Enhancing Tampered Text Detection through Frequency Feature Fusion and Decomposition

Zhongxi Chen[1,2*†], Shen Chen[2†], Taiping Yao[2✉], Ke Sun[1], Shouhong Ding[2], Xianming Lin[1✉], Liujuan Cao[1], and Rongrong Ji[1]

[1] Key Laboratory of Multimedia Trusted Perception and Effcient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China.
[2] Youtu Lab, Tencent, China.
chenzhongxi@stu.xmu.edu.cn

Abstract. Document image tampering poses a grave risk to the veracity of information, with potential consequences ranging from misinformation dissemination to financial and identity fraud. Current detection methods use frequency information to uncover tampering that is invisible to the naked eye. However, these methods often fail to integrate this information effectively, thereby compromising RGB detection capabilities and missing the high-frequency details necessary to detect subtle tampering. To address these gaps, we introduce a Feature Fusion and Decomposition Network (FFDN) that combines a Visual Enhancement Module (VEM) with a Wavelet-like Frequency Enhancement (WFE). Specifically, the VEM makes tampering traces visible while preserving the integrity of original RGB features using zero-initialized convolutions. Meanwhile, the WFE decomposes the features to explicitly retain high-frequency details that are often overlooked during downsampling, focusing on small but critical tampering clues. Rigorous testing on the DocTamper dataset confirms FFDN's preeminence, significantly outperforming existing state-of-the-art methods in detecting tampering.

Keywords: Document Image Tampering Detection · Image Manipulation Detection · Semantic Segmentation

1 Introduction

With text being a primary information medium, document image tampering can lead to serious consequences, ranging from spreading fake news to severe issues like identity theft and financial fraud [31]. Therefore, Document Image Tampering Detection (DITD) has become crucial in various fields, including digital forensics [25], e-commerce [1], and social media platforms [20]. The rapid advancement of image editing tools, while beneficial for legitimate purposes, has also facilitated image manipulation by malicious actors, compromising digital

* Work done during an internship at YouTu Lab, Tencent.
† Equal Contribution. ✉ Corresponding authors.

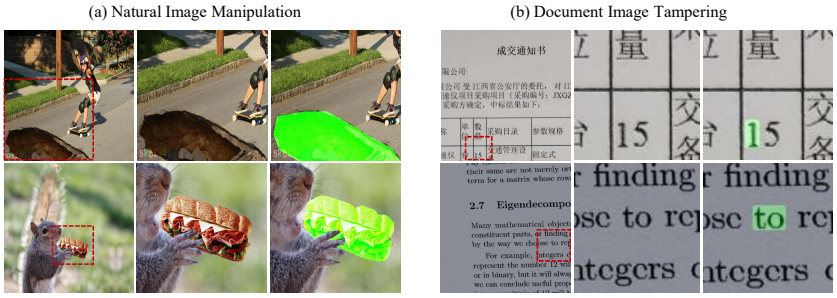

Fig. 1: Comparison of Natural Image Manipulation and Document Image Tampering. From left to right includes the Original Image, a Zoomed-in Detail of the Image, and the Ground Truth. Notably, the areas of document tampering are typically very small, with the tampering traces being quite subtle.

content integrity and posing significant security risks. Consequently, the development of a robust method for detecting tampered text is essential to preserve the trustworthiness of digital content.

Unlike traditional Natural Image Manipulation Detection [46], DITD presents two main challenges: 1) Imperceptible modifications: documents typically have a consistent and relatively simple background, making it easier for malicious actors to create visually imperceptible manipulations that are difficult to detect. 2) Small tampering regions: unlike the large-scale and semantically-based manipulations in natural images, document tampering often occurs in localized characters, resulting in small manipulation regions. Figure 1 illustrates the challenges of the DITD task compared to Natural Image Manipulation Detection.

