

One-Identity-One-Key: Region-aware Watermark for Proactive Deepfake Detection and Robust Source Tracing

Ziyuan He^{ID}, Zhiqing Guo^{ID}, *Member, IEEE*, Liejun Wang^{ID}, *Member, IEEE*,
Xiaoming Tao^{ID}, and Xin Liao^{ID} *Senior Member, IEEE*

Abstract—Proactive forensic methods based on watermarking have attracted increasing attention as an effective paradigm for deepfake detection and source tracing. However, the existing methods still have shortcomings in encryption security, robustness and detection performance when dealing with complex forgery. In this paper, we propose a region-aware watermarking based proactive deepfake detection framework named Segfacemark. Building on global watermark embedding, this method explicitly models the sensitivity of various semantic facial regions to tampering. By employing region-wise partitioning and redundant embedding, the framework strengthens the protection of key facial areas, thereby comprehensively enhancing watermark robustness. In addition, the proposed framework employs a dynamic encryption mechanism that derives personalized keys from identity features to encrypt structural features, thereby realizing a “one-identity-one-key” watermark generation strategy and significantly enhancing system security and robustness against forgery. In the detection stage, the dual verification mechanism of identity decryption consistency and global region-aware watermark consistency not only improves detection accuracy but also enables forgery type identification, and further enhances the interpretability of deepfake detection results through region-aware watermarks. Experimental results show that the proposed method achieves high accuracy and robustness in deepfake detection, while preserving visual fidelity. The code will be released at: <https://github.com/vpseg-research/Segfacemark>.

Index Terms—Deepfake Detection, Source Tracing, Region-aware Watermarking, Dynamic Encryption.

I. INTRODUCTION

THE rapid development of facial deepfake technology has completely changed the generation mode of visual media, these manipulation techniques are mainly divided into two categories: face swapping [1] which replaces one person's

This work was supported in part by the National Natural Science Foundation of China under Grant 62462060, Grant 62302427, and Grant No. U22A2030, in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2023D01C175, in part by the Central Government Guides Local Science and Technology Development Fund Projects under Grant ZYYD2026ZY21, in part by Hunan Provincial Funds for Distinguished Young Scholars under Grant 2024JJ2025, in part by Hunan Provincial Key Science and Technology Program under Grant 2025QK2008. (Corresponding authors: Zhiqing Guo)

Ziyuan He, Zhiqing Guo, Liejun Wang, and Xiaoming Tao are with the School of Computer Science and Technology, Xinjiang University, Urumqi, 830017, China. Zhiqing Guo and Liejun Wang are also with the Xinjiang Multimodal Intelligent Processing and Information Security Engineering Technology Research Center, Urumqi, 830017, China (e-mail: 107552304059@stu.xju.edu.cn; {guozhiqing,wljxju,taoxm}@xju.edu.cn).

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

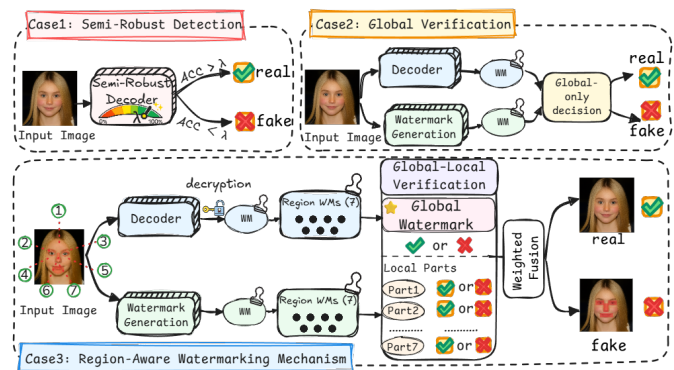


Fig. 1. Comparison of deepfake detection strategies: (a) Semi-robust detection is highly sensitive to threshold setting, leading to misjudgments; (b) global watermark verification lacks sensitivity to local manipulations; (c) the proposed region-aware watermarking mechanism that jointly exploits global and regional watermarks, improving detection accuracy.

identity with another, and face reenactment [2] which transfers facial expressions or motions while preserving the original identity. With the continuous evolution of generative models, the forged content has become almost indistinguishable from real images in terms of visual realism. Although these technologies promote the development of entertainment and virtual interaction, they are also abused for identity forgery and false information dissemination, which seriously threatens social trust [3], [4].

Passive deepfake detection methods aim to distinguish real and forged content by exploiting artifacts or inconsistencies introduced during manipulation. However, with the rapid evolution of generative models, such detectors often suffer from severe generalization degradation when facing unseen forgeries or real-world perturbations [5], [6]. Moreover, these methods typically output only binary predictions, lacking interpretability and traceability, which limits their practical forensic applicability [7]. Recently, research has increasingly focused on proactive forensics, which intervenes before forgery generation by embedding invisible watermarks into original media [8]. This strategy enables the implanted watermark to serve as a unique identity or traceable signature that remains within the content even after manipulation. By extracting or comparing the watermark from potentially forged images, the system can verify authenticity and detect deepfake manipulations.

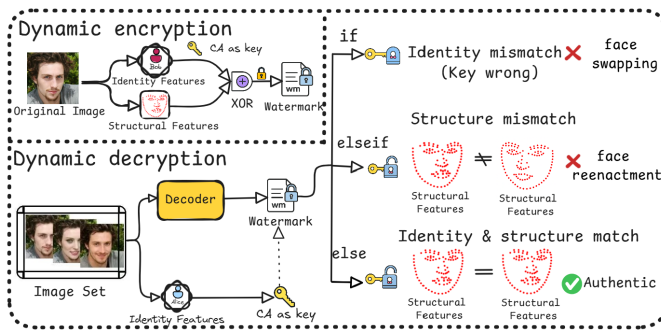


Fig. 2. Dynamic encryption and decryption mechanism. The identity-driven key generated by the Cellular Automata (CA) module combines with structural features for personalized watermark encryption, and the decryption verifies identity–structure consistency for authenticity determination.

Such an approach not only enhances detection reliability but also provides traceability for source attribution, offering a more interpretable solution than traditional passive detection methods.

Existing proactive forensic methods typically rely on semi-robust watermarking designs [9], [10], as shown in Case 1 of Fig. 1. This method distinguishes between benign noise and deepfake images by comparing the accuracy of watermark extraction: watermarks can be fully recovered under benign noise, but cannot be correctly extracted in deepfake images. Since these two tasks are inherently in conflict, it is difficult to achieve an optimal solution for both simultaneously. Therefore, setting a fixed threshold λ to differentiate between the two may lead to misjudgments or missed detections, reducing detection accuracy. Furthermore, most existing methods adopt a global watermark verification paradigm [11], which determines authenticity by evaluating the overall recovery accuracy of the extracted watermark across the entire image. However, as forgery techniques advance, this global matching strategy is less sensitive to localized tampering, often leading to unstable results in fine-grained forgery scenarios. Additionally, the robustness of these methods is limited under various forgery conditions, potentially causing instability in watermark extraction. Existing proactive forensic methods generally overlook watermark security, typically embedding randomly generated watermarks into the original content [12], [13]. Even with the introduction of encryption or key protection mechanisms, many designs still face challenges, especially when these encryption methods rely on fixed rules or static parameters, which create the security issue of the “one-key-fits-all” paradigm [14]. If the encryption parameters are leaked, the watermark becomes vulnerable to tampering, weakening the security of the system. Moreover, encryption frameworks based on cellular automata assign keys randomly through a key pool, but this key distribution is unrelated to the semantic information of the image itself [11], making the defense mechanism lack specificity. Attackers can directly use cracked rules to tamper with or forge watermarks without analyzing the image’s features. In addition, existing encryption technology serves a single purpose, mainly to ensure the security of the watermark, without utilizing the decryption process for any additional functions.

To address the above limitations, this paper proposes a region-aware proactive watermarking framework for reliable deepfake detection and robust source tracing. In Case 3 of Fig. 1, the proposed framework adopts a region-aware watermarking mechanism, where seven corresponding facial-region watermarks are derived from a single decoded global watermark. This design enables the model to jointly leverage global consistency and region-specific cues for more reliable forgery detection. In the embedding stage, the framework strengthens the watermark signals in forgery-prone facial regions, enabling higher robustness against local tampering and various post-processing operations while preserving the visual quality of the image. To further enhance the security and anti-counterfeiting capability of the watermark, this paper proposes a dynamic identity-driven encryption and decryption mechanism, as shown in Fig. 2. The mechanism generates personalized keys based on the identity features of the image and combines them with the structural features of the image to create a content-dependent encrypted watermark. The structural features encode stable spatial geometry and are highly sensitive to local deformations and high-frequency artifacts caused by deepfake manipulations, making them effective cues for improving forgery detection. Since the key dynamically changes with identity and is not reused across different images, this mechanism implements an identity-specific “one-identity-one-key” security paradigm, significantly improving the system’s security against attacks. On this basis, in the detection stage, the framework further improves detection accuracy through a dual verification mechanism that combines identity decryption consistency and global and regional watermark consistency. First, the authenticity of the image is verified by dynamically decrypting the identity features. If decryption fails, the image is classified as a face swapping forgery. If decryption is successful, the consistency between the global and region-aware watermarks is checked, with different weights assigned to each region based on its tampering sensitivity. The final authenticity judgment is made through weighted fusion. If the consistency check passes, it indicates that the identity and structural features are fully consistent, and the image is classified as a real image. If the consistency check fails, it suggests a face reenactment forgery. Moreover, inconsistencies in specific regional watermarks naturally reveal the suspected tampered areas, enhancing the interpretability of deepfake detection results. Our contributions can be summarized as follows:

- We propose a region-aware watermarking framework. While maintaining global embedding, the framework implements region-wise partitioning and redundant embedding specifically for key semantic facial regions, thereby enhancing its resistance to local deepfake tampering and improving overall robustness. During detection, it combines the consistency of global and regional watermarks, which not only increases the accuracy of authenticity verification but also makes the detection results interpretable.
- We propose an identity-driven dynamic encryption mechanism that generates a unique key from facial identity features, enabling a “one-identity-one-key” security design and resisting forgery and reverse analysis. During decryption, identity

and structural consistency are jointly checked, allowing the framework to distinguish different types of deepfake and improving the accuracy and reliability of forgery detection.

