Revealing Photoshop Inpainting Traces Under JPEG Compressions

Yushu Zhang, Lu Zhang, Shuren Qi*, Xiangli Xiao, and Wenying Wen

***Abstract*—Photoshop inpainting has become one of the most challenging targets in image forensics, as its content-aware and patch-based editing mechanisms produce visually coherent manipulations with weak and localized forensic artifacts. This difficulty is further amplified by JPEG compression, which is routinely introduced during online transmission and tends to suppress the high-frequency tampering traces on which existing forensic detectors largely depend. As a result, current methods face an inherent trade-off: Photoshop-oriented detectors provide strong discriminability under clean conditions but lack robustness to compression, whereas compression-robust forensic methods often fail to capture subtle inpainting artifacts. To overcome this tension, this paper proposes a JPEG-resistant Photoshop inpainting localization method based on multi-frequency representation. The proposed framework employs a set of parameterized frequency-selective filters to extract complementary representations across multiple spectral bands. Each frequency branch is trained independently as a dedicated detector, and a fusion module integrates their outputs to generate a comprehensive localization map that balances discriminability and robustness. A theoretical analysis in the frequency domain is further provided to explain how low-frequency representations remain stable under JPEG-induced attenuation, supporting the design rationale of the proposed framework. In addition, a multi-quality JPEG augmentation strategy is adopted during training to mitigate the mismatch between training and testing compression levels. Extensive experiments on both script-created and hand-created Photoshop inpainting datasets demonstrate that the proposed method consistently outperforms representative forgery localization methods under various JPEG compression strengths. We further evaluate the method on images transmitted through Wechat, Weibo, and Twitter, confirming its effectiveness in practical online social network scenarios. These results demonstrate that the proposed multi-frequency representation strategy offers a principled and effective approach to robust image forensic analysis.**

***Index Terms*—Photoshop, inpainting, forensic, robustness, JPEG.**

This work was supported by the National Natural Science Foundation of China under Grant 62522112, the Ganpo Talent Program of Jiangxi Province under Grant gpyc20240012, the Outstanding Youth Fund Program of Jiangxi Province under Grant 20252BAC220008, the Jiangxi Key Research and Development Program under Grant 20261BCE310050, and the Young Talent Support Project of Guangzhou Association for Science and Technology under Grant QT2025-047.

Y. Zhang, X. Xiao, and W. Wen are with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330044, China, and Y. Zhang is also with the Jiangxi Provincial Key Laboratory of Multimedia Intelligent Processing, Nanchang 330044, China (e-mail: zhangyushu@jxufe.edu.cn; xiaoxiangli@jxufe.edu.cn; wenyingwen@sina.cn).

L. Zhang is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: lu_zhang@nuaa.edu.cn).

S. Qi is with the Department of Data Science, City University of Hong Kong, Hong Kong, China (e-mail: shurenqi@cityu.edu.hk).

*(*Corresponding author: Shuren Qi.)*

I. Introduction

IMAGE editing software such as Photoshop has become widely accessible due to its user-friendly interface and powerful image editing capabilities. Among various editing functions, inpainting is particularly attractive because it can remove or replace selected image regions while maintaining visual coherence with the surrounding content. While such techniques provide convenient tools for image restoration and creative editing, they also raise serious concerns in digital forensics, including evidence falsification, copyright infringement, and the malicious manipulation of visual information. Therefore, reliable localization of Photoshop inpainting forgeries has become an important task in image forensic analysis.

Photoshop inpainting differs from fully generative approaches such as GAN-based or diffusion-based image synthesis. Instead of generating an entire image or region purely from learned data priors, Photoshop inpainting is commonly performed through user-guided content-aware filling, patch-based synthesis, and local blending. This editing paradigm is highly representative of practical manual tampering, where forged regions are intentionally made structurally and texturally consistent with their surroundings. As a result, Photoshop-inpainted regions often exhibit weak visual artifacts and subtle statistical inconsistencies, making their localization more challenging than many generic manipulation cases. Studying Photoshop-based inpainting is therefore not merely a special case, but a practically important and technically challenging problem in image forensics.

At the same time, the rapid development of Online Social Networks (OSNs) has greatly facilitated image sharing and distribution, allowing manipulated images to spread quickly and widely. In practical scenarios, tampered images are rarely preserved in their original quality. Instead, they are commonly re-encoded, resized, and compressed during online transmission. Among these operations, JPEG compression is one of the most prevalent and influential degradations. It introduces quantization artifacts, suppresses high-frequency forensic traces, and changes the distribution of spectral energy across frequency bands. Consequently, Photoshop inpainting localization in real-world environments must address two coupled difficulties: the tampering traces are already subtle due to the refined editing mechanism of Photoshop, and these traces can be further weakened or distorted by JPEG compression during transmission.

A key challenge in this task is the trade-off between discriminability and robustness. Discriminability refers to the ability of a detector to distinguish tampered regions from

authentic regions by capturing subtle forensic traces. For Photoshop inpainting, these traces are often reflected in local texture inconsistencies, boundary discontinuities, and abnormal residual patterns, many of which are concentrated in relatively high-frequency components. Therefore, methods that emphasize high-frequency representations usually achieve stronger sensitivity to subtle manipulation artifacts. However, these components are also highly vulnerable to post-processing operations such as JPEG compression, because high-frequency DCT coefficients are more aggressively quantized and may be partially removed during compression.

Robustness, in contrast, refers to the ability of a detector to maintain stable performance when the input image undergoes degradation. Low-frequency components are generally more stable under compression and mild post-processing, and thus provide more robust representations. However, low-frequency features usually contain coarser structural information and may not be sufficiently discriminative for localizing subtle Photoshop inpainting traces. Therefore, a detector relying mainly on high-frequency features may be discriminative but fragile, whereas a detector relying mainly on low-frequency features may be robust but insensitive to weak tampering artifacts.

This frequency-domain perspective reveals an inherent limitation of existing forgery localization methods. General-purpose forensic detectors can localize various types of manipulations, but they are not specifically optimized for the weak traces left by Photoshop inpainting. Photoshop-oriented detectors improve discriminability for such tampering, but they often assume relatively clean testing conditions and suffer from performance degradation after JPEG compression. Conversely, compression-resistant methods improve robustness to post-processing, but they are usually designed for generic manipulations and may fail to capture the highly localized artifacts produced by refined manual editing. As illustrated in Fig. 1, most existing methods implicitly rely on a limited frequency range or a single representation preference, making it difficult to simultaneously achieve strong discriminability and robustness.

To address this problem, this paper proposes a JPEG-resistant Photoshop inpainting localization method based on Multi-Frequency Representation (MFR). The main idea is to explicitly extract and exploit complementary forensic cues from multiple frequency bands. By adjusting the parameters of the proposed frequency representation filters, multiple branches are constructed to capture tampering-related features at different spectral levels. Each branch is trained independently so that it can specialize in the forensic evidence available within its corresponding frequency range. A fusion module is then designed to integrate the localization outputs of all branches, producing a final prediction that balances high-frequency discriminability and low-frequency robustness.

In addition, to reduce the mismatch between training and testing compression levels, we construct a multi-quality JPEG training set. Specifically, 50,000 Photoshop-inpainted images are further compressed with multiple JPEG quality factors, resulting in 250,000 training samples. This strategy exposes the detector to different compression strengths and enables it to learn more stable tampering representations under JPEG degradation. The proposed method is evaluated not only on simulated JPEG-compressed datasets, but also on real-world OSN-transmitted images collected from Wechat, Weibo, and Twitter.

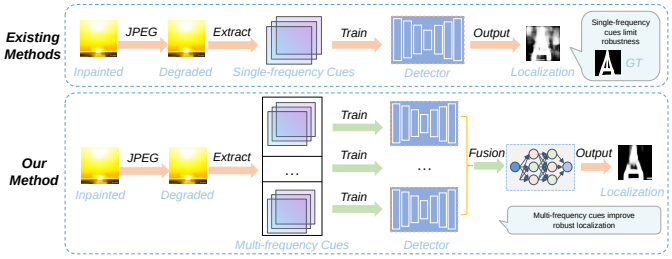

Fig. 1. Existing methods mainly rely on single-frequency cues after image degradation, whereas our method exploits complementary multi-frequency cues for robust localization.

The main contributions of this paper are as follows:

- We propose a frequency-domain perspective for JPEG-resistant Photoshop inpainting localization. By analyzing the complementary roles of high- and low-frequency representations, we formulate the localization problem as a balance between discriminability and robustness under JPEG compression. To the best of our knowledge, this is among the first attempts to explicitly exploit multi-frequency priors for this task.
- We introduce a Multi-Frequency Representation framework for Photoshop inpainting localization. The framework consists of frequency-branch construction, independent branch learning, and multi-branch fusion. This design allows the detector to capture manipulation-sensitive high-frequency cues while preserving compression-resistant low-frequency information, thereby improving localization performance under different JPEG compression strengths.
- We construct a multi-quality JPEG training strategy and evaluate the proposed method under both simulated and real-world transmission conditions. By manually uploading and downloading tampered images through Wechat, Weibo, and Twitter, we further validate the transferability and practical robustness of the proposed method in OSN scenarios.