To address these two challenges, two lines of methods have been proposed. On one hand, some methods [7, 21, 22, 28, 36, 41] employ frequency domain information as auxiliary clues to mine invisible tampering traces. For example, Wang *et al.* [35] adopt Laplacian of Gaussian (LoG) to capture high-frequency information, then the RGB and frequency domain features extracted are fused through element-wise add operation. Further, DTD [22] concatenates the DCT feature with the RGB feature and enhances it using the scSE attention [24]. However, previous research [9, 14, 44] has demonstrated that frequency domain information can sometimes be counterproductive, particularly in images without prior JPEG compression (e.g., PNG images) or those post-processed multiple times after tampering, which may compromise RGB domain detection. Therefore, *effectively balancing and integrating spatial and frequency domain information* is a crucial challenge in DITD. On the other hand, the identification of tampering in documents is challenging due to the the subtle nature of tampering traces and the typically small tampered areas. Previous methods [32, 42] have shown that the forged areas are often hidden in detail regions. To address this, some approaches have proposed specially designed network components, such as Multi-view Iterative Decoders [22] and high-resolution structures [26, 44] to prevent the loss of subtle features. However, due to operations like downsampling, some subtle

tampering cues can still be easily overlooked by the network. Therefore, *how to explicitly enhance small tampering features* remains a crucial challenge.

In response to the challenges in DITD, we introduce our Feature Fusion and Decomposition Network (FFDN). This approach consists of two main components: a Visual Enhancement Module (VEM) and a Wavelet-like Frequency Enhancement (WFE). The VEM enhances visual features by incorporating frequency domain information through an attention mechanism, thus improving the detection of tampering traces that are imperceptible to the naked eye. Additionally, it employs a zero-initialized conv to effectively reduce frequency domain noise by emphasizing visual information. Complementing the VEM, the WFE explicitly decomposes features into high and low-frequency components, amplifying tampering discrepancies across various scales and improving the detection of small or subtle tampered clues. Consequently, our network achieves a more nuanced and effective detection of subtle tampering traces: the VEM makes tampering traces visible, while the WFE resolves the subtleties of these traces, setting our approach apart in the realm of tampering detection technology.

Our research, conducted on the recently released DocTamper [22] dataset, addresses the inherent issues in existing methods, particularly the lack of controlled integration and insufficient utilization of high-frequency information. The experimental results demonstrate that our proposed method significantly outperforms current state-of-the-art competitors. In summary, the key contributions of this paper are as follows:

1. We propose an innovative Visual Enhancement Module. It enhances the model's ability to detect imperceptible tampering through frequency information while maintaining the integrity of RGB.
2. We introduce a Wavelet-like Frequency Enhancement, which explicitly separates multi-level features into high and low-frequency components. This approach ensures the preservation of fine details and the effective exploitation of small tampering cues.
3. Through comprehensive experiments, we demonstrate that our proposed methods surpass existing techniques, especially for small targets, thereby advancing the state-of-the-art in DITD.

2 Related Works

2.1 Natural Image Manipulation Detection

In natural image manipulation, tampering methods mainly fall into three categories: splicing, copy-move, and generation. Splicing involves copying regions from one image to another, copy-move entails shifting objects within the same image, and generation replaces image regions with visually plausible but different content. To address these various forms of image manipulation, researchers have developed numerous detection techniques. Recently, Zhou *et al.* [46] and Bappy *et al.* [2] have integrated SRM kernels into convolutional neural networks for more precise forgery localization. Liu *et al.* [17] utilized an attention-based

progressive network, PSCC-Net, for multi-scale tampering localization. Zhuang *et al.* [48] focused on detecting subtle tampering by pre-training a network with a dataset that simulates common image editing operations. In a dual-domain approach, Kwon *et al.* [15] utilized a DCT branch to detect JPEG compression artifacts, integrating it with spatial features for comprehensive tampering localization. More recently, Transformer models [30] have been applied to this domain, with TransForensics [10] and ObjectFormer [32] being notable examples. These models are designed to detect larger object-level manipulations in natural images. However, the specific challenge of detecting tampered text in images remains significant. The visual similarity between authentic and manipulated text, along with the small tampered areas, indicates a gap in current methodologies that requires specialized attention.