- The experimental results show that the proposed framework achieves excellent detection accuracy, robustness, and imperceptibility across multiple datasets, and even when trained on a single deepfake type, it still demonstrates strong generalization ability.

The remainder of this paper is organized as follows. Section II reviews the related work on proactive forensics. Section III presents the proposed framework in detail. Section IV describes the experimental setup and reports the evaluation results. Section V concludes the paper.

II. RELATED WORKS

Proactive forensics aims to embed verifiable watermarks in authentic images, enabling active authenticity verification and source traceability against malicious manipulations. Early studies primarily focused on identity-driven watermarking. FakeTagger [15] embeds identity-related watermark signals into facial images to support proactive deepfake detection and source attribution. Yu *et al.* [16] further proposed an artificial fingerprinting approach that injects traceable fingerprints into training data, enabling generative models to produce outputs with identifiable marks for active forgery detection and attribution. Building upon this line of research, SepMark [12] introduces a separable watermarking framework based on a single encoder and dual decoders, achieving unified deepfake detection and source tracing. Subsequent works further strengthened identity-aware watermarking designs. Wang *et al.* [14] proposed an identity-aware robust watermarking method that encodes facial identity semantics into the watermark for proactive deepfake face swapping detection and source tracing. Zhao *et al.* [17] explored a proactive detection framework by embedding watermarks into facial identity representations and verifying identity-level consistency to detect face swapping manipulations. From a different perspective, NullSwap [18] adopts a proactive defense strategy that suppresses or obfuscates facial identity cues in advance, aiming to prevent deepfake face swapping attacks rather than performing post-hoc forensic analysis. Beyond identity information, structural cues have also been widely exploited in proactive watermarking. LampMark [11] generates training-free landmark-aware watermarks from facial geometry and performs forgery detection via structural consistency analysis, while EditGuard [19] unifies copyright verification and tamper-agnostic localization within a single structural watermarking framework, achieving high-precision verification and pixel-level localization across diverse AIGC editing scenarios. Similarly, OmniGuard [20] proposes a hybrid proactive-passive watermarking framework for robust tamper localization and copyright protection under diverse AIGC editing conditions. WaterLo [21] introduces a region-wise semi-fragile watermark that disappears only in manipulated regions, enabling both authenticity verification and forgery localization. FractalForensics [22] further explores parameter-driven fractal structures to balance robustness and fragility under black-box conditions, supporting active deepfake detection and fine-grained tampering localization. More

recently, proactive watermarking has been extended to model-driven and generative paradigms. WaveGuard [13] embeds watermarks in the frequency domain and enforces structural consistency to support robust deepfake detection and source tracing. DiffMark [23] injects facial and watermark conditions into the denoising process of diffusion models, enhancing robustness against deepfake manipulations while maintaining high invisibility. KAD-Net [24] combines Kolmogorov–Arnold theory with a differential sensing network in a dual-branch architecture, significantly improving watermark robustness and detection accuracy for deepfake forensics. In the direction of semi-fragile watermarking, FaceSigns [9] achieves selective fragility by keeping the watermark stable under benign transformations while breaking under deepfake operations, thereby improving the reliability of authenticity verification. Despite these advances in robustness, structural sensitivity, fragility control, and traceability, existing proactive forensic methods still face inherent trade-offs among detection accuracy, security, and imperceptibility under complex manipulation scenarios. These limitations motivate the development of more unified watermarking strategies that can jointly enhance robustness and detection reliability while preserving security and visual quality for challenging deepfake forensics tasks.

III. METHODOLOGY

A. Problem Formulation

In this study, we propose a proactive deepfake detection framework based on dynamic encrypted watermark generation and region-aware watermarking. The overall pipeline includes encrypted watermark construction and semantic region watermark embedding during the training stage, followed by identity decryption consistency and global as well as region-aware watermark consistency dual verification during inference, along with forgery type classification. Specifically, in the training stage, the original image I_o and watermark W_o are first processed by the watermark generator to produce an identity-related encrypted watermark, which is then combined with the facial segmentation mask and fed into the encoder to perform hierarchical embedding across seven semantic facial regions, resulting in the watermarked image I_w . This image is subsequently processed by a noise pool to generate a normally degraded image I_{com} and a maliciously forged image I_{mal} , both of which are passed through a unified decoder that simultaneously recovers a global watermark and seven region-specific local watermarks, enabling the model to learn robust watermark extraction and region-level sensitivity under various perturbations. In the inference stage, we first extract the identity features of the test image using an identity recognition network. The Cellular Automata (CA) module then uses these identity features to initialize the automata states and generates a dynamic identity key through iterative evolution, which is applied to decrypt the global encrypted watermark produced by the decoder. If the decryption fails, the identity is inconsistent, and the image is directly classified as a face-swapping forgery. If the decryption succeeds, we compare the decrypted structural watermark with the structural

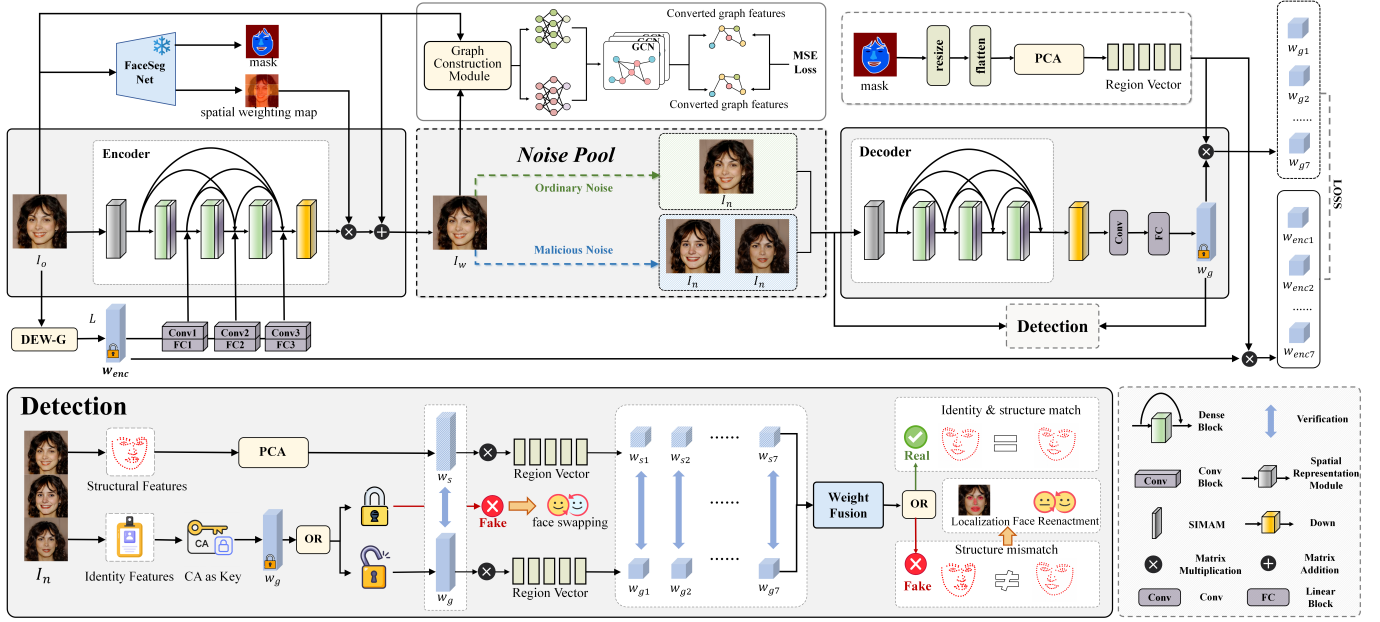


Fig. 3. Overall framework. Given an input face image I_o , **FaceSegNet** generates a spatial weighting map and mask to guide the **encoder** for region-aware redundant embedding of the global watermark w_{enc} , producing the watermarked image I_w . Meanwhile, a graph construction module, along with the **Structural Consistency Graph Neural Network (SC-GNN)**, ensures structural consistency between I_o and I_w during the watermark embedding process. The watermarked image is then processed by the **Noise Pool** to simulate common noise and deepfake manipulations, generating a distorted image I_n . Guided by the mask, the **decoder** recovers both the global watermark and the region-specific watermarks. In the detection stage, the dual verification mechanism of identity decryption consistency and global as well as region-aware watermark consistency enhances detection accuracy, while simultaneously enabling forgery type classification.

watermark that is re-extracted and quantized from the test image. This comparison is performed at the global level and across the seven facial regions, using the same region masks as in the training stage, which come from the face semantic segmentation network. By measuring the global structural consistency together with the local consistency of the seven regions, and by fusing these scores with predefined region weights, the system can determine whether the image contains structure changes caused by expression transfer or motion-driven reenactment. This step also provides interpretability of the tampering results. Overall, this process forms a complete closed loop, moving from identity-driven dynamic watermark encryption and region-aware embedding to globally robust watermark recovery and forgery identification, and enables unified deepfake detection, forgery-type classification, and source tracing.