The remainder of this paper is organized as follows. Section II reviews related work on image forgery localization, Photoshop inpainting detection, and JPEG compression-resistant forensic methods. Section III details the proposed method. Section IV presents the experimental results. Finally, Section V concludes the paper and discusses future work.

II. Related Work

A. Image Forgery Localization and Photoshop Inpainting Detection

Image forgery localization aims to identify tampered regions at the pixel level. Traditional methods mainly rely on handcrafted forensic cues left by specific image acquisition or manipulation processes. For instance, Bianchi et al. [1] detected

tampered regions by analyzing double JPEG compression artifacts in Discrete Cosine Transform (DCT) blocks. Ferrara et al. [2] exploited inconsistencies caused by Color Filter Array (CFA) interpolation for fine-grained forgery localization. Chierchia et al. [3] introduced a Bayesian-Markov Random Field framework to localize forgeries based on Photo Response Non-Uniformity (PRNU) artifacts. Li et al. [4] investigated diffusion-based inpainting detection by analyzing gradient inconsistencies in Laplacian-domain features. Although these handcrafted methods are interpretable and effective under certain assumptions, they usually target specific artifacts and suffer from limited generalization when the manipulation type or post-processing condition changes.

With the development of deep learning, many CNN-based and learning-based forensic methods have been proposed to improve general forgery localization performance. Wu et al. [5] designed ManTra-Net, an end-to-end fully convolutional network that treats image manipulation localization as a local anomaly detection problem. Zhou et al. [6] employed a two-stream Faster R-CNN framework to jointly learn visual tampering artifacts and noise features. Zhuo et al. [7] introduced self-adversarial training to improve attention to tampered regions. Dong et al. [8] proposed a multi-view multi-scale supervised network that captures semantic-agnostic forensic traces from noise and boundary artifacts. Li et al. [9] proposed an edge-aware regional message passing controller to enhance boundary-aware image forgery localization by modeling interactions between manipulated regions and edge information. Several recent studies [10], [11], [12] further adopted contrastive learning to enlarge the feature discrepancy between tampered and authentic regions. Li et al. [13] proposed CLIP-IFDL, a CLIP-based image forgery detection and localization framework that adapts vision-language pretraining to forensic tasks through noise-assisted prompt learning. These methods have achieved promising results on generic manipulation datasets, but their performance can degrade when facing highly realistic and weakly traceable Photoshop inpainting.

Photoshop is one of the most widely used professional image editing tools, and its tampering operations are often designed to conceal visual inconsistencies. Zhuang et al. [14] investigated commonly used Photoshop editing tools and developed detectors for the artifacts introduced by these operations. However, this method does not specifically address Photoshop inpainting. Modern Photoshop provides several powerful inpainting tools, including "Content-Aware Fill" [15], "Content-Aware Patch", and "Content-Aware Move" [16], which can synthesize visually coherent content by sampling and blending surrounding patches. Although some methods [4], [17], [18], [19] have been proposed for inpainting localization, their performance decreases when applied to Photoshop inpainting, where manipulation traces are weaker and more carefully concealed. To address this issue, Zhang et al. [20] proposed PS-Net, a specialized network for localizing Photoshop-inpainted regions.

Although Photoshop-oriented methods improve discriminability for inpainting traces, they are typically developed and evaluated under relatively ideal conditions. In practical scenarios, tampered images are often compressed or transmitted through OSNs before being analyzed. Since Photoshop inpainting traces are already subtle, compression-induced attenuation can further weaken the residual evidence used for localization. Therefore, a detector designed only for clean Photoshop inpainting may not generalize well to JPEG-compressed or OSN-transmitted images. This limitation motivates the need for a method that explicitly considers both Photoshop-specific discriminability and compression-resistant robustness.

B. JPEG Compression-Resistant Forgery Detection

In real-world image dissemination, JPEG compression is one of the most common post-processing operations. It is widely used by cameras, image storage systems, and OSN platforms due to its high compression efficiency. The JPEG pipeline applies block-wise DCT followed by quantization, where high-frequency coefficients are typically assigned larger quantization steps. As a result, fine-grained details and weak forensic traces are more likely to be attenuated or removed. This property is particularly harmful for image forgery localization, because many manipulation artifacts are embedded in high-frequency residuals, boundary inconsistencies, or local texture statistics. When Photoshop-inpainted images are further compressed, the remaining forensic evidence can become extremely weak, making reliable localization more difficult.

To improve robustness against compression and post-processing, several methods have been proposed. Wang et al. [21] introduced a wavelet compression representation learning scheme with contrastive learning to distinguish different compression levels. Rao et al. [22] proposed a self-supervised domain adaptation network that approximates JPEG compression to learn more generalized forensic representations. Wu et al. [23] modeled black-box noise introduced by OSNs and injected such noise into training images to improve robustness in social media scenarios. Zhuang et al. [24] designed a restoration module to recover tampering traces in post-processed images before localization. Shan et al. [25] simulated information loss caused by tampering-induced rescaling and restored high-frequency forensic traces for improved detection.

These studies demonstrate that explicitly considering degradation is essential for practical image forensics. However, most compression-resistant methods are designed for generic manipulation types or broad OSN robustness, rather than Photoshop-specific inpainting. Their representations often prioritize stability under degradation, but this stability may reduce sensitivity to the weak and localized artifacts left by refined manual editing. In other words, robustness-oriented methods can maintain consistent responses after compression, but they may not provide sufficient discriminative power for Photoshop inpainting localization. This creates a complementary limitation to Photoshop-oriented detectors: the former are robust but less sensitive to subtle inpainting traces, whereas the latter are discriminative but fragile under compression.

Different from existing methods, our work explicitly addresses both sides of this problem through multi-frequency representation. Instead of relying on a single spectral preference, the proposed method extracts features from multiple frequency bands and trains independent detectors for different branches. Low-frequency branches provide more stable

responses under JPEG compression, while high-frequency branches preserve sensitivity to subtle manipulation artifacts. The final fusion module integrates these complementary predictions to produce a localization map that maintains both robustness and discriminability. This design makes the proposed method particularly suitable for Photoshop inpainting localization under JPEG compression and OSN transmission, where both weak tampering traces and compression-induced degradation must be handled simultaneously.

III. Method

In this section, we present the proposed JPEG-resistant Photoshop inpainting localization method based on Multi-Frequency Representation (MFR). The overall design is motivated by the frequency-domain trade-off discussed in Section I: high-frequency representations provide stronger discriminability for subtle Photoshop inpainting traces, whereas low-frequency representations are generally more stable under JPEG compression. Therefore, instead of relying on a single representation, the proposed method explicitly constructs multiple frequency-aware branches and integrates their predictions to balance discriminability and robustness.

The proposed framework consists of three main components. First, a set of Multi-Frequency Representation Filters is used to extract tampering-related representations from different spectral bands. Second, a Photoshop inpainting detector is employed as the branch-wise localization network, where each branch learns the forensic cues available at a specific frequency level. Third, a fusion module combines the outputs of all branches to generate the final localization result. In addition, a multi-quality JPEG training strategy is adopted to reduce the mismatch between training and testing compression levels. The details of these components are described in the following subsections.

A. Multi-Frequency Representation Filters

Tampering localization requires representations that are both discriminative and robust. For Photoshop inpainting, discriminative cues are often associated with subtle residual artifacts, local texture inconsistencies, and boundary-level statistical variations, which are more evident in relatively high-frequency components. However, such components are also more vulnerable to JPEG compression and other post-processing operations. In contrast, low-frequency components are more stable under degradation but may be less sensitive to weak manipulation traces. To explicitly model this trade-off, we introduce a tunable filter that extracts representations from different frequency bands by adjusting its parameters.

The proposed tunable filter $C^w_{n,m}$ contains two categories of parameters: the scale-related parameter w and the frequency-order parameters (n, m). These parameters jointly determine the spectral characteristics of the extracted representation. Specifically, w controls the effective support size of the basis function in the spatial domain and thus affects the amount of image information involved in the representation. A larger w generally preserves richer structural details and increases representation capacity, whereas a smaller w yields more compact responses that are relatively stable under degradation. In contrast, (n, m) determines the frequency order of the filter and governs the corresponding response band in the spectral domain. Lower-order configurations mainly emphasize low-frequency components, while higher-order configurations capture finer high-frequency details.