2.2 Document Image Tampering Detection

DITD, a task distinct from text segmentation, is designed to identify and pinpoint tampered sections within document images. Inspired by Natural Image Manipulation Detection methodologies, a multitude of techniques have been devised. Previous methods [13, 29] that relied solely on visual clues have proven ineffective in detecting tampered areas. For instance, Joren *et al.* [13] redefined the conventional image splice detection problem as a node classification problem, where Optical Character Recognition (OCR) bounding boxes form nodes and edges are added based on a text-specific distance heuristic. More recent studies have made substantial advancements by incorporating frequency-based methods. Xu *et al.* [41], for instance, leveraged the residual filter in the second stream to learn manipulation traces in pixel correlation. Yanikoglu *et al.* [43] utilize Fast Fourier Transform features for signature verification, and others [22, 36, 41] employ the Discrete Cosine Transform to extract spectral features, aiding in the detection of discontinuities in Block Artifact Grids (BAG) between tampered and genuine areas. Additionally, DTD [36] proposed the Curriculum Learning strategy to fit the difficulty of compressed image during training.

However, despite the benefits of incorporating frequency domain information to discern forgery traces, we observed that the direct fusion of both modalities can result in a decline in detection capability within the RGB domain. We posit that DCT information needs to be introduced into the detection model in a more controlled and adaptive manner.

3 Methodology

3.1 Overall Framework

In this section, we introduce the Feature Fusion and Decomposition Network (FFDN), a novel architecture for DITD, as illustrated in Figure 2. The process is initiated by the Visual and Frequency Perception Head, which extracts visual and frequency information from the image and DCT coefficients, respectively. This information is then integrated by the Visual Enhancement Module

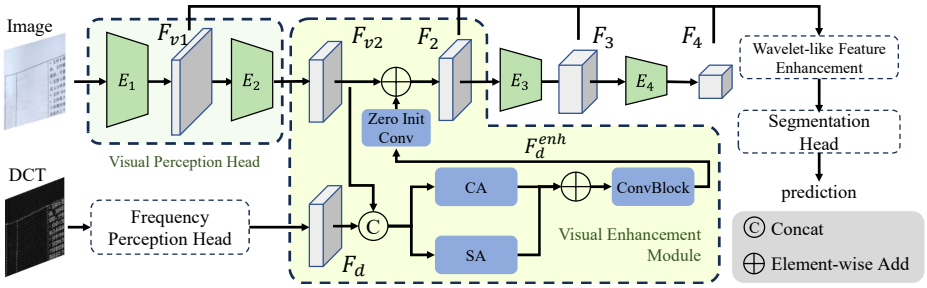

Fig. 2: The overall framework of our method. It starts with the Visual and Frequency Perception Head extracting features, followed by the Visual Enhancement Module for balanced data integration. The Wavelet-like Frequency Enhancement then sharpens the feature separation, and finally, a segmentation head like the FPN provides detailed tampering detection.

(VEM), which brings the frequency information into the visual feature space, and ensures the RGB information is the main guide. Subsequently, the Wavelet-like Frequency Enhancement (WFE) takes over, meticulously separating the features into high and low-frequency details to retain the small tampering artifacts. Lastly, an FPN [16] Head enables pixel-wise tampering detection. Collectively, these components constitute a robust and efficient framework for detecting even subtle signs of document image tampering. We first introduce the two heads in Section 3.2, then explore the structure of VEM in Section 3.3, and finally, discuss the design details of the WFE in Section 3.4.

3.2 Visual and Frequency Perception Head

Following Qu *et al.* [22], our method utilizes a dual-branch structure to analyze both visual and frequency information. Given an RGB image $I \in \mathbb{R}^{3\times H\times W}$, which is converted to the YCbCr color space, where H and W represent the height and width of the image, respectively. From the Y channel, we compute the DCT coefficients $D \in \mathbb{R}^{H\times W}$ and the quantization table $T \in \mathbb{R}^{8\times 8}$. The Visual branch consists of a two-layer encoder (E_1 and E_2) that focuses on detecting clear tampering indicators such as color and font inconsistencies. The Frequency branch, inherited from DTD [22], targets more subtle anomalies like BAG. It takes D and T as input, multiplies them to obtain decoded frequency characteristics, and uses dilated convolutions to capture grid features from D, yields a frequency representation F_d. This process can be represented as:

$$F_{v1} = \mathrm{E}_1(I), \quad F_{v2} = \mathrm{E}_2(F_{v1}), \quad F_d = \mathrm{FPH}(D, T), \tag{1}$$

where $F_{v1} \in \mathbb{R}^{C_1\times \frac{H}{4}\times \frac{W}{4}}$, $F_{v2} \in \mathbb{R}^{C_2\times \frac{H}{8}\times \frac{W}{8}}$ and $F_d \in \mathbb{R}^{C_d\times \frac{H}{8}\times \frac{W}{8}}$, represent the two-tiered visual and frequency representations, respectively.