B. Model Architecture

1) Dynamic Encrypted Watermark Generator (DEW-G):

We design a training-free Dynamic Encrypted Watermark Generator (DEW-G) to create a unique encrypted watermark for each face image. The main idea is to jointly encode facial geometry and identity information and then combine them through a dynamic key mechanism. DEW-G consists of four steps: structural watermark generation, identity digest construction, dynamic key evolution, and final structure–identity encryption.

a) *Structural Watermark Generation.*: Given an input face, $d_{lm} = 68$ facial landmarks are detected using the

dlib [25] landmark detector and arranged into a vector:

$$\mathbf{e}_{lm} = [x_0, y_0, x_1, y_1, \dots, x_{d_{lm}-1}, y_{d_{lm}-1}]^T \in \mathbb{R}^{2d_{lm}}. \quad (1)$$

where (x_i, y_i) denotes the image-plane coordinates of the i -th landmark. We then project $\mathbf{e}_{lm}$ into a compact structural space using principal component analysis (PCA) [26]:

$$\mathbf{z} = P(\mathbf{e}_{lm}), \quad \mathbf{z} \in \mathbb{R}^{d_w} \quad (2)$$

where $P(\cdot)$ denotes the PCA projection learned on a reference dataset and fixed during inference. Each dimension of $\mathbf{z}$ is normalized using the statistics estimated from the reference set and binarized to obtain a structural watermark:

$$\tilde{z}_i = \frac{z_i - z_i^{\min}}{z_i^{\max} - z_i^{\min} + \varepsilon} \quad (3)$$

$$m_i^{str} = \mathbb{I}(\tilde{z}_i \geq 0.5), \quad \mathbf{w}^{str} \in \{0, 1\}^{d_w}. \quad (4)$$

where $z_i^{\min}$ and $z_i^{\max}$ are computed on the reference dataset and reused for subsequent samples.

b) *Identity Digest and Initial Key.*: We extract the identity embedding $\mathbf{f}^{id}$ using ArcFace [27]. To construct an identity-dependent initial key, the embedding is first serialized and mapped to a binary sequence via a SHA-256 hash:

$$s = \text{SHA256}(\text{Serialize}(\mathbf{f}^{id})). \quad (5)$$

The resulting binary sequence is then truncated or zero-padded to obtain an initial key of length d_w :

$$\mathbf{k}^{(0)} \in \{0, 1\}^{d_w}. \quad (6)$$

c) *Dynamic Identity Key Evolution.*: Starting from the initial key $\mathbf{k}^{(0)}$, we evolve the identity key using a one-dimensional Rule-30 cellular automaton:

$$k_i^{(t+1)} = k_{(i-1) \bmod l}^{(t)} \oplus (k_i^{(t)} \vee k_{(i+1) \bmod l}^{(t)}), \quad i = 0, \dots, l-1. \quad (7)$$

where $l = d_w$. This process produces a sequence of evolved keys $\{\mathbf{k}^{(t)}\}_{t=0}^T$. A pseudo-random seed derived from the identity embedding is used to deterministically select an index subset $\mathcal{S}$ from the key sequence, and the final dynamic identity key is computed by XOR aggregation:

$$\hat{\mathbf{k}} = \bigoplus_{t \in \mathcal{S}} \mathbf{k}^{(t)}. \quad (8)$$

d) *Final Structure–Identity Encryption.*: Before encryption, we concatenate the identity verification bits with the structural watermark to form the plaintext watermark:

$$\mathbf{m} = [\mathbf{v}_{\text{id}} \parallel \mathbf{w}^{\text{str}}], \quad (9)$$

where $\mathbf{v}_{\text{id}}$ denotes the identity verification bits derived from a predefined verification sequence, and $\parallel$ denotes concatenation. The verification bits are placed at the beginning of the watermark bitstream and are used to verify the success of identity-driven decryption during inference. The encrypted watermark is then obtained by

$$\mathbf{w}_{\text{enc}} = \mathbf{m} \oplus \hat{\mathbf{k}}, \quad \mathbf{w}_{\text{enc}} \in \{0, 1\}^{d_w}. \quad (10)$$

In this way, the encrypted watermark is jointly bound to facial geometry and identity information. During training, $\mathbf{w}_{\text{enc}}$ is used as the embedding target of the encoder.

2) *Encoder–Decoder Network.*: The encoder aims to embed the dynamic encrypted watermark into the U channel of a face image while preserving high perceptual quality. The U channel represents chrominance information, which has a smaller impact on visual quality compared to luminance (Y channel). Therefore, embedding the watermark in this channel minimizes interference with perceptual quality to the greatest extent. Given an input image, the model first employs a Spatial Representation Module to extract low-level spatial features from the U channel, capturing fundamental facial structural information. A face semantic segmentation network [28] parses the input image to obtain seven predefined facial regions, including the left eye, right eye, nose, mouth, face skin, left eyebrow, and right eyebrow. These regions cover the main semantic components of a face and are selected to capture local inconsistencies caused by facial manipulations. The mouth, eyes, and eyebrows are sensitive to expression changes and local motion, while the nose and face-skin regions provide complementary cues related to facial structure and texture consistency. Therefore, this seven-region partition supports region-aware watermark consistency verification. Based on the segmentation results, a single-channel spatial attention map A is generated together with the corresponding semantic masks used for subsequent regional watermark recovery and detection. The one-dimensional watermark vector generated by DEW-G is projected into the spatial domain through multiple linear mappings followed by convolutional processing, producing watermark feature maps aligned with the spatial resolution

of the input image. These watermark features are progressively fused with the image features through a series of DenseBlocks, enabling gradual injection of watermark information into the feature representations. A SimAM attention module [29] is applied after each DenseBlock to refine local feature responses and enhance spatially informative structures. The fused features are further processed by a Down Module, which compresses the feature channels and produces a single-channel residual watermark signal R . The watermark is embedded into the original U channel through segmentation-guided spatial modulation:

$$U_w = U_o + \alpha AR \quad (11)$$

where U_o and U_w denote the original and watermarked U channels respectively, A represents the segmentation-guided spatial attention map, and α controls the global embedding strength. The modified U channel is then recombined with the remaining color channels to reconstruct the final watermarked image.

The decoder follows a backbone structure similar to the encoder and first reconstructs the global watermark w_g from the watermarked image. The recovered global watermark serves as the main reference for identity consistency verification and robust watermark recovery. After obtaining the global watermark, the decoder further generates region-level watermarks using the facial region masks produced by the face segmentation network. For each predefined facial region, the corresponding binary mask is first downsampled and flattened into a vector. The vector is then passed through a pretrained PCA model to obtain a low-dimensional region vector $v_{\text{region}}^{(r)}$ with the same dimensionality as the watermark. The PCA output is normalized using the stored minimum and range values to ensure numerical stability. The normalized values are further clipped and binarized to obtain the final region vector. The local watermark of region r is then computed by element-wise modulation between the global watermark and the region vector:

$$w^{(r)} = w_g \odot v_{\text{region}}^{(r)}, \quad r = 1, \dots, 7. \quad (12)$$

where $\odot$ denotes element-wise multiplication.

With this design, region-level watermarks are derived from the global watermark through region-vector modulation. This avoids using separate decoding branches for each facial region and keeps the network structure compact. The decoder finally outputs one global watermark and seven region-specific watermarks. The global watermark is used for identity-level verification, while the regional watermarks are used for structure-level consistency analysis to distinguish face-swapping and reenactment manipulations.

3) *Structural Consistency Graph Neural Network (SC-GNN)*: To further enhance watermark imperceptibility, we introduce a Structure Consistency Graph Neural Network (SC-GNN) that constrains the watermarked image to preserve the fine-grained structural patterns of the original face. In this module, each image is represented as a graph where every pixel location becomes a node, and the flattened 3×3 local patch around it is used as the node feature. Each node is

connected to its eight spatial neighbors, enabling the graph to capture local texture variations and geometric continuity. The constructed graph is processed by several graph convolution layers followed by global pooling, producing a compact structural descriptor of the image. SC-GNN is applied to both the original and the watermarked image, and their graph-level embeddings are enforced to remain consistent during training. This encourages the encoder to avoid disturbing important facial textures and to embed the watermark primarily in visually insensitive regions, thereby improving imperceptibility without compromising robustness.

C. Identity–Structure Dual Verification

The framework employs an identity-structure dual verification mechanism during inference, enabling it to not only determine whether an image has been tampered with but also distinguish between face swapping and face reenactment attacks, while providing localized tampering cues. The images are categorized into three types: common distortions(common), face swap(swap), and face reenactment(reenact), with these images being processed together as a whole. In this dual verification mechanism, structural features and identity features are assigned different roles. The structural features are encoded into watermarks to describe facial geometry, while the identity features are used to generate dynamic keys that bind the watermark to a specific identity, enabling the framework to handle both identity replacement and structural manipulation in a unified manner. Structural features are extracted from the image, upon which an unencrypted watermark W_s is generated based on these features. Concurrently, identity features are extracted from the image to generate a decryption key. The generated identity key is first used to decrypt the watermark recovered by the decoder. After decryption, the first several bits of the watermark sequence are regarded as identity verification bits and are compared with the predefined verification sequence for consistency. When the identity feature is consistent with the identity bound during encryption, the identity-driven key can correctly recover these verification bits. In contrast, face swapping changes the identity feature of the image, causing the regenerated decryption key to mismatch the original key and thus breaking the consistency of the verification bits. Therefore, the consistency of the verification bits serves as direct evidence for identifying face-swapping forgeries. If the identity verification succeeds, the remaining part of the decrypted sequence is regarded as the structural watermark W_g , which is further used for subsequent structural consistency verification. The verification of structural features is carried out by calculating the seven region-specific watermarks of W_s and W_g , namely W_s^1 through W_s^7 and W_g^1 through W_g^7 . The differences between the global watermark and the seven region-specific watermarks are compared to assess structural consistency. As W_s and W_g reflect the structural features of the image, this step effectively verifies the structural consistency of the image. The BER between the global watermark and the region-specific watermarks is calculated and weighted by the region weights λ_r , resulting in an overall **score** D :

$$D = \lambda_0 \cdot \text{BER}(W_s, W_g) + \sum_{r=1}^7 \lambda_r \cdot \text{BER}(W_s^{(r)}, W_g^{(r)}) \quad (13)$$

where $\text{BER}(W_s, W_g)$ represents the global watermark BER, $\text{BER}(W_s^{(r)}, W_g^{(r)})$ denotes the BER for the r -th region's watermark, and λ_0 and λ_r are the corresponding weight coefficients.