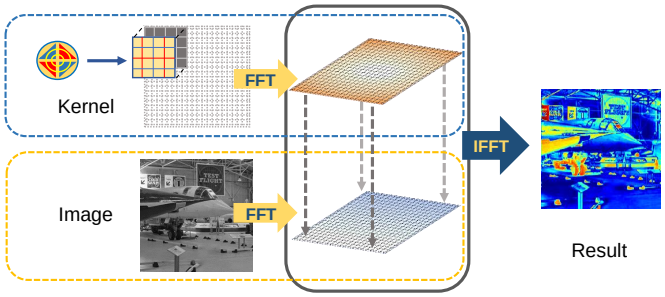

Fig. 2. Architecture of the multi-frequency representation filter.

From the perspective of image forensics, these parameters provide a controllable way to navigate the trade-off between robustness and discriminability. Low-frequency representations are more resistant to JPEG-induced attenuation, whereas high-frequency representations preserve more manipulation-sensitive details. Therefore, the parameter selection of $C^w_{n,m}$ can be regarded as a tunable frequency-basis design problem, where multiple representative configurations are required rather than a single fixed setting.

To extract image representations across multiple frequencies, we introduce a parameter-adjustable filter kernel [26]. The complete workflow is shown in Fig. 2. Given an input tampered image, the image is first transformed into the frequency domain using the Fast Fourier Transform (FFT). The resulting frequency representation is then multiplied element-wise with the Fourier transform of the filter kernel. Finally, the filtered result is converted back to the spatial domain using the Inverse Fast Fourier Transform (IFFT). This process can be written as

$$I' = F^{-1}\left(F(I) \odot F\left((C^w_{nm})^T\right)\right), \tag{1}$$

where I denotes the input tampered image, F and F^{-1} represent the FFT and IFFT, respectively, I' denotes the filtered output, and $\odot$ indicates element-wise multiplication. The term $(C^w_{nm})^T$ denotes the parameter-adjustable filter kernel, where $w \in \mathbb{Z}$ controls the representation scale, and $n, m \in \{0, 1, 2, \ldots, w\}$ determine the frequency order.

Lower-order (n, m) configurations mainly respond to low-frequency components and are therefore more stable under lossy operations such as noise, blur, and JPEG compression. This is because these operations tend to suppress or distort high-frequency image information more strongly [27]. To provide an intuitive illustration, Fig. 3 shows heatmaps of filtered images under different w and (n, m) settings. When (n, m) is fixed, increasing w preserves richer image structures. When w is fixed, increasing (n, m) produces responses with more fine-grained high-frequency details.

In practical applications, tampered images are often degraded by JPEG compression and other unknown post-

processing operations. The degraded image can be generally expressed as

$$I_d = f(I) + \eta, \tag{2}$$

where I is the tampered image before additional compression, f represents the JPEG compression operation, and η denotes bounded perturbations introduced by other unknown processing factors. Ideally, $\eta = 0$, but in real-world scenarios such as OSN transmission, η accounts for additional platform-dependent distortions.

The robustness of the proposed filter can be analyzed from a frequency-domain perspective. JPEG compression is based on block-wise Discrete Cosine Transform (DCT), followed by quantization of transform coefficients. Since high-frequency coefficients are typically assigned larger quantization steps, they are more likely to be rounded toward zero, resulting in stronger attenuation of fine-grained image details [28]. Although JPEG compression is not a strictly linear low-pass filtering operation due to quantization and entropy coding, its dominant statistical effect can be interpreted as a low-pass-like frequency-selective attenuation process. In other words, JPEG compression tends to preserve low-frequency components while suppressing high-frequency components that often contain weak forensic traces.

Based on this observation, we approximate the dominant frequency attenuation effect of JPEG compression using an implicit frequency-selective operator. This approximation does not explicitly model blocking artifacts or nonlinear quantization errors. Nevertheless, it provides a useful analytical framework for explaining why low-frequency representations are more stable under JPEG compression and why multi-frequency representation is beneficial for robust localization.

Definition 1 (Degraded Image). *For analytical convenience, the dominant frequency attenuation effect of JPEG compression is approximated by an implicit frequency-selective operator. Under this approximation, the degraded image can be written as*

$$I_d = p \otimes I + \eta, \tag{3}$$

where p denotes the implicit frequency-selective attenuation operator, $\otimes$ denotes convolution, and η represents bounded perturbations introduced by other unknown processing operations.

Definition 2 (Filtered Result of the Degraded Image). *The filtered result of the degraded image is obtained by substituting Eq. (3) into Eq. (1), yielding*

$$I'_d = F^{-1}\left(F(p \otimes I) \odot F((C^w_{nm})^T) + F(\eta) \odot F((C^w_{nm})^T)\right), \tag{4}$$

where I'_d represents the filtered result of the degraded image.

According to the convolution theorem, Eq. (4) can be rewritten as

$$\begin{aligned} I'_d = & F^{-1}\left(F(p) \odot F(I) \odot F((C^w_{nm})^T)\right. \\ & \left. + F(\eta) \odot F((C^w_{nm})^T)\right). \end{aligned} \tag{5}$$

When the values of (n, m) are small, the filter primarily responds to low-frequency components. In this case, $F((C^w_{nm})^T)$ is concentrated in a limited low-frequency region

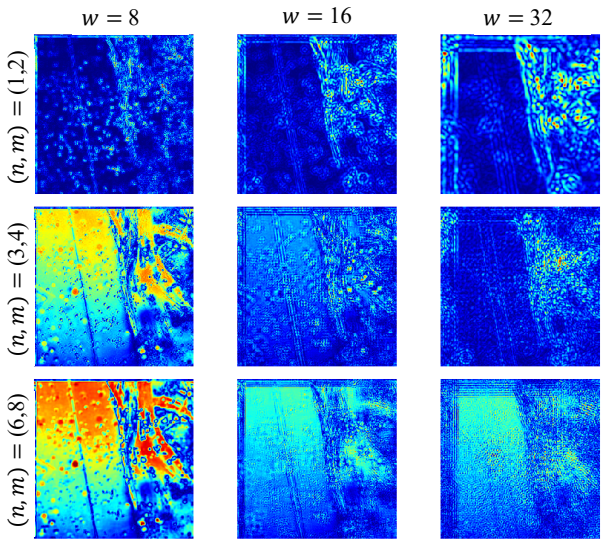

Fig. 3. Heatmaps of filtered images under different w and (n, m) parameters.

[29]. Since JPEG compression has a weaker effect on low-frequency components, the filtered response is less influenced by the degradation operator.

Proposition 1 (Robustness Approximation). *Under the low-frequency filter configuration, i.e., when (n, m) is small, considering the low-pass-like attenuation behavior of p and the bounded nature of η, the filtered output of the degraded image approximates the filtered output of the original image:*

$$I'_d \simeq I'. \tag{6}$$

This result follows from two key approximations:

$$F(p) \odot F((C^w_{nm})^T) \simeq F((C^w_{nm})^T), \tag{7}$$

because p has limited impact on low-frequency responses, and

$$F(\eta) \approx 0, \tag{8}$$

because η contributes negligible Fourier components under mild degradation. Substituting Eqs. (7) and (8) into Eq. (5) gives

$$\begin{aligned} I'_d &= F^{-1}\left(F(p) \odot F(I) \odot F((C^w_{nm})^T) + F(\eta) \odot F((C^w_{nm})^T)\right) \\ &\simeq F^{-1}\left(F(I) \odot F((C^w_{nm})^T)\right) \\ &= I', \end{aligned} \tag{9}$$

which supports the robustness approximation in Eq. (6).

The above proposition indicates that, under low-frequency filter configurations, the filtered output of a JPEG-compressed tampered image can approximate the filtered output before additional compression. Therefore, smaller values of (n, m) are generally associated with stronger robustness to JPEG-induced frequency attenuation.

In summary, larger (n, m) values produce richer high-frequency responses and provide stronger discriminability for subtle inpainting traces, while smaller (n, m) values yield more stable low-frequency responses and provide stronger robustness under compression. To balance these two properties, the proposed framework adopts a theory-guided empirical strategy. Multiple representative parameter configurations

spanning different frequency levels are explored within a bounded range, and the final settings are selected according to their localization performance on compressed validation samples. In practice, we select three multi-frequency representations as inputs to the detector.

Additionally, due to the effectiveness of the filter kernels in [20] for Photoshop inpainting localization, we incorporate three fixed convolution kernels as complementary residual extractors: a 3×3 first-order derivative kernel, a 3×3 second-order derivative kernel, and a 3×3 SRM kernel [30]. These fixed kernels provide additional high-frequency residual cues that are useful for detecting Photoshop inpainting traces. The three filter kernels are shown in Fig. 4.

Fig. 4. Three fixed filter kernels used to capture Photoshop inpainting traces.

B. Mechanism Analysis in the Frequency Domain

The preceding subsection has introduced the multi-frequency representation filters from an architectural perspective. Here, we provide a deeper mechanism analysis to explain why different frequency bands exhibit varying forensic utility for Photoshop inpainting detection and how JPEG compression differentially affects them. This analysis further justifies the multi-branch design of the proposed framework.