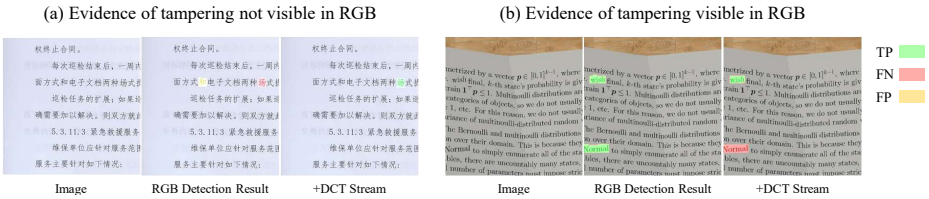

Fig. 3: Visualization of the tampering detection results. The left image shows a case where tampering is difficult to discern, while the right image displays evident tampering, noticeable through inconsistencies in position and font. The colors green, red, and yellow represent True Positives (TP), False Negatives (FN), and False Positives (FP), respectively. The introduction of DCT information has enhanced the model's ability to detect tampering traces that are invisible to the naked eye. However, paradoxically, this has led to a decrease in detection performance for some samples where tampering could be identified using RGB information alone.

3.3 Visual Enhancement Module

Our Visual Enhancement Module (VEM) is specifically designed to selectively incorporate frequency information as needed, enhancing the model's capability to detect imperceptible tampering while preserving the integrity of RGB-based detection. We recognize from research [9, 44] that an excessive dependence on DCT information can hinder the model's ability to detect tampering, especially for PNG images or those post-processed multiple times, sometimes even impairing overall performance. Moreover, our observations suggest that while frequency features can improve the detection of nuanced tampering, they may also interfere with the model's ability to detect tampering that is readily apparent in RGB data, as demonstrated in our easy and hard case analysis, as illustrated in Figure 3. To address these challenges, we propose an adaptive fusion mechanism. We first extract frequency features F_d and RGB features F_{v2}, and then apply spatial and channel attention mechanisms, to produce an enhanced frequency feature F_d^{enh} guided by the RGB features. The fusion is executed as follows:

$$F_d^{enh} = \text{ConvBlock}(\text{CA}([F_{v2}, F_d]) + \text{SA}([F_{v2}, F_d])), \tag{2}$$

where $[\cdot]$ denotes the concatenation operation, CA and SA represent channel and spatial attention mechanisms, and ConvBlock is a combination of 1×1 convolution, batch normalization and ReLU operations. To prioritize RGB features in the fusion and allow frequency information to be adaptively introduced, we pass the frequency features through a zero-initialized convolution [45] prior to merging them with the RGB features:

$$F_2 = \text{Conv}(F_d^{enh}) + F_{v2}. \tag{3}$$

The zero-initialized convolution ensures that the model starts by relying primarily on RGB features and gradually learns to incorporate frequency information only when it proves beneficial. Following the VEM, the enhanced feature F_2

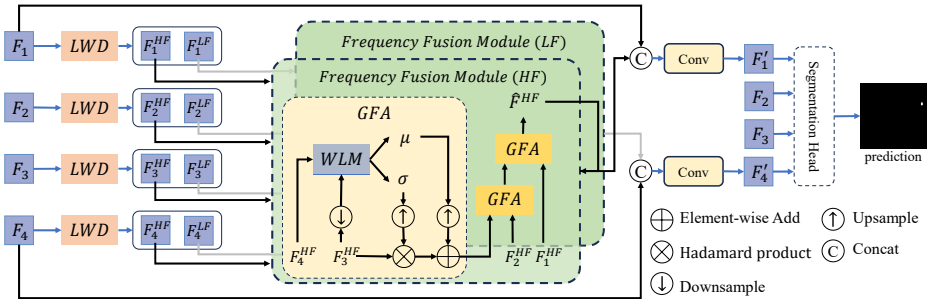

Fig. 4: The structure of the Wavelet-like Frequency Enhancement (WFE) module. It receives multi-level features $\{F_i\}_{i=1}^{4}$, outputting improved features F_1' and F_4'. The process initiates with the Learnable Wavelet-like Decomposition (LWD), which partitions the input features into their frequency constituents. Following this, the Frequency Fusion Module (FFM) integrates these separated frequency elements. The process concludes by merging the enhanced features with the original F_1 and F_4 to produce the final enhanced multi-level features. 'WLM' denotes the Window-based Linear Model, and GFA is Guidance-based Feature Aggregation modules.