Based on the score value D , reenact and common noise are distinguished. Specifically, a threshold D_{thresh} is set, and when the score value D is smaller than this threshold, the image is classified as common. When the score value D is greater than the threshold, the image is classified as reenact forgery type. The framework employs an identity-structure dual verification mechanism, enabling efficient detection of deepfake images and distinguishing between different types of forgeries. It not only determines whether an image has been tampered with but also accurately distinguishes the forgery types. Additionally, it utilizes region-specific Bit Error Rate to assist in locating tampered areas, thereby improving the accuracy and interpretability of the detection.

D. Loss Functions

To jointly optimize the invisibility, robustness, and region-level structural consistency of the watermark, we define several loss functions for the encoder–decoder framework. The purposes and formulations of each loss component are described as follows.

- *Embedding Loss*: The image reconstruction loss restricts the watermarked image to remain visually consistent with the original, preventing noticeable artifacts and thereby ensuring the invisibility of the embedded watermark.

$$\mathcal{L}_{\text{img}} = L_2(I_o, \text{Enc}(\theta, I_o, w)) = \|I_o - I_w\|_2^2. \quad (14)$$

where $\text{Enc}(\theta, I_o, w)$ denotes the encoder parameterized by θ , which takes the original image I_o and the watermark vector w as inputs and produces the watermarked image I_w . The function $L_2(\cdot, \cdot)$ represents the mean squared error (MSE) loss between two images.

- *Global Watermark Loss*: The global watermark loss supervises the decoder to accurately recover the complete global watermark from the distorted watermarked image w_g . This loss serves as the core foundation for identity consistency verification and robust watermark recovery.

$$\mathcal{L}_{\text{global}} = L_2(w_{\text{enc}}, \text{Dec}(\theta, I_w)) = \|w_{\text{enc}} - \hat{w}_g\|_2^2, \quad (15)$$

where $\text{Dec}(\theta, I_w)$ denotes decoder parameterized by θ , which takes the watermarked image I_w as input and predicts the recovered global watermark $\hat{w}_g$.

- *Regional Watermark Loss*: The regional watermark loss supervises the decoder to recover the local watermark corresponding to each semantic facial region. For each region r , the original watermark is mapped into the region's semantic space to obtain the target regional watermark $W_{\text{enc}}^{(r)}$, which is matched against the predicted regional watermark $W_g^{(r)}$. This loss enforces structural consistency at the regional level,

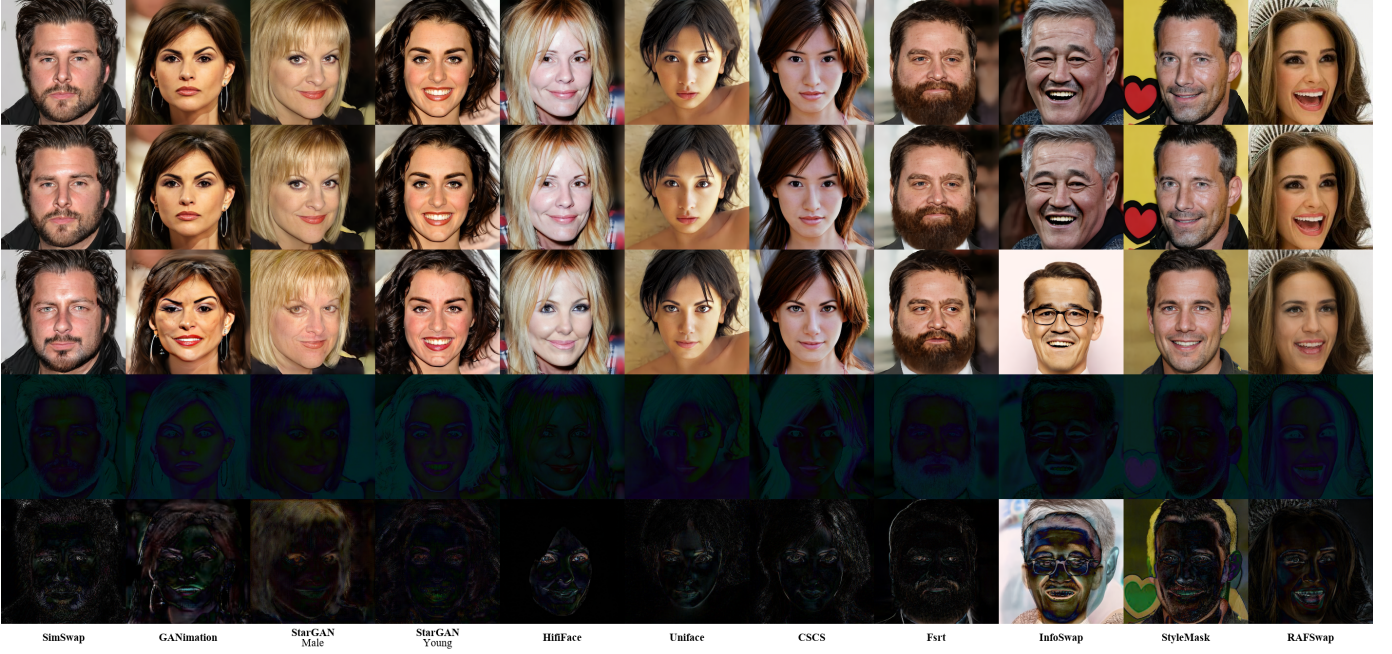


Fig. 4. More malicious distortions. From top to bottom are the origin image I_o , the watermarked image I_w , the noised image I_n , the residual signal $|I_o - I_w|$, and $|I_n - I_w|$. All images are resized to 256×256 for consistency.

enabling fine-grained forgery discrimination and region-wise structural analysis.

$$\mathcal{L}_{regional} = \frac{1}{R} \sum_{r=1}^R \|W_{enc}^{(r)} - W_g^{(r)}\|_2^2. \quad (16)$$

where R denotes the number of predefined regions, $W_{enc}^{(r)}$ is the target regional watermark derived from the original watermark W_{enc} , and $W_g^{(r)}$ is the predicted regional watermark reconstructed by the decoder.

- **SC-GNN Loss:** The SC-GNN loss enforces structural consistency between the original image and the watermarked image by comparing their graph-based structural representations. Both images are transformed into graphs and processed by the GNN to extract structural feature embeddings. This loss penalizes their discrepancies and mitigates structural distortions introduced during watermark embedding.

$$\mathcal{L}_{gnn} = \|f_o - f_w\|_2^2. \quad (17)$$

where f_o and f_w denote the GNN-derived structural features of the original image I_o and the watermarked image I_w , respectively.

- **Total Loss:** The total loss integrates all objectives to jointly optimize watermark invisibility, robustness, and regional structural consistency. A weighted combination of the individual loss components is employed to balance their contributions during training.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_{global} + \lambda_3 \mathcal{L}_{regional} + \lambda_4 \mathcal{L}_{gnn}. \quad (18)$$

where λ_1 , λ_2 , λ_3 , and λ_4 are hyperparameters that control the contribution of each loss term.

IV. EXPERIMENTS

A. Implementation Details

Datasets. The CelebA-HQ dataset [39], [40] is adopted in this study. It contains 30,000 facial images from 6,217 identities and follows the official training, validation, and test split. CelebA-HQ also provides the facial images used for encrypted watermark generation in the DEW-G. To evaluate the cross-dataset generalization ability, we further employ LFW [41], FaceForensics++ [42], DFDC [43], and Celeb-DF v2 [44]. For the video datasets, frames are sampled from videos and evaluated at the image level. All experiments are conducted at two resolutions, 128×128 and 256×256 .

Manipulation Pools. To comprehensively evaluate the stability of our model under various degradation and manipulation scenarios, we categorize all operations into two groups: a Common Pool and a Malicious Pool. The Common Pool consists of typical image degradations frequently encountered in practice, including JPEG compression, resizing, Gaussian blur, median blur, dropout, salt-and-pepper noise, and Gaussian noise. The Malicious Pool comprises two categories of deepfake manipulations: face swapping and face reenactment. For face-swapping perturbations, we employ SimSwap [1], Uniface [35], CSCS [33], HifiFace [34], and RAFSwap [32]. For face reenactment manipulations, we adopt GANimation [36], StarGAN [38], InfoSwap [37], FSRT [31] and StyleMask [2]. In the face-swapping setting, target identities are randomly drawn from the CelebA validation split. All manipulated images are generated using publicly available pretrained models to ensure the authenticity and diversity of the forged samples.

Parameters. The proposed method is implemented using PyTorch, and all experiments are conducted on a single cloud-based virtual GPU with 48GB of memory. The model is

TABLE I
ROBUSTNESS EVALUATION OF WATERMARK RECOVERY ON CELEBA-HQ UNDER COMMON AND DEEFAKE DISTORTIONS AT 128×128 RESOLUTION, MEASURED BY GLOBAL BER AND AVERAGE REGIONAL BER.