Natural images typically exhibit a power-law spectral decay, commonly referred to as the $1/f^{\alpha}$ distribution, where most signal energy is concentrated in low-frequency bands. These components mainly encode global luminance, coarse structures, and semantic layouts. In contrast, high-frequency bands contain fine-grained textures, sharp transitions, and local statistical variations. From a forensic perspective, manipulation operations often introduce localized anomalies, including boundary discontinuities, texture mismatches, and abnormal noise residuals. These artifacts are usually more evident in high-frequency components because they represent deviations from the original local statistics of the image. Therefore, high-frequency representations provide stronger discriminative information for detecting subtle tampering traces [31].

Photoshop-based inpainting further complicates this problem. Unlike naive copy-move or simple splicing, Photoshop inpainting relies on content-aware filling, patch-based synthesis, and adaptive blending to reconstruct missing regions. These mechanisms aim to maintain structural continuity and texture coherence with surrounding pixels. By sampling and blending neighboring patches, the inpainting process suppresses abrupt local discontinuities that would otherwise reveal manipulation boundaries. As a result, the remaining forensic traces are weak, localized, and often embedded in subtle residual patterns rather than obvious visual artifacts.

JPEG compression intensifies this difficulty. The JPEG pipeline applies DCT followed by quantization, where high-frequency coefficients are more aggressively quantized because they are less perceptible to human vision [32]. When applied to Photoshop-inpainted images, compression acts as a secondary degradation process that further suppresses the already weak high-frequency forensic cues. The combined effect of Photoshop-induced smoothing and JPEG-induced attenuation leads to a substantial loss of manipulation evidence, making reliable localization significantly more challenging.

This mechanism explains the necessity of multi-frequency representation. A detector relying only on high-frequency features can be highly sensitive to inpainting artifacts under clean conditions, but it becomes fragile once these features are attenuated by JPEG compression. Conversely, a detector relying only on low-frequency features may be robust to compression, but it lacks sufficient sensitivity to weak inpainting traces. The proposed MFR framework addresses this dilemma by combining branches tuned to different frequency bands. High-frequency branches preserve discriminative manipulation cues, low-frequency branches provide stable responses under compression, and the fusion module integrates their complementary predictions. Therefore, the proposed framework achieves a more principled balance between discriminability and robustness than single-frequency representations.

C. Detector for Photoshop Inpainting

This subsection introduces the branch-wise Photoshop inpainting detector used in the proposed MFR framework. The detector is designed to localize tampered regions at the pixel level. Its main component is an encoder-decoder network that learns forensic features from the filtered input. Since filtering may suppress part of the spatial-domain visual information, a spatial feature supplement module is further introduced. This module takes the raw RGB image as input and provides complementary spatial cues to the frequency-domain network. The overall detector architecture is illustrated in Fig. 5.

1) Fast Dense Block: To capture subtle traces of Photoshop tampering, we employ the dense block proposed in [33]. The dense block consists of four internal convolutional layers and one transition layer. After each convolution, Batch Normalization (BN) and ReLU activation are applied. The transition layer uses a 1×1 convolution to reduce the dimensionality of the output. The output X of the dense block is expressed as

$$X = T(\sum_{l=1}^{4} D_l(X_0 + \cdots + X_{l-1})),$$

where D_l denotes the convolution operation of the l-th layer, X_l represents the feature map output by the l-th layer, X_0 is the input, and T denotes the 1×1 convolution operation in the transition layer.

Since the proposed framework contains multiple frequency branches, directly training all branches increases computational cost. To improve efficiency, we adopt ghost convolution [34], which generates additional feature maps through low-cost linear transformations. Specifically, the standard convolutions in the 2nd, 3rd, 6th, and 7th dense blocks are replaced by ghost convolutions, reducing the number of parameters and

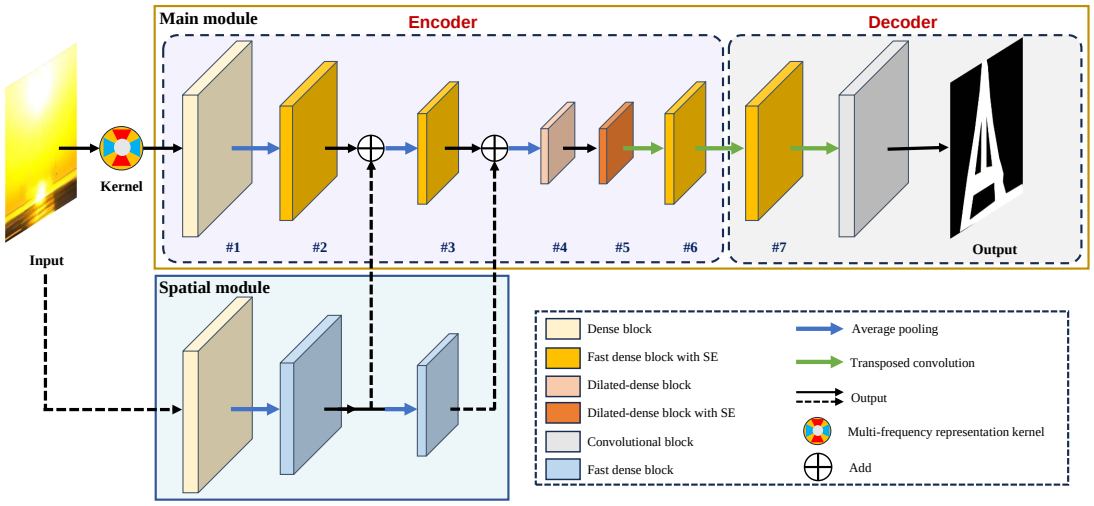

Fig. 5. Architecture of the proposed Photoshop inpainting detection. The main encoder-decoder structure contains seven dense blocks and one convolutional block. The spatial feature supplementary module contains three dense blocks and provides complementary spatial cues to the main network.

computational complexity in each branch. The output of the modified fast dense block is formulated as

$$X = T(X_0 + \sum_{l=1}^{2} G_l(X_0 + \cdots + X_{l-1})),$$

where G_l denotes the ghost convolution in the l-th layer, and T denotes the transition operation.

2) Main Encoder-Decoder Structure: The encoder extracts Photoshop inpainting-related forensic features from the filtered representation. It consists of one standard dense block, two fast dense blocks, and two dilated dense blocks [14]. The dilated dense blocks enlarge the receptive field, allowing the network to preserve contextual information during downsampling. Each dilated dense block contains four convolutions with a kernel size of 3 and a dilation rate of 2, followed by a transition layer.

The decoder converts the extracted feature maps into a tampering localization probability map. It consists of two fast dense blocks and one 5×5 convolution layer. Average pooling with a kernel size of 2 and a stride of 2 is used for downsampling, while transposed convolution with a kernel size of 4 and a stride of 2 is used for upsampling. A softmax layer is applied to generate pixel-wise probabilities, and a threshold of 0.5 is used to obtain the final binary localization map.

Since Photoshop inpainting traces are subtle and may be further weakened by compression, we introduce the Squeeze-and-Excitation (SE) module [35] into the main encoder-decoder structure. The SE module models channel-wise dependencies and enhances informative feature responses. The parameters and configurations of each block are summarized in Table I.

3) Spatial Domain Feature Supplement Module: The filtered representation emphasizes forensic residuals but may discard part of the spatial-domain information that is useful for localization, such as object structure and region continuity. To compensate for this loss, we design a spatial-domain feature supplement module. This module takes the raw RGB image as input and consists of one dense block and two fast dense blocks.

TABLE I
PARAMETERS OF DENSE BLOCKS IN THE MAIN MODULE

Dense block	Convolution type	Kernel	Stride	Output depth
#1	Normal	3	1	16
#2	Fast with SE	3	1	32
#3	Fast with SE	3	1	64
#4	Dilated	3	2	96
#5	Dilated with SE	3	2	96
#6	Fast with SE	3	1	64
#7	Fast with SE	3	1	48

The output feature maps of the second and third blocks in the supplement module are added element-wise to the corresponding feature maps of the second and third blocks in the main encoder-decoder network. This design allows the detector to preserve useful spatial context while still focusing on frequency-domain forensic traces, thereby improving the stability and accuracy of the localization results.

D. Multi-Frequency Representation Fusion Framework

This subsection describes how different frequency branches are integrated into the final MFR framework. Based on the analysis in Section III-A, different filter configurations provide different trade-offs between discriminability and robustness. High-frequency branches are more sensitive to subtle Photoshop inpainting artifacts, while low-frequency branches are more resistant to JPEG compression. Therefore, instead of selecting only one frequency representation, the proposed framework jointly utilizes multiple branches.

