversariality only on autoencoder-based face swapping model. However, face swapping technology is rapidly evolving, with many high-quality emerging face swapping models being widely applicated, such as GAN-based and diffusion-based face swapping models. This necessitates further enhancement of adversariality on various victim models in future work and the enhancement of black-box model adversarial transferability based on universal features for different types of models. Furthermore, the trade-off between the adversariality, invisibility, and traceability of watermarks is also a crucial issue hindering the improvement of model performance. Enhancing the generalized adversariality of the model necessarily enhances the embedding strength of the watermark. How to further improve the universal performance of the model across various face swapping techniques while maintaining the perceptual quality of watermarked images is the focus of our future work.

VII. CONCLUSION

In this paper, we present the first dual-effect active defense method for face swapping models, named Dual Defense. Dual Defense employs a one-time watermark embedding for copyright tracking during image propagation. It is simultaneously capable of thwarting malicious face swapping operations during dissemination and ensuring the integrity of the imperceptible watermark. Specifically, we propose a novel robust adversarial watermark network that employs the OFEA method for adversarial optimization to make watermarked images adversarial. Simultaneously, we introduce the structural information compensation loss in the wavelet domain to further restrict image quality degradation, effectively balancing the invisibility and adversariality of the watermark, resolving conflicts in multi-objective optimization. Furthermore, we specifically design a more comprehensive evaluation method, incorporating additional evaluation metrics, and conduct a comprehensive assessment of Dual Defense on three large-scale face datasets.

Experimental results demonstrate the excellent adversariality and traceability accuracy of Dual Defense across various complex scenarios, and demonstrate remarkable task and dataset generalization capabilities.

REFERENCES

[1] X. Tu, Z. Ma, J. Zhao, G. Du, M. Xie, and J. Feng, "Learning generalizable and identity-discriminative representations for face anti-spoofing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–19, 2020.

[2] R. Han, X. Wang, N. Bai, Q. Wang, Z. Liu, and J. Xue, "Fcd-net: Learning to detect multiple types of homologous deepfake face images," *IEEE Trans. Inf. Forensics Security.*, vol. 18, pp. 2653–2666, 2023.

[3] X. Zhang, X. Zhang, W. Liu, X. Zou, M. Sun, and J. Zhao, "Waveform level adversarial example generation for joint attacks against both automatic speaker verification and spoofing countermeasures," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105469, 2022.

[4] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," in *Proc. IEEE/CVF Conf. on Comput Vis. Pattern Recognit.(CVPR)*, 2023, pp. 4490–4499.

[5] L. Chen, D. Ye, Y. Shang, and J. Huang, "Robust video hashing based on local fluctuation preserving for tracking deep fake videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. IEEE, 2022, pp. 2894–2898.

[6] Y. Jeong, D. Kim, Y. Ro, and J. Choi, "Frepgan: Robust deepfake detection using frequency-level perturbations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 1060–1068.

[7] J. Wang, B. Tondi, and M. Barni, "Classification of synthetic facial attributes by means of hybrid classification/localization patch-based analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[8] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Trans. Inf. Forensics Security.*, pp. 1–1, 2023.

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

[10] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *Proc. IEEE Int. Conf. Comput. Vis.(ICCV)*, 2021, pp. 14 448–14 457.

[11] Q. Huang, J. Zhang, W. Zhou, W. Zhang, and N. Yu, "Initiative defense against facial manipulation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1619–1627.

[12] R. Wang, Z. Huang, Z. Chen, L. Liu, J. Chen, and L. Wang, "Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations," *arXiv:2206.00477*, 2022.

[13] J. Dong, Y. Wang, J. Lai, and X. Xie, "Restricted black-box adversarial attack against deepfake face swapping," *IEEE Trans. Inf. Forensics Security.*, vol. 18, pp. 2596–2608, 2023.

[14] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proc. IEEE/CVF Conf. on Comput Vis. Pattern Recognit.(CVPR)*, 2018, pp. 6713–6722.

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

[16] R. Natsume, T. Yatagawa, and S. Morishima, "Rsgan: Face swapping and editing using face and hair representation in latent spaces," *arXiv:1804.03447*, 2018.

[17] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.(ICCV)*, 2017, pp. 3677–3685.

[18] M. Kowalski, "Faceswap github," in *github: MarekKowalski/FaceSwap*, 2016.

[19] J. Zhao, J. Li, F. Zhao, X. Nie, Y. Chen, S. Yan, and J. Feng, "Marginalized CNN: learning deep invariant representations," in *British Machine Vision Conference 2017, BMVC*, 2017.

[20] FaceSwap, "Deepfakes github," in *github: Deepfakes/faceswap*, 2017.

[21] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Conf. on Comput Vis. Pattern Recognit.(CVPR)*, 2019, pp. 7184–7193.

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

[23] H. Zhu, C. Fu, Q. Wu, W. Wu, C. Qian, and R. He, "Aot: Appearance optimal transport based identity swapping for forgery detection," *Adv. in Neural Info. Proc. Sys.*, vol. 33, pp. 21 699–21 712, 2020.

[24] Q. Guan, P. Liu, W. Zhang, W. Lu, and X. Zhang, "Double-layered dual-syndrome trellis codes utilizing channel knowledge for robust steganography," *IEEE Trans. Inf. Forensics Security.*, vol. 18, pp. 501–516, 2023.

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

[26] Z. Jia, H. Fang, and W. 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*, 2021, pp. 41–49.