Distortion	SepMark [12]	EditGuard [19]	LampMark [11]	KAD-Net [24]	DiffMark [23]	WaveGuard [13]	LIDMark [30]	Ours	
								Global	Regional
Identity	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
JpegTest	0.2136%	1.4152%	0.0000%	0.0967%	1.2501%	0.0000%	1.4705%	0.0000%	0.0000%
Resize	0.0059%	0.0017%	0.0000%	0.0000%	0.0212%	0.0000%	0.0025%	0.0000%	0.0000%
GaussianBlur	0.0024%	1.6027%	0.2104%	0.0000%	0.0000%	0.0021%	0.0000%	0.0011%	0.0000%
MedianBlur	0.0012%	1.6212%	0.1023%	0.0000%	0.0000%	0.0011%	0.0018%	0.0022%	0.0009%
Dropout	1.9125%	0.9257%	0.0000%	0.0000%	0.3104%	0.0000%	0.0000%	0.0000%	0.0000%
SaltPepper	0.0413%	0.0925%	5.1365%	0.0047%	0.0904%	0.0187%	0.0021%	0.0070%	0.0016%
GaussianNoise	0.7460%	1.6772%	0.2561%	0.2632%	1.3501%	0.0253%	0.0000%	0.0054%	0.0013%
Average	0.3654%	0.9170%	0.7132%	0.0456%	0.3778%	0.0059%	0.1846%	0.0020%	0.0005%
FSRT [31]	0.7814%	5.7701%	2.9307%	2.0145%	0.1201%	0.0785%	0.6125%	0.0048%	0.0025%
RAFSwap [32]	13.2547%	12.0971%	4.0974%	11.2541%	5.9706%	0.2093%	0.5781%	0.0188%	0.0095%
SimSwap [1]	20.0295%	16.0721%	0.1975%	13.1480%	5.5802%	0.1637%	0.0297%	0.0781%	0.0201%
CSCS [33]	0.6801%	0.9901%	2.3015%	1.0214%	0.1321%	0.0091%	0.0109%	0.0087%	0.0024%
HifiFace [34]	0.9904%	6.7104%	2.9713%	1.0211%	0.1606%	0.0085%	0.8475%	0.0092%	0.0057%
Uniface [35]	0.3430%	9.1702%	3.1452%	0.7204%	0.0191%	0.0102%	0.5574%	0.0065%	0.0019%
GANimation [36]	0.0000%	0.7152%	0.0109%	0.0082%	0.0025%	0.0000%	0.0108%	0.0000%	0.0000%
InfoSwap [37]	6.2144%	11.2457%	4.2371%	4.0425%	7.0210%	3.0879%	0.2241%	0.0237%	0.0064%
StyleMask [2]	8.0221%	11.9012%	1.2563%	19.9024%	8.0147%	1.9125%	2.0852%	0.0025%	0.0007%
StarGAN (Male) [38]	0.1125%	7.6204%	0.1265%	0.0212%	4.6601%	0.0045%	15.0119%	0.0065%	0.0010%
StarGAN (Young) [38]	0.1560%	8.1745%	0.8771%	0.0094%	4.3225%	0.0064%	15.2140%	0.0071%	0.0008%
Average	4.5986%	8.2243%	2.0138%	4.8330%	3.2730%	0.4991%	3.1984%	0.0151%	0.0046%

TABLE II
EVALUATION OF WATERMARKED IMAGE QUALITY ON CELEBA-HQ USING PSNR AND SSIM ACROSS DIFFERENT MODELS, IMAGE RESOLUTIONS, AND MESSAGE LENGTHS

Model	Image Size	Message Length	PSNR $\uparrow$	SSIM $\uparrow$
SepMark [12]	128×128	30	38.3112	0.9588
EditGuard [19]	128×128	30	37.1664	0.9796
LampMark [11]	128×128	64	44.6251	0.9921
DiffMark [23]	128×128	30	41.2869	0.9776
KAD-Net [24]	128×128	30	38.6025	0.9564
WaveGuard [13]	128×128	30	45.6025	0.9871
LIDMark [30]	128×128	152	41.0319	0.9855
Ours	128×128	30	46.3421	0.9948
SepMark [12]	256×256	128	38.5646	0.9328
EditGuard [19]	256×256	128	37.8254	0.9512
LampMark [11]	256×256	128	42.7825	0.9869
DiffMark [23]	256×256	128	41.9572	0.9769
KAD-Net [24]	256×256	128	38.9801	0.9334
WaveGuard [13]	256×256	128	42.2509	0.9829
LIDMark [30]	256×256	152	44.1974	0.9906
Ours	256×256	128	44.3901	0.9911

optimized with the Adam optimizer, with the initial learning rate set to 5×10^{-4} . The watermark embedding strength is fixed to $\alpha = 0.04$, a value determined through preliminary experiments to achieve a desirable balance between visual quality and embedding robustness. As shown in Formula 18, the loss weights are set to $\lambda_1 = 15$, $\lambda_2 = 1.0$, $\lambda_3 = 0.01$, and $\lambda_4 = 0.01$. The watermark length is set as a fixed hyperpa-

rameter, mainly determined by the image resolution and the available embedding capacity. In this work, we adopt a 30-bit watermark for 128×128 images and a 128-bit watermark for 256×256 images to balance structural representation capacity, embedding robustness, and visual imperceptibility. The watermark length is not dynamically adjusted according to the structural complexity of each face image or any specific forgery type.

Baseline. To comprehensively evaluate the performance of the proposed model in real-world scenarios, we select representative baseline methods from existing face watermarking and proactive forensics literature. Due to the limited availability of open-source implementations for proactive watermarking and forensic frameworks, we additionally include several robust watermarking approaches to ensure sufficient evaluation coverage. In the category of robust watermarking, we adopt EditGuard [19], which span different technical paradigms including adversarial training, feature modeling, and generative protection. For proactive watermarking, we include SepMark [12], LampMark [11], DiffMark [23], WaveGuard [13], KAD-Net [24] and LIDMark [30] representing diverse design principles such as identity binding, region-wise encoding, and diffusion-based recovery. Since these baseline methods differ in their applicable resolutions and training configurations, we report each model only under the settings in which it can be properly executed.

B. Visual Quality

To quantitatively evaluate the impact of watermark embedding on image authenticity and structural preservation, this study adopts Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) as the primary visual quality

TABLE III
ROBUSTNESS EVALUATION OF WATERMARK RECOVERY ON CELEBA-HQ UNDER COMMON AND DEEFAKE DISTORTIONS AT 256×256 RESOLUTION, MEASURED BY GLOBAL BER AND AVERAGE REGIONAL BER.

Distortion	SepMark [12]	EditGuard [19]	LampMark [11]	KAD-Net [24]	DiffMark [23]	WaveGuard [13]	LIDMark [30]	Ours	
								Global	Regional
Identity	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
JpegTest	0.0071%	3.5801%	0.0000%	0.0038%	1.8101%	0.0012%	0.5485%	0.0000%	0.0000%
Resize	0.0000%	0.2815%	1.5809%	0.0000%	0.0011%	0.0000%	0.0071%	0.0000%	0.0000%
GaussianBlur	0.0000%	3.3701%	0.3151%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
MedianBlur	0.0000%	3.8504%	0.1825%	0.0125%	0.0000%	0.0000%	0.0047%	0.0000%	0.0000%
Dropout	0.0061%	0.9752%	2.4312%	0.0691%	0.6081%	0.0019%	0.0000%	0.0021%	0.0006%
SaltPepper	0.0011%	0.1210%	5.2514%	0.0073%	0.4802%	0.0096%	0.0101%	0.0018%	0.0000%
GaussianNoise	0.0525%	1.0274%	0.8254%	0.0803%	2.3704%	0.0055%	0.0000%	0.0012%	0.0000%
Average	0.0084%	1.6507%	1.3233%	0.0216%	0.6587%	0.0023%	0.0713%	0.0006%	0.0000%
FSRT [31]	0.7814%	37.1724%	3.2541%	1.0254%	4.0502%	0.0112%	0.2504%	0.0082%	0.0021%
RAFSwap [32]	5.3730%	46.0711%	4.1041%	10.0233%	2.1032%	1.0209%	0.6901%	0.0125%	0.0054%
SimSwap [1]	7.9125%	47.5625%	0.7125%	4.1217%	5.9621%	0.0096%	0.9708%	0.0089%	0.0018%
CSCS [33]	0.6801%	1.6173%	5.1402%	2.0632%	0.5606%	0.0181%	0.0036%	0.0054%	0.0010%
HifiFace [34]	0.9904%	45.0121%	9.1074%	0.8709%	0.8754%	0.0102%	0.7754%	0.0091%	0.0021%
Uniface [35]	0.3430%	49.4123%	7.0297%	2.0178%	2.2015%	0.0205%	0.0120%	0.0068%	0.0017%
GANimation [36]	0.0020%	7.5911%	0.0165%	0.0000%	0.0015%	0.0005%	0.0000%	0.0000%	0.0000%
InfoSwap [37]	4.9605%	10.2744%	3.0833%	3.0422%	2.5178%	2.0085%	0.8509%	0.0074%	0.0027%
StyleMask [2]	8.0221%	7.0329%	1.0634%	9.5685%	6.2564%	2.0118%	1.8725%	0.0202%	0.0097%
StarGAN (Male) [38]	0.1125%	2.1241%	3.1040%	0.4751%	3.8205%	0.0056%	6.0875%	0.0013%	0.0000%
StarGAN (Young) [38]	0.1560%	3.5414%	3.2167%	0.7546%	3.8751%	0.0079%	6.1622%	0.0009%	0.0000%
Average	2.6667%	23.4011%	3.6211%	3.0875%	2.9295%	0.4659%	1.6069%	0.0073%	0.0024%

TABLE IV
COMPARISON OF EMBEDDING STRATEGIES IN TERMS OF BER (%).

Strategy	128×128		256×256	
	Common	Malicious	Common	Malicious
Region-only	0.0362%	0.1908%	0.0297%	0.1412%
Global-regional	0.0020%	0.0151%	0.0006%	0.0073%

metrics. PSNR measures pixel-level deviations, whereas SSIM reflects structural and perceptual consistency. As shown in Table II, our method achieves the highest PSNR and SSIM scores at both 128×128 and 256×256 resolutions, indicating superior performance in controlling pixel distortion and maintaining structural fidelity. This advantage primarily stems from the SG-GNN-based structural geometric consistency module, which stabilizes facial topology through graph-structured modeling and geometric constraints, effectively suppressing local structural disturbances introduced during embedding. In addition, the fine-grained modeling of local facial shapes allows the watermark to align more closely with the original geometry, thereby enhancing robustness while preserving high imperceptibility. From a subjective perspective, the watermarked images in Fig.4 appear almost identical to the original ones, with no visible texture or structural discrepancies. The agreement between subjective observations and objective metrics further demonstrates the strong visual invisibility achieved by our method.