Specifically, three parameter-adjustable MFR filters are selected according to validation performance, and three fixed residual kernels from [20] are further incorporated. This produces six frequency-related input representations in total. Each representation is fed into an independent detector M_i, where $i = 1, 2, \ldots, 6$. Each branch is trained separately so that it

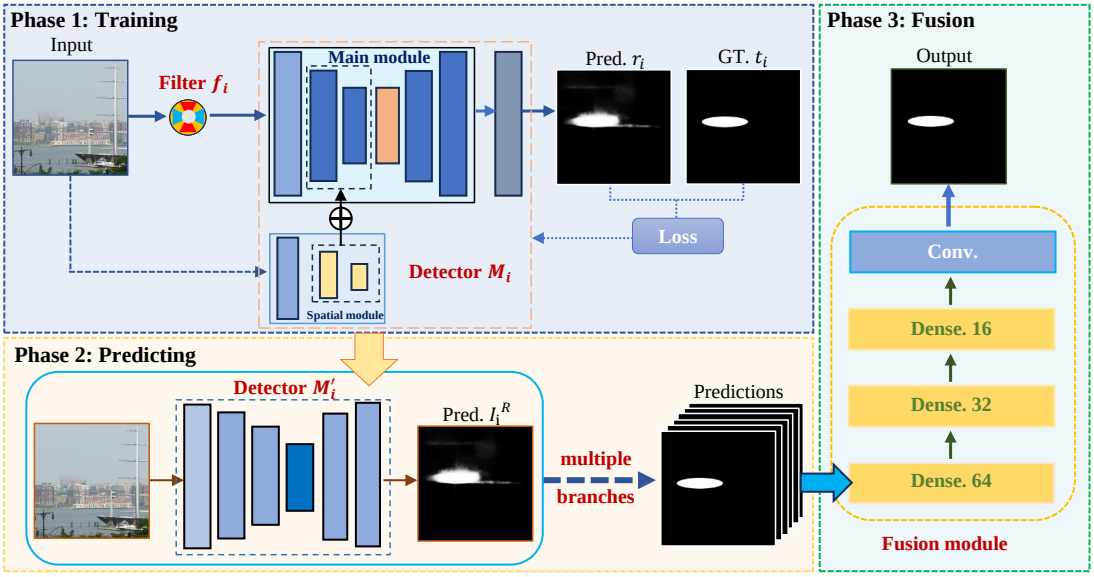

Fig. 6. Overview of the proposed training scheme. The six branches are trained individually, and each branch outputs a localization map. The fusion module merges the six branch outputs to generate the final prediction.

can specialize in the forensic features of its corresponding frequency representation. After each branch is optimized, its parameters are fixed. The trained branch detectors then generate six localization probability maps I_i^R for each input image.

After obtaining the six localization maps, we design a fusion module to integrate the complementary predictions from different branches. The six maps I_i^R are concatenated along the channel dimension to form a six-channel input tensor. The fusion module contains three fully connected layers with 64, 32, and 16 neurons, respectively, each followed by a ReLU activation function. The output is then processed by a 5×5 convolution layer for local normalization and refined prediction, followed by a softmax function to generate pixel-wise tampering probabilities. A threshold of 0.5 is used to classify each pixel as tampered or authentic.

The branch-wise training and fusion strategy has two advantages. First, independent training prevents different frequency representations from interfering with each other during feature learning, allowing each detector to specialize in its own spectral domain. Second, the fusion module learns to combine complementary evidence from different branches, thereby improving the balance between high-frequency discriminability and low-frequency robustness.

To further address the mismatch between training and testing compression levels, we construct a multi-quality JPEG training set. Motivated by the recompression and resizing operations commonly involved in online image sharing pipelines [36], five JPEG quality factors, i.e., QF = 95, 90, 85, 80, and 75, are applied to the original 50,000 Photoshop-inpainted training images, resulting in 250,000 compressed training samples. This strategy increases the diversity of compression-related artifacts and exposes each detector to different degradation strengths, improving robustness under practical JPEG compression and OSN transmission scenarios.

IV. Experiments

In this section, we evaluate the proposed Multi-Frequency Representation (MFR) framework from four aspects. First, we compare it with representative image forgery localization methods under simulated JPEG compression with different quality factors. Second, we evaluate its robustness in real-world OSN transmission scenarios. Third, we conduct ablation studies to analyze the contribution of multi-frequency fusion and multi-quality JPEG training. Finally, we further examine its robustness to common post-processing operations. All evaluations are conducted at the pixel level.

A. Experimental Setup

1) Datasets: For training and validation, we use the PS-scripted Places dataset [20], which is constructed from 50,000 randomly selected JPEG images from the Places database [37]. Each image is cropped to 256×256 pixels and tampered using the content-aware fill tool in Adobe Photoshop. The manipulated regions are generated by one to three regular masks and occupy approximately 4% to 21% of the image area. The resulting Photoshop-inpainted dataset without additional JPEG recompression is denoted as D_0.

To improve robustness against compression, we further compress D_0 using five JPEG quality factors, i.e., QF = 95, 90, 85, 80, and 75. This produces an augmented compressed training set D_1 containing 250,000 tampered images. The dataset D_1 is randomly split into training and validation subsets with a ratio of 9:1. This multi-quality construction allows the detector to observe Photoshop inpainting traces under different compression strengths and reduces the mismatch between training and testing degradation levels.

For testing, we randomly select 400 non-overlapping images from the Places database and crop them to 256×256 pixels. These images are not used during training or validation. Based

on the way tampered regions are generated, we construct two test sets:

- **Script-created dataset:** This dataset contains 100 images. The inpainting masks are regular regions selected by scripts, and the manipulated content is generated using the content-aware fill tool in Photoshop. Since its generation process is close to that of the training data, this dataset mainly evaluates in-distribution localization performance.
- **Hand-created dataset:** This dataset contains 300 images manually tampered with Photoshop tools. Specifically, we use three inpainting-related tools, including content-aware fill [15], patch, and content-aware move [16]. Each tool is used to create 100 tampered images, and the original image sets used by different tools do not overlap. For the content-aware move tool, rotation and scaling are further applied to the selected regions. Compared with the script-created dataset, this dataset better reflects practical manual editing scenarios and is therefore more challenging.

2) Evaluation Protocol: We evaluate the proposed method under both simulated and real-world degradation settings. For simulated JPEG compression, the script-created and hand-created datasets are used as base test sets. "Origin" denotes images without additional JPEG recompression, while "Weak", "Middle", and "Strong" denote JPEG-compressed images with QF = 95, 85, and 75, respectively. These settings represent progressively stronger compression levels. A higher QF generally preserves more image details, whereas a lower QF introduces stronger quantization artifacts and greater information loss. Therefore, the selected QF values allow us to evaluate the localization performance under mild, moderate, and relatively strong JPEG degradation.

For real-world evaluation, we transmit the hand-created dataset through three popular OSN platforms, namely Weibo, Wechat, and Twitter. The images are manually uploaded and downloaded using the default settings of each platform to better simulate common user behavior. Since OSN pipelines may involve resizing, recompression, watermarking, and other platform-dependent transformations, this setting provides a more practical evaluation of robustness beyond controlled JPEG compression.

3) Comparative Methods: We compare the proposed method with six representative image forgery localization methods:

- **ManTra-Net** [5]: An end-to-end fully convolutional network that formulates image manipulation localization as a local anomaly detection problem.
- **SATFL** [7]: A self-adversarial training method with a coarse-to-fine localization network for detecting forged regions.
- **IF-OSN** [23]: A robustness-oriented forgery detection method that models noise introduced by OSN transmission.
- **FOCAL** [11]: A forgery localization method based on contrastive learning and unsupervised clustering.
- **DFCN** [14]: A fully convolutional encoder-decoder network designed for detecting common Photoshop editing operations.
- **PS-Net** [20]: A primary-secondary network specifically designed to localize Photoshop-inpainted regions.

For methods whose source code and training procedures are available, including SATFL [7], FOCAL [11], DFCN [14], and PS-Net [20], we retrain them using the same augmented training set D_1 for a fair comparison under JPEG-compressed conditions. For ManTra-Net and IF-OSN, we use their official pretrained weights because their training pipelines are not directly compatible with our Photoshop inpainting dataset. This setting is further discussed in Section IV-F.

4) Implementation Details: The proposed method is implemented using TensorFlow. Adam [38] is adopted as the optimizer, and weighted cross-entropy [39] is used as the loss function. The initial learning rate is set to 0.001 and decays by 80% every three epochs. The batch size is set to 16. Each frequency branch is trained for 60 epochs, and the fusion module is trained for 4 epochs after the branch parameters are fixed.

For the adjustable filters, we select the three configurations that achieve the highest validation F1-scores, namely $w = 4$, $(n, m) = (3, 4)$; $w = 4$, $(n, m) = (4, 4)$; and $w = 8$, $(n, m) = (6, 8)$. Adobe Photoshop CC 2019 is used to generate all Photoshop-inpainted images. All experiments are conducted on a GeForce RTX 4090 GPU.