[27] X. Wu, X. Liao, and B. Ou, "Sepmark: Deep separable watermarking for unified source tracing and deepfake detection," in *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, 2023, pp. 1190–1201.

[28] Z. Chen, L. Xie, S. Pang, Y. He, and B. Zhang, "Magdr: Mask-guided detection and reconstruction for defending deepfakes," in *Proc. IEEE/CVF Conf. on Comput Vis. Pattern Recognit.(CVPR)*, 2021, pp. 9014–9023.

[29] C.-Y. Yeh, H.-W. Chen, S.-L. Tsai, and S.-D. Wang, "Disrupting image-translation-based deepfake algorithms with adversarial attacks," in *Proc. IEEE/CVF Conf. on Comput Vis. Pattern Recognit.(CVPR) Workshops*, 2020, pp. 53–62.

[30] H. Huang, Y. Wang, Z. Chen, Y. Zhang, Y. Li, Z. Tang, W. Chu, J. Chen, W. Lin, and K.-K. Ma, "Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 989–997.

[31] Z. Li, N. Yu, A. Salem, M. Backes, M. Fritz, and Y. Zhang, "Unganable: Defending against gan-based face manipulation," *arXiv:2210.00957*, 2022.

[32] C. Yang, L. Ding, Y. Chen, and H. Li, "Defending against gan-based deepfake attacks via transformation-aware adversarial faces," in *2021 Int. Joint Conf. on Neural Net.*, 2021, pp. 1–8.

[33] W. Sun, J. Zhou, Y. Li, M. Cheung, and J. She, "Robust high-capacity watermarking over online social network shared images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1208–1221, 2021.

[34] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Adv. in neural info. proc. sys.*, vol. 31, 2018.

[35] X. Yang, W. Xiang, H. Zeng, and L. Zhang, "Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme," in *Proc. IEEE Int. Conf. Comput. Vis.(ICCV)*, 2021, pp. 4781–4790.

[36] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, p. 600612, Apr 2004.

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv:1706.06083*, 2017.

[38] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE Int. Conf. on Auto. Face & Gesture Recognit.*, 2018, pp. 67–74.

[39] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014.

[40] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[42] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. on Comput Vis. Pattern Recognit.(CVPR)*, 2015, pp. 815–823.

Yunming Zhang received the MS degree from the School of Information Science and Engineering, Shandong Normal University, Jinan, China, in 2022. She is currently working toward the doctors degree in the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. She research interests include Artificial intelligence security, multimedia forensics, and deepwatermarking.

Dengpan Ye (Member, IEEE) received the B.Sc. degree in automatic control from SCUT in 1996 and the Ph.D. degree from NJUST in 2005. He was a Post-Doctoral Fellow in information system with Singapore Management University. Since 2012, he has been a Professor with the School of Cyber Science and Engineering, Wuhan University. He has authored or coauthored over 70 refereed journal and conference papers. His research interests include machine learning and multimedia security.

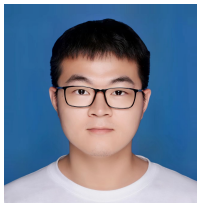
Xie Caiyun received her bachelor's degree from the School of Computer Science, Wuhan University, Hubei, China, in 2023. She is currently pursuing a master's degree in the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. Her research interests include artificial intelligence security and deep watermarking technology.

Long Tang received the MS degree in computer science from Xidian University, Xian, P.R. China, in 2021. He is currently working toward the doctors degree in the School of Cyber Science and Engineering, Wuhan University, Wuhan, P.R. China. His research interests include theories in Adversarial attacking and defending, DeepFake generation and detection, and Verification of neural networks.

Xin Liao (Senior Member, IEEE) received the B.E. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications in 2007 and 2012, respectively. He is currently a Professor and a Doctoral Supervisor with Hunan University, China. He worked as a Post-Doctoral Fellow with the Institute of Software, Chinese Academy of Sciences, and also a Research Associate with The University of Hong Kong. From 2016 to 2017, he was a Visiting Scholar with the University of Maryland, College Park, USA. His current research interests include multimedia forensics, steganography, and watermarking. He is the Secretary and a member of Technical Committee (TC) on Multimedia Security and Forensics of AsiaPacific Signal and Information Processing Association, a member of TC on Computer Forensics of the Chinese Institute of Electronics, and a member of TC on Digital Forensics and Security of the China Society of Image and Graphics. He is serving as an Associate Editor for the IEEE Signal Processing Magazine.

Ziyi Liu received the B.S. degree from Chongqing Jiaotong University, Chongqing, China, in 2017, and the M.S. degree from the Guilin University Of Electronic Technology, Guilin, China, in 2022. Currently studying for a PhD at Wuhan University. His research interests include machine learning, network security, and federated learning.

Chuanxi Chen received the MS degree from the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China, in 2019. And he received the doctors degree in the School of Cyber Science and Engineering, Wuhan University, Wuhan, China, in 2023. Now, he is the teacher in College of Computer and Cyber Security, Fujian Normal University, Fuzhou, China. His research interests include artificial intelligence security, machine learning, and privacy protection.



Jiacheng Deng received the MS degree in computer science from Ningbo University, Ningbo, P.R. China, in 2023. He is currently working toward the doctors degree in the School of Cyber Science and Engineering, Wuhan University, Wuhan, P.R. China. His research interests include theories in Synthetic Speech Detection, Adversarial attacking, and Explanation of Neural Networks.