C. Robustness Test

To evaluate the robustness of the framework under different perturbation conditions, the noise is categorized into common distortions and deepfake manipulations, and the bit error rate

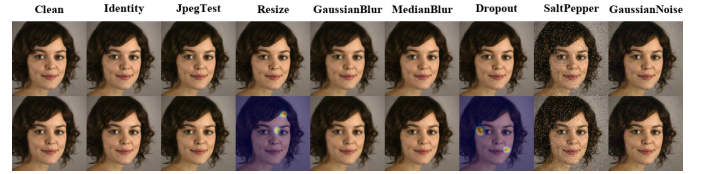


Fig. 5. Visualization of detection results for common noise based on regional BER. The face is divided into seven semantic regions, and the BER of each region is computed. Regions with higher error rates are shown with stronger warm colors to indicate reduced watermark consistency.

(BER) of watermark recovery is adopted as the evaluation metric. The BER is defined as

$$\text{BER}(w_{rec}, w_o) = \frac{1}{L} \sum_{i=1}^L \mathbf{1}(\varphi(w_{rec}^i) \neq w_o^i) \times 100\% \quad (19)$$

where w_o denotes the original embedded watermark, w_{rec} represents the recovered watermark, L is the watermark length, $\varphi(\cdot)$ is a binarization function, and $\mathbf{1}(\cdot)$ is the indicator function. The BER measures the proportion of incorrectly recovered bits, and a lower BER indicates stronger robustness under the corresponding perturbation.

As shown in Tables I and III, which report the robustness results at resolutions of 128 and 256, respectively. Under common noise conditions, the framework achieves a near-zero BER across all tested degradations, indicating that the embedded watermark can be stably recovered under various benign distortion scenarios. Under deepfake noise conditions, the framework also maintains a low BER, especially for high-fidelity face swapping and expression-driven attacks, where its watermark recovery performance is significantly superior to that of the baseline methods, demonstrating its effectiveness in resisting identity and structural disturbances introduced by malicious forgeries. Furthermore, the last column of Tables I

TABLE V

DEEFAKE DETECTION PERFORMANCE IN AUC SCORES AGAINST DIFFERENT FACE MANIPULATION ALGORITHMS ON CELEBA-HQ AT 128 AND 256 RESOLUTIONS. THE ROW OF "MIXED" EVALUATES THE DETECTION PERFORMANCE ON A MIXED TESTING SET OF ALL SEVEN ALGORITHMS.

Resolution	SepMark [12]		KAD-Net [24]		WaveGuard [13]		LIDMark [30]		Ours	
	128	256	128	256	128	256	128	256	128	256
FSRT [31]	98.37%	97.15%	90.91%	93.66%	98.58%	98.66%	94.51%	95.13%	98.80%	99.01%
RAFSwap [32]	96.51%	96.98%	90.68%	95.37%	98.83%	98.78%	95.24%	93.66%	99.72%	99.52%
SimSwap [1]	92.82%	95.50%	96.29%	96.39%	98.47%	98.67%	93.09%	94.32%	98.59%	99.18%
CSCS [33]	98.79%	99.34%	92.16%	92.73%	98.74%	98.19%	94.82%	96.72%	98.76%	98.03%
HifiFace [34]	98.40%	98.70%	95.93%	95.13%	98.81%	98.32%	93.44%	95.41%	99.99%	99.99%
UniFace [35]	98.40%	98.70%	92.93%	93.13%	98.79%	98.91%	95.07%	94.06%	98.99%	99.27%
GANimation [36]	99.40%	99.70%	98.93%	98.13%	99.25%	99.54%	97.63%	96.82%	98.99%	99.10%
InfoSwap [37]	87.14%	86.30%	87.41%	85.34%	95.65%	95.51%	93.38%	93.41%	99.96%	99.32%
StyleMask [2]	85.41%	84.23%	84.87%	82.73%	90.74%	89.64%	92.59%	93.31%	98.62%	98.98%
StarGAN (Male) [38]	98.96%	98.27%	97.87%	97.87%	98.65%	99.73%	93.42%	93.56%	99.87%	99.02%
StarGAN (Young) [38]	98.28%	98.64%	97.15%	97.34%	98.07%	98.75%	93.51%	93.42%	99.81%	99.27%
Mixed	95.68%	95.77%	93.19%	93.44%	97.69%	97.79%	94.55%	95.04%	99.46%	99.24%

TABLE VI

CROSS-DATASET EVALUATION ON LFW, FACEFORENSICS++, DFDC, AND CELEB-DF v2. PSNR/SSIM, BER, AND AUC ARE USED TO EVALUATE VISUAL QUALITY, ROBUSTNESS, AND DETECTION PERFORMANCE, RESPECTIVELY. Δ_{BER} AND Δ_{AUC} DENOTE DIFFERENCES FROM CELEBA-HQ.

128 × 128				
Evaluation Metric	LFW	FF++	DFDC	Celeb-DF v2
PSNR ↑	46.2405	46.3081	46.0925	47.0597
SSIM ↑	0.9948	0.9921	0.9909	0.9917
Avg. Common BER ↓	0.0024%	0.0032%	0.0017%	0.0029%
Δ_{BER}	↑0.0004%	↑0.0012%	↓0.0003%	↑0.0009%
Avg. Malicious BER ↓	0.0118%	0.0208%	0.0138%	0.0196%
Δ_{BER}	↓0.0033%	↑0.0057%	↓0.0013%	↑0.0045%
Avg. Malicious AUC ↑	98.68%	99.25%	98.47%	99.59%
Δ_{AUC}	↓0.78%	↓0.21%	↓0.99%	↑0.13%
256 × 256				
Evaluation Metric	LFW	FF++	DFDC	Celeb-DF v2
PSNR ↑	43.7811	43.4107	43.3725	44.0254
SSIM ↑	0.9908	0.9895	0.9897	0.9894
Avg. Common BER ↓	0.0018%	0.0017%	0.0010%	0.0009%
Δ_{BER}	↑0.0012%	↑0.0011%	↑0.0004%	↑0.0003%
Avg. Malicious BER ↓	0.0061%	0.0058%	0.0095%	0.0102%
Δ_{BER}	↓0.0012%	↓0.0015%	↑0.0022%	↑0.0029%
Avg. Malicious AUC ↑	98.68%	98.74%	99.15%	98.87%
Δ_{AUC}	↓0.56%	↓0.50%	↓0.09%	↓0.37%

and III reports the average regional BER across seven facial semantic regions to evaluate the robustness of watermark recovery in local areas. The results show that the framework maintains consistently low BER across all seven regions, indicating that the proposed region-aware watermark embedding mechanism enables reliable watermark recovery under localized tampering and region-specific deepfake attacks.

To further clarify the motivation behind the proposed embedding strategy, we compare the facial-region-only embedding scheme with the proposed global embedding combined with regional redundancy enhancement. As shown in Table IV, facial-region-only embedding over-concentrates watermark signals in facial regions, making watermark recovery more

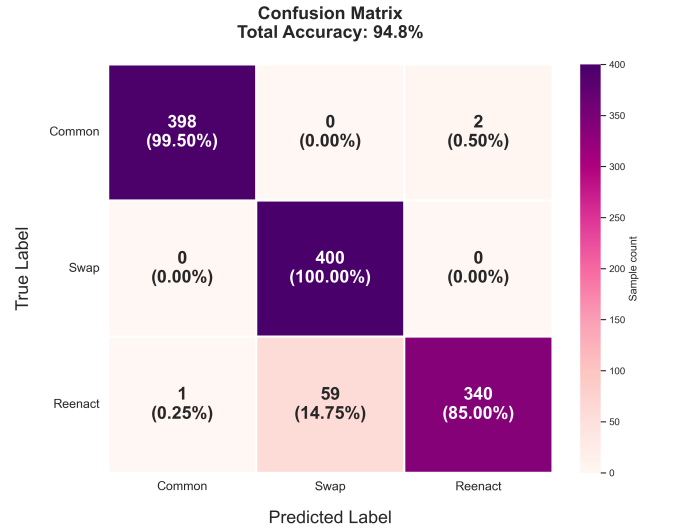


Fig. 6. Confusion Matrix of the Forgery Classification Model

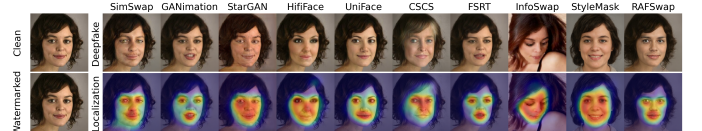


Fig. 7. Visualization of detection results for deepfake manipulations based on regional BER. The watermark BER is computed for each facial semantic region. Regions with high BER are visualized with stronger warm colors, indicating areas significantly affected by malicious manipulation.

vulnerable to localized tampering and spatial perturbations; in contrast, our method embeds the watermark over the whole image and further enhances key facial regions, thereby achieving more stable watermark recovery.

In addition, Table VI reports the cross-dataset robustness results on LFW, FaceForensics++, DFDC, and Celeb-DF v2. For video datasets, frames are sampled and evaluated at the image level. Δ_{BER} denotes the BER difference compared

TABLE VII
ABLATION STUDY ON REGIONAL REDUNDANCY AND SC-GNN AT
128 × 128 RESOLUTION.

Evaluation Metric	Regional Redundancy		SC-GNN	
	w/o	w/	w/o	w/
Avg. Common BER ↓	0.0025%	0.0020%	0.0024%	0.0020%
Avg. Malicious BER ↓	0.0249%	0.0151%	0.0184%	0.0151%
PSNR ↑	47.0412	46.3421	44.3084	46.3421
SSIM ↑	0.9956	0.9948	0.9722	0.9948
Avg. AUC ↑	99.24%	99.46%	99.47%	99.46%

TABLE IX
COMPUTATIONAL COMPLEXITY COMPARISON OF PROACTIVE DEEPPAKE
DETECTION METHODS.