5) Performance Metrics: Since image forgery localization is a pixel-level binary classification problem, we adopt three widely used pixel-level metrics: Area Under the Receiver Operating Characteristic Curve (AUC), F1-score (F1), and Intersection over Union (IoU). AUC measures the overall ranking ability of predicted probabilities, while F1 and IoU evaluate the quality of binary localization maps. Following common practice, a threshold of 0.5 is used to convert probability maps into binary masks for F1 and IoU calculation.

B. Performance on Simulated JPEG Compression

Table II reports the localization performance of different methods on the script-created and hand-created test sets under four compression settings. Overall, the proposed method achieves the best or highly competitive results across most metrics and compression levels, demonstrating its effectiveness in balancing discriminability and compression robustness.

On the script-created dataset, our method consistently obtains the highest F1, AUC, and IoU values under all JPEG settings. Without additional recompression, it achieves an F1-score of 0.961 and an IoU of 0.929, indicating strong discriminability for Photoshop inpainting traces. As the compression strength increases, all methods exhibit performance degradation, confirming that JPEG compression weakens forensic evidence. Nevertheless, the proposed method maintains clear advantages under weak, middle, and strong compression. In particular, under strong compression, our method achieves an F1-score of 0.680 and an IoU of 0.542, outperforming PS-Net by 0.096 and 0.103, respectively. This verifies that multi-frequency representation can preserve useful localization cues even when high-frequency traces are partially suppressed.

TABLE II
LOCALIZATION PERFORMANCE OF DIFFERENT METHODS ON THE SCRIPT-CREATED AND HAND-CREATED TEST SETS UNDER SIMULATED JPEG COMPRESSION. "ORIGIN" DENOTES IMAGES WITHOUT ADDITIONAL JPEG RECOMPRESSION, WHILE "WEAK", "MIDDLE", AND "STRONG" CORRESPOND TO QF = 95, 85, AND 75, RESPECTIVELY. THE BEST RESULT IN EACH METRIC COLUMN IS HIGHLIGHTED IN BOLD.

Models	Test Datasets	JPEG Compression Levels											
		Origin			Weak			Middle			Strong		
		F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU
ManTra-Net[5]	Script-Created Dataset	0.080	0.494	0.042	0.080	0.494	0.042	0.080	0.494	0.042	0.080	0.494	0.042
SATFL[7]		0.002	0.443	0.001	0.694	0.949	0.558	0.462	0.899	0.324	0.214	0.796	0.132
IF-OSN[23]		0.016	0.528	0.001	0.006	0.487	0.003	0.005	0.458	0.003	0.004	0.459	0.002
DFCN[14]		0.916	0.992	0.861	0.773	0.976	0.649	0.704	0.964	0.567	0.507	0.905	0.359
PS-Net[20]		0.941	0.995	0.897	0.873	0.988	0.782	0.806	0.975	0.691	0.584	0.929	0.439
FOCAL[11]		0.528	0.567	0.094	0.512	0.535	0.071	0.498	0.522	0.053	0.504	0.526	0.062
Ours		**0.961**	**0.996**	**0.929**	**0.902**	**0.991**	**0.828**	**0.850**	**0.981**	**0.752**	**0.680**	**0.945**	**0.542**
ManTra-Net[5]	Hand-Created Dataset	0.167	0.493	0.097	0.167	0.493	0.097	0.167	0.493	0.097	0.167	0.493	0.097
SATFL[7]		0.002	0.456	0.001	0.605	0.876	0.469	0.364	0.807	0.238	0.116	0.663	0.066
IF-OSN[23]		0.180	0.720	0.143	0.108	0.665	0.084	0.059	0.621	0.046	0.040	0.574	0.031
DFCN[14]		0.836	0.935	0.748	0.744	0.907	0.627	0.660	0.861	0.526	0.424	0.719	0.219
PS-Net[20]		0.880	0.953	0.811	0.834	0.941	0.740	0.760	0.908	0.645	0.494	0.758	0.357
FOCAL[11]		0.615	0.644	0.256	0.590	0.613	0.219	0.580	0.602	0.204	**0.566**	0.585	0.182
Ours		**0.891**	**0.954**	**0.830**	**0.847**	**0.942**	**0.759**	**0.787**	**0.912**	**0.678**	0.529	**0.763**	**0.395**

On the hand-created dataset, the localization task becomes more challenging because the tampered regions are manually selected and generated using multiple Photoshop tools. This dataset also introduces greater diversity in object structures, boundary shapes, and editing patterns. As a result, the performance of all methods is generally lower than that on the script-created dataset. Even under this more realistic setting, our method achieves the best F1, AUC, and IoU under Origin, Weak, and Middle compression levels. For example, under Middle compression, our method obtains an F1-score of 0.787 and an IoU of 0.678, outperforming PS-Net by 0.027 and 0.033, respectively.

Under the strongest compression on the hand-created dataset, FOCAL obtains a slightly higher F1-score than our method. However, its AUC and IoU are substantially lower, indicating that its binary predictions are less reliable at the pixel level and that its localization regions are less accurate. In contrast, our method achieves the highest AUC and IoU in this setting, suggesting better probability calibration and spatial localization quality under severe degradation. These results indicate that the proposed MFR framework does not merely improve a single metric, but maintains a more balanced localization performance across different evaluation criteria.

The performance trends in Table II also reveal the limitations of existing methods. General-purpose detectors such as ManTra-Net and SATFL have difficulty capturing Photoshop-specific inpainting traces. IF-OSN is designed for OSN robustness but does not effectively model the weak artifacts produced by Photoshop inpainting. DFCN and PS-Net achieve strong performance under clean or weakly compressed conditions, but their performance drops substantially as compression becomes stronger. This supports our motivation that Photoshop inpainting localization under JPEG compression requires both manipulation-specific discriminability and degradation-resistant robustness. By integrating frequency branches with different sensitivities, the proposed method better addresses this trade-off.

C. Performance Under Real-World OSN Transmission

To further evaluate practical robustness, we test all methods on images transmitted through real OSN platforms. Unlike simulated JPEG compression, OSN transmission involves platform-dependent processing pipelines, including possible resizing, recompression, metadata removal, enhancement, and watermarking. These operations introduce more complex distribution shifts and may further obscure the already subtle traces left by Photoshop inpainting.

We use the hand-created dataset as the base test set because it better reflects realistic manual editing scenarios. Each image is manually uploaded to and downloaded from Weibo, Wechat, and Twitter using the default platform settings. This process yields three OSN-transmitted test sets. For example, images downloaded from Weibo may contain platform watermarks, which further increases the difficulty of localization.

As shown in Table III, the proposed method achieves the best F1 and IoU scores across all three OSN platforms. Specifically, our method obtains F1-scores of 0.829, 0.838, and 0.805 on Weibo, Wechat, and Twitter, respectively. Compared with the Origin setting on the hand-created dataset in Table II, the average F1-score decreases from 0.891 to 0.824, indicating that OSN processing does introduce noticeable degradation. Nevertheless, the proposed method still maintains strong localization performance, demonstrating its robustness under real-world transmission conditions.

Compared with Photoshop-oriented methods such as DFCN and PS-Net, our method achieves more stable performance after OSN transmission. Although PS-Net obtains a comparable AUC on Weibo and a slightly higher AUC on Wechat, our method consistently provides higher F1 and IoU values, suggesting more accurate binary localization masks and better spatial overlap with ground-truth regions. This indicates that the multi-frequency fusion strategy improves not only the ranking quality of probability maps but also the final localization accuracy.

Fig. 7 provides qualitative comparisons under OSN transmission. General-purpose methods such as ManTra-Net,

Fig. 7. Qualitative localization results of different methods on the hand-created dataset after OSN transmission through Weibo, Wechat, and Twitter.

TABLE III
LOCALIZATION PERFORMANCE OF DIFFERENT METHODS UNDER REAL-WORLD OSN TRANSMISSION SCENARIOS. THE BEST RESULT IN EACH METRIC COLUMN IS HIGHLIGHTED IN BOLD.

Models	Weibo			Wechat			Twitter		
	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU
ManTra-Net[5]	0.167	0.493	0.097	0.167	0.493	0.097	0.167	0.493	0.097
SATFL[7]	0.599	0.872	0.462	0.534	0.836	0.398	0.446	0.825	0.314
IF-OSN[23]	0.167	0.706	0.131	0.180	0.720	0.143	0.083	0.641	0.066
DFCN[14]	0.736	0.903	0.617	0.712	0.892	0.586	0.678	0.873	0.546
PS-Net[20]	0.814	**0.936**	0.714	0.828	**0.937**	0.733	0.785	0.917	0.674
FOCAL[11]	0.605	0.635	0.240	0.622	0.655	0.264	0.597	0.622	0.230
Ours	**0.829**	**0.936**	**0.735**	**0.838**	0.936	**0.749**	**0.805**	**0.920**	**0.702**

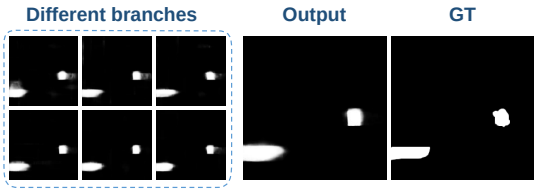

Fig. 8. Visualization of individual branch outputs and the final fused output.