is fed into two encoder layers, resulting in multi-level features $\{F_1, F_2, F_3, F_4\}$, where F_1 corresponds to F_{v1} from Visual Perception Head.

3.4 Wavelet-like Frequency Enhancement

The distinguishing features of tampered images, such as blending artifacts, are primarily found in the high-frequency domain. And, low-frequency components like color and illumination also contribute. Given the typically diminutive and subtle nature of tampered regions in manipulated documents, an approach that effectively addresses both types of information is imperative. Conventional document image tampering detection methods often utilize high-resolution features to capture small details but lack a targeted approach for their explicit enhancement. Our method, as illustrated in Figure 4, utilizes Wavelet-like Frequency Enhancement (WFE) to enhance document tampering detection by focusing on both high and low-frequency details.

To implement this, we use Learnable Wavelet-like Decomposition (LWD) to decompose multi-level features into their high and low-frequency components. These components are then fused using the Frequency Fusion Module (FFM) to create a unified representation for frequency enhancement. We further merge this enhanced representation with F_1 and F_4, which contain the most high-frequency details and larger scale information respectively. This strategy ensures a more precise and effective detection of subtle tampering traces.

Learnable Wavelet-like Decomposition. Our method employs the deep wavelet decomposition [23], a technique known for decomposing images or features into various frequency components. Specifically, we utilize the Haar wavelet transform, which is the simplest form of wavelet transform. Given the multi-scale

features $\{F_i\}_{i=1}^4$, the Haar wavelet transform breaks each down into four components: F_i^{LL}, F_i^{LH}, F_i^{HL}, and F_i^{HH}. However, our method only utilizes F_i^{LL} and F_i^{HH}, as they represent the two-dimensional low-frequency and high-frequency components, respectively. The low-frequency component, F_i^{LL}, encapsulates the image's overall structure, while the high-frequency component, F_i^{HH}, accentuates detailed textures, edges, and subtle clues. This is formally represented as:

$$F_i^{LL} = \left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} * F_i \right) \downarrow 2, \quad F_i^{HH} = \left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} * F_i \right) \downarrow 2, \tag{4}$$

where $*$ denotes the convolution operation, and $\downarrow 2$ denotes the down-sampling operation with a reduction factor of 2. Following decomposition, each component undergoes enhancement via an adaptive attention module, resulting in the enhanced low-frequency and high-frequency features F_i^{LF} and F_i^{HF}. This module integrates both channel and spatial attention mechanisms, focusing on the most informative regions of the feature map:

$$F_i^{LF} = \text{SA}(\text{CA}(\text{ConvBlock}(F_i^{LL}))), \quad F_i^{HF} = \text{SA}(\text{CA}(\text{ConvBlock}(F_i^{HH}))), \tag{5}$$

where SA and CA denote the spatial attention and channel attention respectively, and ConvBlock refers to a 3×3 convolution layer with batch normalization and ReLU activation. The final output, comprising eight components, offers a comprehensive image representation, crucial for detecting subtle tampering clues.

Frequency Fusion Module. We employ two Frequency Fusion Modules (FFM) to combine multi-scale frequency features ($\{F_i^{LF}\}_{i=1}^4$ and $\{F_i^{HF}\}_{i=1}^4$) to enhance document tampering detection. The FFM incorporates three Guidance-based Feature Aggregation (GFA) modules [11], which refine features by promoting inter-level feature interaction. The GFA uses a window-based linear model (WLM) [18] to align a lower-level feature F_{i-1} with its higher-level counterpart F_i, optimizing scaling and shifting parameters σ and μ, which describe the difference between these two features. These parameters are then seamlessly blended with F_{i-1}, then recursively applied to the next level, obtaining the frequency-enhanced representation $\hat{F}^{LF}$ and $\hat{F}^{HF}$ respectively. Finally, we leverage the high-resolution details preserved in F_1 by fusing it with the high-frequency enhanced feature map $\hat{F}^{HF}$ to further accentuate subtle tampering traces. Conversely, F_4 encapsulates broader scale information and is thus merged with the low-frequency enhanced feature map $\hat{F}^{LF}$ to capture low-frequency characteristics. Formally, the enhanced feature maps are computed as:

$$F_1' = \text{Conv}([\hat{F}^{LF}, F_1]), \quad F_4' = \text{Conv}([\hat{F}^{HF}, F_4]), \tag{6}$$

where Conv refers to a 1×1 convolution layer. The resulting multi-scale feature set $\{F_1', F_2, F_3, F_4'\}$, enhanced via WFE, is then processed through an FPN Head, to predict the segmentation map:

$$p = \text{SegHead}(\{F_1', F_2, F_3, F_4'\}), \tag{7}$$

where $p \in \mathbb{R}^{H \times W}$ denotes the predicted segmentation map.

3.5 Loss Function

Our segmentation model employs a composite loss function, defined as:

$$\mathcal{L} = \mathcal{L}_{CE}(p, y) + \mathcal{L}_{LS}(p, y), \tag{8}$$

where $\mathcal{L}_{CE}$, the Cross-Entropy loss, measures the pixel-wise classification accuracy, and $\mathcal{L}_{LS}$, the Lovasz-Softmax loss, directly targets the segmentation performance by focusing on the quality of the predicted boundaries. Here, p denotes the predicted segmentation maps, and y represents the corresponding ground truth.

4 Experiment

4.1 Experimental Setup

Datasets. The primary dataset employed for our study is DocTamper [22], comprising 170,000 tampered document images (contracts, invoices, and receipts in Chinese and English) using techniques like copy-move, splicing, and generation. The DocTamper dataset is partitioned into a training subset of 120,000 images and a same-domain test subset, DocTamper-T, with 30,000 images. Additionally, it includes cross-domain FCD and SCD datasets, containing 2,000 and 18,000 images respectively, which are derived from the Noisy Office Dataset [3] and Huawei Cloud [5]. To assess the model's generalization capabilities, we also evaluated on the T-SROIE [36] dataset, comprising 986 tampered images derived from the SROIE [12] dataset. This dataset primarily focuses on receipt scenarios and is tampered with using generative methods. This additional experiment allows us to gauge the model's performance in recognizing tampered text across a wider range of document contexts.

Evaluation Metrics. Consistent with established benchmarks [8, 14, 22], we assess our model using four key metrics: Precision, Recall, F1-score, and the Intersection over Union (IoU) for tampered regions. These metrics provide a comprehensive evaluation of our model's performance.

Implementation Details. Our model implementation was based on the MMSegmentation Toolbox [6] and trained utilizing 4 NVIDIA Tesla V100 GPUs. We use the pretrained ConvNextV2-base model as the backbone, supplemented with an FPN head for segmentation, and configured the input image size to 512x512. For optimization, we employed the AdamW optimizer with a learning rate of 0.0001, with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight decay of 0.05. We trained our model for 100K iterations with a batch size of 16, and using the poly learning rate policy with a power of 0.9. For testing, we followed the DTD [22] pipeline, where a JPEG compression is first applied to the image before it is fed into the network. The JPEG compression quality settings range from 75 to 100 to maintain consistency with DTD.

Table 1: Model comparison on three datasets. Adhering to the pipeline outlined by Qu *et al*. [22], each image in the test set has been compressed with a quality factor specified by their public repository. 'P' represents precision, 'R' stands for recall, and 'F' signifies the F1-score. The '*' marks instances where we have reimplemented the training code and retrained the models under identical settings. The best results are highlighted in **bold**, and the second-best results are marked with an underline.