Model	128 × 128		256 × 256	
	Params (M)	FLOPs (G)	Params (M)	FLOPs (G)
SepMark [12]	28.310	4.031	28.310	16.059
EditGuard [19]	8.138	2.996	18.551	11.992
DiffMark [23]	16.941	62.919	75.721	337.774
KAD-Net [24]	8.542	7.835	48.542	31.357
WaveGuard [13]	5.085	27.026	15.475	27.032
LIDMark [30]	5.518	12.221	21.023	60.622
Ours	3.426	53.797	11.755	134.452

with CelebA-HQ, where the up and down arrows indicate an increase and a decrease, respectively. As shown in Table VI, the proposed method maintains low BER under both common distortions and malicious deepfake manipulations across different datasets. The small BER variations demonstrate that the proposed region-aware watermarking framework has strong cross-dataset robustness and can stably recover watermarks under diverse data distributions. Based on these results, the framework enhances the stability of watermark recovery under complex deepfake conditions by strengthening watermark embedding in key facial semantic regions. Overall, the method exhibits reliable robustness against both common noise and deepfake attacks, providing a solid foundation for subsequent deepfake detection.

D. Detection Performance

In this section, we evaluate the proposed detection framework through experiments conducted on the CelebA-HQ dataset and in a cross-dataset setting. The framework utilizes an identity-structure dual verification mechanism to detect deepfake images and distinguish between different types of forgeries, including face swapping(swap) and face reenactment(reenact). We use AUC (Area Under Curve) as the primary evaluation metric to assess the model's detection capability. According to the identity-structure dual verification strategy described in Section III, the detection score D is computed and used for classification. For swap type forgeries, the score D is directly set to 1, indicating an identity mismatch, and the image is classified as a swap forgery. For other types of noise, the score D is computed using formula 13. The score D for reenact is typically higher, while

TABLE VIII
ABLATION STUDY ON REGIONAL REDUNDANCY AND SC-GNN AT
256 × 256 RESOLUTION.

Evaluation Metric	Regional Redundancy		SC-GNN	
	w/o	w/	w/o	w/
Avg. Common BER ↓	0.0057%	0.0006%	0.0011%	0.0006%
Avg. Malicious BER ↓	0.0211%	0.0073%	0.0091%	0.0073%
PSNR ↑	44.9810	44.3901	42.0856	44.3901
SSIM ↑	0.9941	0.9911	0.9703	0.9911
Avg. AUC ↑	98.75%	99.24%	99.22%	99.24%

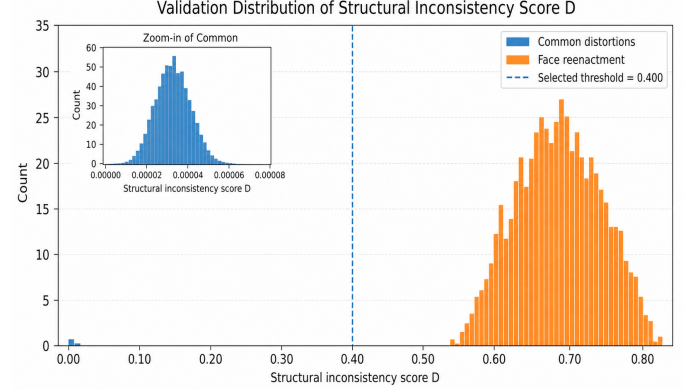


Fig. 8. Threshold verification for distinguishing common distortions from face reenactment using the structural inconsistency score D .

the score for real images is lower. To determine the threshold for distinguishing common distortions from face reenactment, we analyze the validation-set distribution of the structural inconsistency score D . Specifically, we randomly select 200 common-distortion samples and 200 face-reenactment samples, and compute their corresponding D scores. As shown in Fig.8, common distortions produce near-zero D values, while face reenactment samples are mainly distributed between 0.55 and 0.80, indicating clear separability between the two classes. Therefore, we set the threshold to 0.400, which lies in the low-density interval between the two distributions.

As shown in Table V, the proposed method achieves nearly 99% AUC across different image resolutions, demonstrating the robustness of the framework in detecting deepfake images under varying conditions. Regardless of the image resolution, the performance remains high, indicating that the framework can effectively handle variations in image quality. Additionally, we conduct cross-dataset experiments to evaluate the generalization capability of the proposed detection mechanism. As shown in Table VI, the proposed method maintains high Avg. Malicious AUC on the other datasets, with only slight performance variations compared with CelebA-HQ, demonstrating its strong cross-dataset generalization ability under different data distributions.

Additionally, to verify the forgery model classification ability of our framework, we evaluated its classification performance on three types of forged images: common distortions(common), face swapping(swap), and face reenact-

ment(reenact). The CelebA-HQ test set was used, with 400 samples for each image type to maintain a balanced dataset. The score value D is computed based on the formula 13, and the classification is performed based on this value. For images where decryption fails, they are directly classified as swap forgery. For images where the score value D exceeds the set threshold D_{thresh} , they are classified as reenact forgery. If the score value D is smaller than D_{thresh} , the image is classified as common (real image). This design ensures that the framework can effectively distinguish between different types of forgery images. According to the confusion matrix shown in Fig. 6, the proposed framework achieves an overall accuracy of 94.8% on the test set. Most samples in the common and swap categories are correctly classified, while the majority of misclassifications occur between Reenact and Swap. This is mainly because some reenactment forgeries exhibit relatively low synthesis quality, making their visual characteristics closer to swap forgeries, which leads to misclassification. Overall, the framework can effectively distinguish different types of forged images.

E. Region-wise Contribution Analysis

To quantify the contribution of different facial regions, we conduct a region-wise AUC analysis at the 256×256 resolution. In our framework, face swapping is mainly detected by identity decryption consistency verification, while the region-aware watermarking module is mainly used to distinguish common distortions from face reenactment manipulations. Specifically, based on the detection score in Eq. 13, we first compute a global-only baseline using only the global BER term. Then, we keep the global BER term unchanged and combine it with the BER of each individual facial region using the original weights. The resulting region-specific score is used to compute the AUC for distinguishing common distortions from face reenactment manipulations. The improvement over the global-only baseline reflects the marginal contribution of the corresponding regional BER. As shown in Fig. 9, the mouth region achieves the highest AUC, followed by the eyebrow and eye regions. This is consistent with the fact that face reenactment mainly affects expression-related and motion-sensitive facial regions. In contrast, the nose and face-skin regions are relatively stable under expression-driven manipulations and therefore show smaller improvements. These results demonstrate that different facial regions contribute differently to reenactment detection and further verify the effectiveness of the proposed region-aware watermarking design.

F. Visualization of Detection Results

As an auxiliary function of the detection process, Region-aware watermarking can provide interpretability for detection results. During detection, when identity consistency is satisfied, the system computes the BER of the structural watermark for each facial semantic region and maps regions with high BER back to the corresponding semantic masks for localization. In the visualization results, different semantic regions are highlighted using a color-coded scheme according to their BER values, where regions with higher errors indicate

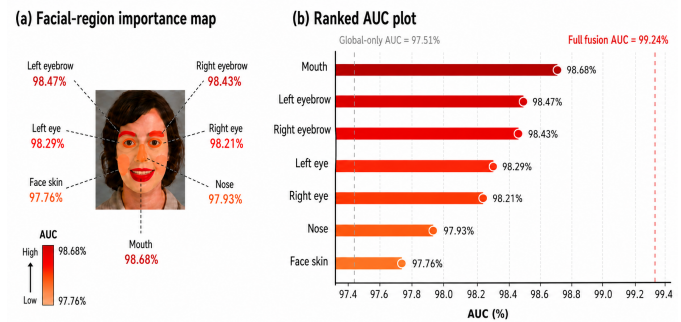


Fig. 9. Region-wise contribution analysis at the 256×256 resolution. (a) Facial-region importance map, where darker colors indicate higher AUC values. (b) Ranked AUC results of different facial regions.

locations with degraded structural consistency. As shown in Fig. 5, under common noise conditions such as JPEG compression, resizing, blurring, and additive noise, the structural watermarks of all semantic regions can be stably recovered, and no obvious high-error regions appear in the localization masks. This demonstrates that the proposed method is robust to benign distortions and does not introduce false localization. In contrast, as shown in Fig. 7, under malicious manipulation scenarios such as face swapping and face reenactment, the semantic regions corresponding to actual tampering exhibit significantly higher structural watermark error rates and are clearly highlighted in the localization masks, thereby effectively indicating the tampered areas.

G. Complexity Analysis

To further evaluate the computational cost of the proposed framework, we report the number of parameters and FLOPs of different proactive forensics methods in Table IX. Params denotes the number of learnable parameters, and FLOPs denotes the computational cost required for single-image inference under the corresponding input resolution. As shown in Table IX, Segfacemark achieves the smallest number of parameters at both 128×128 and 256×256 resolutions, indicating the high parameter efficiency of the proposed encoder-decoder architecture. Meanwhile, the FLOPs of Segfacemark remain within a comparable range to existing proactive forensic methods. The computational cost mainly comes from region-aware watermark embedding and global-region joint watermark recovery. Unlike methods that only rely on global watermark verification, our method recovers both the global watermark and region-level watermarks, which are further used for identity-structure dual verification, robust source tracing, forgery-type classification, and region-level interpretability. Overall, Segfacemark achieves a favorable balance among model size, computational complexity, robustness, and detection performance.