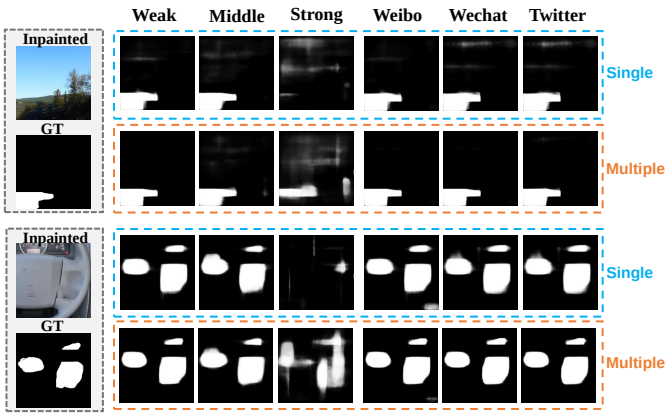

Fig. 9. Comparison between models trained with a single JPEG quality factor and models trained with multiple JPEG quality factors. "Single" denotes training with one fixed JPEG quality factor, while "Multiple" denotes training with data augmented across multiple JPEG compression strengths.

SATFL, IF-OSN, and FOCAL often fail to localize Photoshop-inpainted regions accurately. DFCN and PS-Net can detect some manipulated areas but tend to produce false negatives or false positives after platform processing. In contrast, the proposed MFR-based method yields more complete and spatially accurate localization results. These visual observations are consistent with the quantitative results in Table III.

D. Ablation Study

We conduct ablation studies to verify the effectiveness of two key designs in the proposed framework: multi-frequency branch fusion and multi-quality JPEG training. The former evaluates whether combining different frequency branches improves localization performance, while the latter examines whether training with multiple compression levels improves generalization to unseen degradation strengths.

1) Effectiveness of Multi-Frequency Fusion: Table IV reports the performance of each individual frequency branch and the final fused model. The results show that no single branch consistently achieves the best performance across all test conditions. This is expected because different branches emphasize different frequency ranges and therefore capture different types of forensic evidence. Some branches are more sensitive to manipulation-specific high-frequency traces, while others provide more stable responses under compression and OSN transmission.

After fusing the six branches, the proposed model achieves the best F1, AUC, and IoU scores under all evaluated settings in Table IV. On Hand[1], the fused model improves the F1-score to 0.847, compared with the best single-branch F1-score of 0.839. On Hand[3], where compression is strongest, the fused model achieves an F1-score of 0.529 and an IoU of 0.395, outperforming all individual branches. Under Twitter transmission, the fused model reaches an F1-score of 0.805, whereas the average F1-score of the individual branches is approximately 0.778. These results demonstrate that different frequency branches provide complementary information and that the fusion module effectively integrates this information for more robust localization.

Fig. 8 shows a visual example of branch-wise predictions and the final fused output. Individual branches may miss parts of the tampered region or introduce false alarms due to their limited frequency preference. In contrast, the fused output provides a more complete localization mask with clearer manipulation boundaries. The visualization further indicates that multi-frequency fusion effectively reduces false positive regions and improves boundary consistency, demonstrating better robustness and localization accuracy than any single-branch prediction.

2) Effectiveness of Multi-Quality JPEG Training: We further evaluate the influence of multi-quality JPEG training. A detector trained with only one compression level may overfit to the corresponding degradation distribution and fail to generalize to other JPEG quality factors. This problem is particularly important for real-world applications because the compression strength used by OSNs or image storage systems is often unknown.

To examine this issue, we compare models trained with a single JPEG quality factor against the model trained with the multi-quality dataset D_1. As shown in Fig. 9, models trained with a single compression level exhibit unstable performance when tested under different compression strengths. In contrast, the model trained with multiple JPEG qualities maintains more

consistent performance across different testing conditions. This indicates that the augmented dataset D_1 exposes the detector to a broader range of compression-induced variations, allowing it to learn more generalizable forensic representations.

These results confirm that the performance improvement of the proposed method is not solely due to network architecture. The multi-quality training strategy also plays an important role in improving robustness to compression mismatch between training and testing. When combined with multi-frequency fusion, it enables the detector to maintain discriminability for Photoshop inpainting while reducing sensitivity to specific JPEG quality factors.

E. Robustness to Additional Post-Processing

In addition to JPEG compression and OSN transmission, we further evaluate the robustness of the proposed method under several common post-processing operations in Photoshop. These operations include color adjustment, contrast adjustment, sharpening, rotation, flipping, mirroring, Gaussian blur, and Gaussian noise. All tests are conducted on the hand-created dataset. The model is trained only on D_1, which does not include these additional post-processing operations, so this experiment evaluates the generalization ability of the proposed method beyond JPEG compression.

Fig. 10 reports the pixel-level AUC scores under different post-processing settings. For reference, Type 0 denotes the performance on tampered images without additional post-processing. The results show that the proposed method maintains competitive robustness under color adjustment, contrast adjustment, sharpening, rotation, flipping, and mirroring. These operations modify either photometric properties or image orientation, but they do not completely remove the frequency-domain inconsistencies introduced by Photoshop inpainting. Therefore, the proposed multi-frequency representation can still preserve useful forensic cues.

For Gaussian noise, the proposed method also maintains a certain degree of robustness. Even when the noise amount increases to 1.2%, the AUC score remains above 0.7, indicating that the detector can tolerate moderate random perturbations. However, the performance decreases more noticeably under Gaussian blur. When the blur radius increases to 0.4, the AUC score drops significantly. This is because Gaussian blur directly suppresses high-frequency components, which are critical for localizing subtle inpainting traces. This observation is consistent with our frequency-domain analysis: although low-frequency branches improve robustness, excessive removal of high-frequency forensic evidence still limits localization performance.

Overall, the proposed method demonstrates robustness not only to JPEG compression and OSN transmission, but also to several common post-processing operations. At the same time, the results reveal that strong blur remains a challenging case, suggesting that future work should further enhance the recovery or preservation of manipulation-sensitive high-frequency cues under severe smoothing.

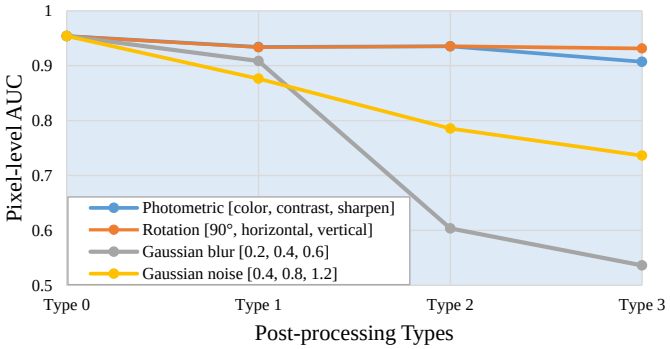

Fig. 10. Pixel-level AUC scores on the hand-created dataset under additional post-processing operations, including color adjustment, contrast adjustment, sharpening, rotation, Gaussian noise, and Gaussian blur.

F. Discussion

The comparison in this work involves two evaluation settings for baseline methods. For methods with available source code and reproducible training procedures, we retrain them using the same augmented dataset D_1. For ManTra-Net and IF-OSN, we use official pretrained models because their training pipelines are not directly compatible with our Photoshop inpainting training data. This difference may affect the absolute performance of some baselines and should be considered when interpreting the results. Nevertheless, we follow the most consistent protocol possible under practical constraints: retrainable methods are trained with identical data and preprocessing, while official checkpoints are evaluated using their original inference settings to avoid implementation bias.

The experimental results reveal complementary limitations of existing methods. General-purpose forgery localization methods are not specifically optimized for Photoshop inpainting and therefore struggle to capture the weak traces left by content-aware editing. Photoshop-oriented methods such as DFCN and PS-Net show strong performance under clean or weakly compressed conditions, but their performance decreases substantially when JPEG compression or OSN processing weakens high-frequency forensic evidence. Robustness-oriented methods such as IF-OSN consider platform-induced degradation, but they do not explicitly model Photoshop-specific inpainting artifacts. These observations support the necessity of jointly considering manipulation-specific discriminability and degradation-resistant robustness.

Although the proposed MFR framework consistently improves localization performance under JPEG compression and OSN transmission, its performance still declines as degradation becomes severe. This decline is expected because Photoshop inpainting traces are inherently weak and localized, and JPEG compression or OSN processing can further suppress the high-frequency residuals that are important for localization. In particular, severe Gaussian blur remains challenging because it directly removes fine-grained forensic cues. Therefore, the proposed method improves the robustness-discriminability balance but does not completely eliminate the difficulty caused by strong information loss.