	DocTamper-T				**DocTamper-FCD**				**DocTamper-SCD**			
	Iou	P	R	F	Iou	P	R	F	Iou	P	R	F
PSCC-Net [17]	0.17	0.25	0.83	0.39	0.13	0.19	0.82	0.30	0.11	0.15	**0.83**	0.25
UperNet [39]	0.70	0.66	0.60	0.62	0.30	0.57	0.35	0.43	0.48	0.57	0.58	0.57
CAT-Net [14]	0.78	0.75	0.69	0.72	0.66	0.85	0.70	0.76	0.58	0.65	0.65	0.65
Swin-UPer [18]	0.79	0.75	0.72	0.73	0.64	0.80	0.70	0.75	0.57	0.66	0.68	0.67
SegFormer [40]	0.81	0.77	0.74	0.75	0.69	0.82	0.74	0.78	0.61	0.68	0.70	0.69
Mask2Former [4]	0.84	0.82	0.83	0.82	0.66	0.81	0.75	0.78	0.59	0.70	0.79	0.74
ConvNext [19]	0.84	0.81	0.78	0.79	0.62	0.76	0.71	0.74	0.63	0.71	0.74	0.73
ConvNextV2 [37]	0.86	0.82	0.79	0.81	0.65	0.79	0.75	0.77	0.67	0.74	0.76	0.75
InternImage [33]	0.84	0.81	0.77	0.79	0.72	0.83	0.79	0.81	0.64	0.73	0.74	0.73
DTD* [22]	0.84	0.81	0.77	0.79	0.79	0.88	0.82	0.85	0.68	0.75	0.76	0.75
Ours	**0.90**	**0.87**	**0.84**	**0.86**	**0.88**	**0.93**	**0.91**	**0.92**	**0.75**	**0.81**	0.82	**0.81**

4.2 Quantitative and Qualitative Results

Our model's performance was rigorously evaluated against seven leading semantic segmentation frameworks and specialized tampering detection models, as detailed in Table 1. Among the competitors were UperNet [39], SegFormer [40], Swin-UPer [18], Mask2Former [4], ConvNext [19], ConvNextV2 [37] and InterImage [33], alongside tampering detection-specific models including PCSS-Net [17], CAT-Net [14], and DTD [22]. Notably, CAT-Net and DTD incorporate DCT information in addition to RGB data for tampering detection, while the other models rely solely on RGB. The competitive performance achieved by RGB-based models underscores the rationale behind our RGB-centric approach. Overall, our model outperformed others on the three DocTamper datasets, achieving a comprehensive improvement of 8.58% and 12.87% over DTD and ConvNextV2, respectively. Our method demonstrates balanced performance improvements across different test sets and performs particularly well on the FCD dataset, where most models show poor results. These results highlight the effectiveness of our VEM and WFE modules, validating their contribution.

Figure 5 visually compares the detection results of our model with other state-of-the-art methods across three datasets: DocTamper-T, FCD, and SCD. The visualizations clearly demonstrate our model's enhanced capability to detect subtle tampering features, thanks to the WFE module's heightened sensitivity to high-frequency details and the informative feature produced by the VEM. This qualitative evidence further validates the quantitative results, solidifying our model's position as a new benchmark in the field.