H. Ablation Study

To further validate the effectiveness of the key components in Segfacemark, we conduct ablation studies on regional redundant embedding and SC-GNN. In each experiment, only

the target component is removed while all other settings remain unchanged, where $w/$ denotes the full model and w/o denotes the model without the corresponding component. As shown in Tables VII and VIII, removing regional redundant embedding leads to a clear increase in the average malicious BER and a decrease in Average AUC at both resolutions. This indicates that regional redundant embedding enhances watermark recovery in key facial semantic regions, thereby improving robustness and detection reliability under malicious manipulations. In contrast, SC-GNN mainly contributes to the visual fidelity of watermarked images. After removing SC-GNN, PSNR and SSIM decrease significantly at both resolutions, while Average AUC remains nearly unchanged. This demonstrates that SC-GNN effectively preserves facial structural information through structural consistency constraints, improving watermark imperceptibility without degrading detection performance. Overall, regional redundant embedding and SC-GNN improve the overall performance of Segface-mark from the perspectives of robustness and visual fidelity, respectively.

V. CONCLUSION

This paper propose a region-aware watermarking based proactive deepfake detection framework named Segfacemark. By modeling the tampering sensitivity of facial regions and strengthening the protection of key areas, the robustness of the watermark is enhanced. The framework adopts dynamic encryption mechanism to realize the watermark generation of “one-identity-one-key” and enhance the security. In the detection stage, through the double verification mechanism, the accuracy is improved, the forgery type identification is realized, and the interpretability of the detection results is enhanced. However, the current method only provides interpretability for detecting tampered regions, without fine-grained localization. Future work will focus on improving tampered region localization and enhancing detection granularity to increase detection accuracy and applicability.

REFERENCES

- [1] R. Chen, X. Chen, B. Ni, and Y. Ge, “Simswap: An efficient framework for high fidelity face swapping,” in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2003–2011.
- [2] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, “Stylemask: Disentangling the style space of stylegan2 for neural face reenactment,” in *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)*. IEEE, 2023, pp. 1–8.
- [3] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, “Deepfake detection: A comprehensive survey from the reliability perspective,” *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–35, 2024.
- [4] M. Zou, B. Yu, Y. Zhan, S. Lyu, and K. Ma, “Semantics-oriented multitask learning for deepfake detection: A joint embedding approach,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 10, pp. 9950–9963, 2025.
- [5] J. Li, L. Yu, R. Liu, and H. Xie, “A detail-aware transformer to generalizable face forgery detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 4, pp. 3262–3275, 2024.
- [6] Z. Guo, L. Wang, W. Yang, G. Yang, and K. Li, “Ldfnet: Lightweight dynamic fusion network for face forgery detection by integrating local artifacts and global texture information,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1255–1265, 2023.
- [7] S. Han, S. Li, S. Wang, L. Yuan, Y. Zhang, and X. Gao, “Deepfake detection leveraging self-blended artifacts guided by facial embedding discrepancy,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- [8] H. Zheng, Y. Li, L. Wang, Y. Diao, and Z. Guo, “Boosting active defense persistence: A two-stage defense framework combining interruption and poisoning against deepfake,” *IEEE Transactions on Information Forensics and Security*, vol. 21, pp. 3080–3092, 2026.
- [9] P. Neekhar, S. Hussain, X. Zhang, K. Huang, J. McAuley, and F. Koushanfar, “Facesigns: Semi-fragile watermarks for media authentication,” *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.
- [10] Y. Yang, C. Liang, H. He, X. Cao, and N. Z. Gong, “Faceguard: Proactive deepfake detection,” *arXiv preprint arXiv:2109.05673*, 2021.
- [11] T. Wang, M. Huang, H. Cheng, X. Zhang, and Z. Shen, “Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10515–10524.
- [12] 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*, 2023, pp. 1190–1201.
- [13] Z. He, Z. Guo, L. Wang, G. Yang, Y. Diao, and D. Ma, “Waveguard: Robust deepfake detection and source tracing via dual-tree complex wavelet and graph neural networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.
- [14] T. Wang, M. Huang, H. Cheng, B. Ma, and Y. Wang, “Robust identity perceptual watermark against deepfake face swapping,” *arXiv preprint arXiv:2311.01357*, 2023.
- [15] R. Wang, F. Juefei-Xu, M. Luo, Y. Liu, and L. Wang, “Faketagger: Robust safeguards against deepfake dissemination via provenance tracking,” in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3546–3555.
- [16] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, “Artificial fingerprinting for generative models: Rooting deepfake attribution in training data,” in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 14448–14457.
- [17] Y. Zhao, B. Liu, M. Ding, B. Liu, T. Zhu, and X. Yu, “Proactive deepfake defence via identity watermarking,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 4602–4611.
- [18] T. Wang, S. Niu, H. Cheng, X. Zhang, and Y. Wang, “Nullswap: Proactive identity cloaking against deepfake face swapping,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 9945–9954.
- [19] X. Zhang, R. Li, J. Yu, Y. Xu, W. Li, and J. Zhang, “Editguard: Versatile image watermarking for tamper localization and copyright protection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11964–11974.
- [20] Y. Zhang, Z. Liu, Z. Wang, H. Chen, and X. Yu, “Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. —.
- [21] Y. Liu, K. Guo, Y. Gao, Y. Li, C. Lin, and J. Li, “Waterlo: Watermarking for localized deepfake detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 1881–1890.
- [22] T. Wang, H. Cheng, M.-H. Liu, and M. Kankanhalli, “Fractalforensics: Proactive deepfake detection and localization via fractal watermarks,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 7210–7219.
- [23] C. Sun, H. Sun, Z. Guo, Y. Diao, L. Wang, D. Ma, G. Yang, and K. Li, “Diffmark: Diffusion-based robust watermark against deepfakes,” *Information Fusion*, 2025.
- [24] S. He, Y. Diao, Y. Li, C. Sun, L. Wang, and Z. Guo, “Kad-net: Kolmogorov-arnold and differential-aware networks for robust and sensitive proactive deepfake forensics,” *Knowledge-Based Systems*, 2025.
- [25] D. E. King, “Dlib-ml: A machine learning toolkit,” *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [26] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
- [27] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
- [28] K. Narayan, V. Vs, and V. M. Patel, “Segface: Face segmentation of long-tail classes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6182–6190.
- [29] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, “Simam: A simple, parameter-free attention module for convolutional neural networks,” in *Internat-*

tional conference on machine learning. PMLR, 2021, pp. 11 863–11 874.

- [30] J. Wu, L. Wang, and Z. Guo, “All in one: Unifying deepfake detection, tampering localization, and source tracing with a robust landmark-identity watermark,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.
- [31] A. Rochow, M. Schwarz, and S. Behnke, “Fsr: Facial scene representation transformer for face reenactment from factorized appearance head-pose and facial expression features,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 7716–7726.
- [32] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, and Y. Liu, “Region-aware face swapping,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7632–7641.
- [33] Z. Huang, F. Tang, Y. Zhang, J. Cao, C. Li, S. Tang, J. Li, and T.-Y. Lee, “Identity-preserving face swapping via dual surrogate generative models,” *ACM Transactions on Graphics*, vol. 43, no. 5, pp. 1–19, 2024.
- [34] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, “Hififace: 3d shape and semantic prior guided high fidelity face swapping,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization*, 2021, pp. 1136–1142.
- [35] C. Xu, J. Zhang, Y. Han, G. Tian, X. Zeng, Y. Tai, Y. Wang, C. Wang, and Y. Liu, “Designing one unified framework for high-fidelity face reenactment and swapping,” in *European conference on computer vision*. Springer, 2022, pp. 54–71.
- [36] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Ganimation: Anatomically-aware facial animation from a single image,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 818–833.
- [37] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, “Information bottleneck disentanglement for identity swapping,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3404–3413.
- [38] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “Stargan v2: Diverse image synthesis for multiple domains,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8188–8197.
- [39] T. Karras, “Progressive growing of gans for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.
- [40] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5549–5558.
- [41] G. B. Huang, M. Mattar, T. 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.
- [42] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
- [43] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge (dfdc) dataset,” *arXiv preprint arXiv:2006.07397*, 2020.
- [44] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.



Ziyuan He received the B.S. degree from the School of Information Technology, Beijing Normal University, Zhuhai, China, in 2023. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Xinjiang University, Urumqi, China. His research interests include deepfake detection and image processing.



Zhiqing Guo (Member, IEEE) received the Ph.D. degree in Computer Science and Technology from Hunan University in 2023. He is currently an Associate Professor at the School of Computer Science and Technology, Xinjiang University, Urumqi, China. At present, he is the PI of several projects, such as the Natural Science Foundation of China (NSFC), with a total fund of more than 2 million yuan. His research interests include multimedia forensics, image/video content security, visual scene understanding. He has published more than 40 papers in international journals or conferences such as IEEE Transactions on Information Forensics and Security (TIFS), IEEE Transactions on Multimedia (TMM), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), ACM Transactions on Multimedia Computing Communications and Applications (TOMM), IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), etc.



Liejun Wang (Member, IEEE) received his Ph.D. degree in the School of Information and Communication Engineering from Xi'an Jiaotong University in 2012. He is currently a full professor with the School of Information Science and Engineering, Xinjiang University. His research interests include wireless sensor networks, computer vision, and natural language processing. He has published research papers in international journals such as IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Cybernetics, IEEE Transactions on Vehicular Technology, etc.



Xiaoming Tao received a Ph.D. degree in information and communication systems from Tsinghua University, Beijing, China, in 2008. She is currently a Professor with the Department of Electronic Engineering at Tsinghua University and School of Computer Science and Technology at Xinjiang University. Prof. Tao served as a workshop General Co-Chair for IEEE INFOCOM 2015 and the volunteer leadership for IEEE ICIP 2017. She has been the Editor of the J. Communications and Information Networks and China Communication since 2016.

She is also the recipient of the National Science Foundation for Outstanding Youth and many national awards, for example, the 2017 China Young Women Scientists Award, 2017 Top Ten Outstanding Scientists and Technologists from China Institute of Electronics, 2017 First Prize of Wu Wen Jun AI Science and Technology Award, 2016 National Award for Technological Invention Progress, and 2015 Science and Technology Award of China Institute of Communications, and so on.



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