TABLE IV
ABLATION RESULTS OF INDIVIDUAL FREQUENCY BRANCHES AND THE FINAL FUSED MODEL. "ALL" DENOTES THE FUSION OF ALL SIX BRANCHES.

Branch	Hand[1]			Hand[2]			Hand[3]			Twitter		
	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU
1	0.828	0.934	0.732	0.743	0.893	0.625	0.499	0.757	0.361	0.767	0.902	0.654
2	0.839	0.941	0.748	0.766	0.912	0.648	0.480	0.753	0.345	0.790	0.921	0.681
3	0.832	0.940	0.739	0.760	0.908	0.644	0.504	0.761	0.367	0.783	0.916	0.673
4	0.825	0.937	0.727	0.755	0.907	0.636	0.493	0.756	0.358	0.773	0.914	0.658
5	0.826	0.939	0.729	0.754	0.904	0.639	0.491	0.758	0.354	0.779	0.916	0.668
6	0.822	0.931	0.726	0.756	0.901	0.642	0.491	0.750	0.356	0.774	0.909	0.664
ALL	**0.847**	**0.942**	**0.759**	**0.787**	**0.912**	**0.678**	**0.529**	**0.763**	**0.395**	**0.805**	**0.920**	**0.702**

Future work may focus on enhancing the discriminability and adaptability of forensic representations under severe degradation. One promising direction is to introduce contrastive learning to enlarge the feature separation between tampered and authentic regions, enabling the model to preserve manipulation-sensitive cues even when compression weakens forensic traces. Another direction is to develop adaptive cross-frequency fusion mechanisms, allowing the detector to dynamically assign higher weights to the most reliable spectral cues under different degradation levels. These improvements may further strengthen robustness while maintaining accurate localization in real-world scenarios.

V. CONCLUSION

This paper presents a JPEG-resistant Photoshop inpainting localization method based on multi-frequency representation. The work is motivated by a fundamental observation: discriminability and robustness in forgery localization correspond to conflicting spectral preferences, and existing methods that operate at a fixed frequency band cannot simultaneously satisfy both requirements. By designing parameterized frequency-selective filters to extract features at multiple spectral levels, training independent branch detectors for each frequency band, and integrating their outputs through a dedicated fusion module, the proposed framework explicitly exploits the complementarity between high-frequency discriminative cues and low-frequency robustness. A theoretical analysis further supports this design by showing that lower-frequency filter configurations are less susceptible to JPEG-induced attenuation, providing a principled basis for the multi-branch strategy.

A multi-quality JPEG augmentation strategy is also incorporated to reduce the domain gap between training and testing compression levels, enabling the model to generalize across a wide range of practical distortion conditions. Extensive experiments on both script-created and hand-created Photoshop inpainting datasets demonstrate consistent superiority over representative forgery localization methods across multiple JPEG compression strengths. Evaluations on images transmitted through Wechat, Weibo, and Twitter further confirm that the method maintains strong localization performance in real-world OSN scenarios.

Beyond the specific task of Photoshop inpainting, the broader insight of this work is that the spectral frequency of a forensic representation is a principled axis along which robustness and discriminability can be jointly optimized. Future research directions include developing adaptive frequency fusion mechanisms that dynamically reweight spectral branches based on estimated distortion severity, as well as extending the multi-frequency representation framework to handle a broader class of manipulations and more complex OSN-induced degradations without requiring task-specific retraining.

REFERENCES

[1] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 3, pp. 1003–1017, 2012.

[2] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of cfa artifacts," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 5, pp. 1566–1577, 2012.

[3] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A bayesian-mrf approach for prnu-based image forgery detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 4, pp. 554–567, 2014.

[4] H. Li, W. Luo, and J. Huang, "Localization of diffusion-based inpainting in digital images," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 12, pp. 3050–3064, 2017.

[5] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9543–9552.

[6] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1053–1061.

[7] L. Zhuo, S. Tan, B. Li, and J. Huang, "Self-adversarial training incorporating forgery attention for image forgery localization," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 819–834, 2022.

[8] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3539–3553, 2023.

[9] D. Li, J. Zhu, M. Wang, J. Liu, X. Fu, and Z.-J. Zha, "Edge-aware regional message passing controller for image forgery localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8222–8232.

[10] F. F. Niloy, K. K. Bhaumik, and S. S. Woo, "Cfl-net: Image forgery localization using contrastive learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4642–4651.

[11] H. Wu, Y. Chen, and J. Zhou, "Rethinking image forgery detection via contrastive learning and unsupervised clustering," 2023. [Online]. Available: https://arxiv.org/abs/2308.09307

[12] Y. Zeng, B. Zhao, S. Qiu, T. Dai, and S.-T. Xia, "Toward effective image manipulation detection with proposal contrastive learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4703–4714, 2023.

[13] D. Li, J. Zhu, X. Fu, X. Guo, Y. Liu, G. Yang, J. Liu, and Z.-J. Zha, "Noise-assisted prompt learning for image forgery detection and localization," in *European Conference on Computer Vision*. Springer, 2024, pp. 18–36.

[14] P. Zhuang, H. Li, S. Tan, B. Li, and J. Huang, "Image tampering localization using a dense fully convolutional network," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2986–2999, 2021.

[15] content-aware-fill. [Online]. Available: https://helpx.adobe.com/photoshop/using/content-aware-fill.html

[16] content-aware-patch-move. [Online]. Available: https://helpx.adobe.com/photoshop/using/content-aware-patch-move.html

[17] H. Li and J. Huang, "Localization of deep inpainting using high-pass fully convolutional network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8301–8310.

[18] H. Wu and J. Zhou, "Iid-net: Image inpainting detection network via neural architecture search and attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1172–1185, 2022.

[19] Y. Zhang, Z. Fu, S. Qi, M. Xue, Z. Hua, and Y. Xiang, "Localization of inpainting forgery with feature enhancement network," *IEEE Trans. Big Data*, vol. 9, no. 3, pp. 936–948, 2023.

[20] Y. Zhang, Z. Fu, S. Qi, M. Xue, X. Cao, and Y. Xiang, "Ps-net: A learning strategy for accurately exposing the professional photoshop inpainting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13 874–13 886, 2024.

[21] M. Wang, X. Fu, J. Liu, and Z.-J. Zha, "Jpeg compression-aware image forgery localization," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 5871–5879.

[22] Y. Rao, J. Ni, W. Zhang, and J. Huang, "Towards jpeg-resistant image forgery detection and localization via self-supervised domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–12, 2022.

[23] H. Wu, J. Zhou, J. Tian, J. Liu, and Y. Qiao, "Robust image forgery detection against transmission over online social networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 443–456, 2022.

[24] P. Zhuang, H. Li, R. Yang, and J. Huang, "Reloc: A restoration-assisted framework for robust image tampering localization," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 5243–5257, 2023.

[25] W. Shan, A. Liu, J. Qiu, and J. Li, "Slrid: A robust image tampering localization framework for extremely scaled forgery images," *IEEE Signal Process. Lett.*, vol. 31, pp. 2095–2099, 2024.

[26] S. Qi, Y. Zhang, C. Wang, J. Zhou, and X. Cao, "A principled design of image representation: Towards forensic tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5337–5354, 2023.

[27] ——, "A survey of orthogonal moments for image representation: Theory, implementation, and evaluation," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–35, 2021.

[28] G. K. Wallace, "The jpeg still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[29] J. Flusser, B. Zitova, and T. Suk, *Moments and Moment Invariants in Pattern Recognition*. John Wiley & Sons, 2009.

[30] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 3, pp. 868–882, 2012.

[31] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[32] J. Fridrich, "Digital image forensics," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 26–37, 2009.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[34] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[36] C. Erdmann, "Better Image Quality in Your Twitter Tweets," Compress-Or-Die, 2022, last updated: Dec. 21, 2022. [Online]. Available: https://compress-or-die.com/Better-image-quality-in-your-Twitter-tweets. Accessed: May 28, 2026.

[37] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2018.

[38] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 191–198.

Yushu Zhang (Senior Member, IEEE) received the Ph.D. degree from Chongqing University, Chongqing, China, in 2014. He is currently a Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include privacy, security, and trustworthy AI. He serves on the Editorial Boards of *IEEE Transactions on Dependable and Secure Computing*, *Signal Processing*, and *Information Sciences*.

Lu Zhang received the M.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2026. Her research interests include media forensics and security.

Shuren Qi received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2024. He is currently a Postdoctoral Fellow with the Department of Data Science, City University of Hong Kong, Hong Kong, China. He was previously a Postdoctoral Fellow with The Chinese University of Hong Kong, Hong Kong, China. His research interests include invariants, representations, and geometric deep learning.

Xiangli Xiao received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2024. He is currently a Lecturer with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include multimedia security, digital watermarking, blockchain, and cloud computing security.

Wenying Wen received the Ph.D. degree from Chongqing University, Chongqing, China, in 2013. She is currently a Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. Her research interests include image processing, multimedia security, and artificial intelligence security.