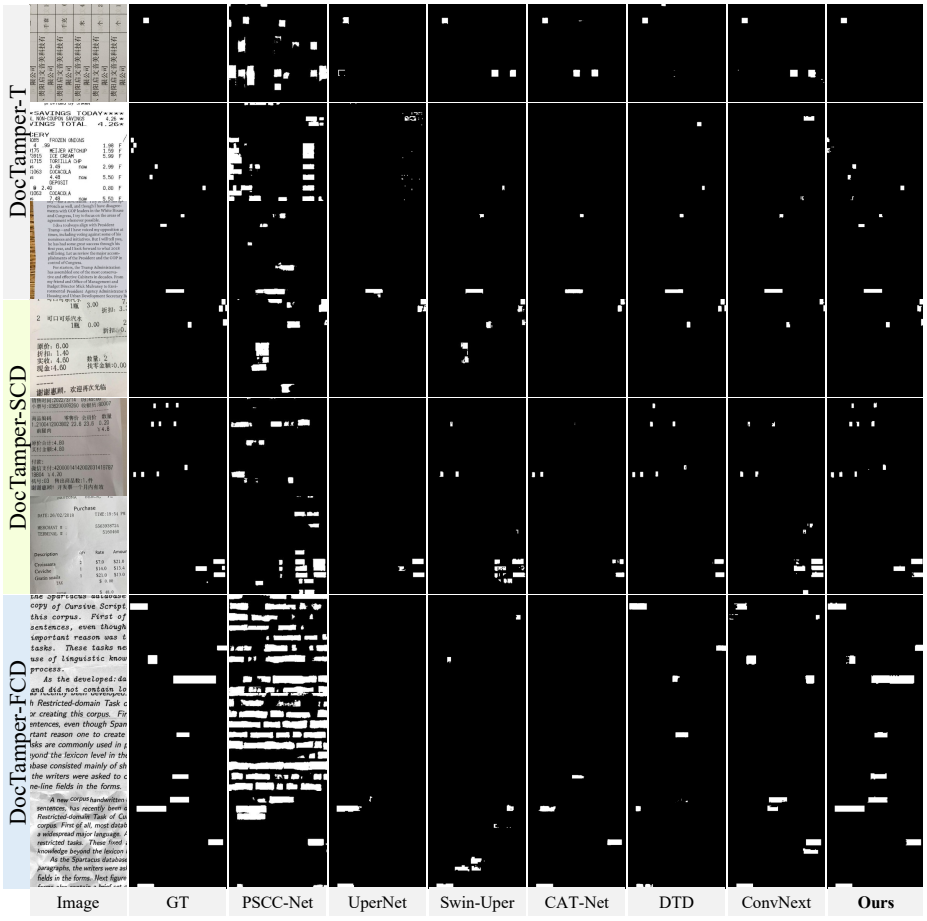

Fig. 5: Visualization of the detection results of our model and other state-of-the-art models on three datasets.

4.3 Ablation Study

In this subsection, we present an ablation study to evaluate the impact of each component in our framework. The study's findings, detailed in Table 2, confirm the effectiveness of our individual modules. The Frequency Perception Head [22], by incorporating DCT information, extends the model's capability from RGB-based tampering clues to detecting anomalies in the frequency domain, significantly improving tampering trace detection. Building upon the foundation provided by the FPH, our two innovative contributions—the VEM and WFE—further advance the model's performance. Each module independently contributes to incremental gains in detection accuracy and precision. When these modules are integrated, they operate synergistically, resulting in even more substantial performance improvements. This collaborative effect is manifested in the

Table 2: Ablation study of the proposed components. Adhering to DTD [22], each image in the test set has been compressed with a quality factor specified by the their public repository. FPH, VEM, and WFE denote the Frequency Perception Head, Visual Enhancement Module, and Wavelet-like Feature Enhancement, respectively.

			DocTamper-T				DocTamper-FCD				DocTamper-SCD			
FPH	VEM	WFE	IoU	P	R	F	IoU	P	R	F	IoU	P	R	F
			0.83	0.81	0.76	0.78	0.80	0.89	0.83	0.86	0.66	0.73	0.76	0.74
✓			0.86	0.84	0.81	0.83	0.83	0.91	0.89	0.90	0.73	0.77	0.79	0.78
✓	✓		0.88	0.86	0.83	0.84	0.87	0.92	0.90	0.91	0.74	0.80	0.80	0.80
✓		✓	0.88	0.85	0.82	0.84	0.84	0.90	0.89	0.90	0.73	0.80	0.80	0.80
✓	✓	✓	**0.90**	**0.87**	**0.84**	**0.86**	**0.88**	**0.93**	**0.91**	**0.92**	**0.75**	**0.81**	**0.82**	**0.81**

Table 3: Ablation of the Visual Enhancement Module on the DocTamper-T dataset.

	IoU	F1
(a) Atten → CBAM	0.877	0.838
(b) Atten → Effi-Atten	0.887	0.847
(c) Random Init Conv	0.882	0.844
(d) Max Fusion	0.881	0.844
Ours	**0.895**	**0.857**

Table 4: Ablation of the Wavelet-like Frequency Enhancement on the DocTamper-T dataset.

	IoU	F1
(a) LWT -> Fix DWT	0.867	0.827
(b) w/o GFA	0.883	0.843
(c) FFM w/o LF	0.892	0.854
(d) FFM w/o HF	0.874	0.833
Ours	**0.895**	**0.857**

improved metrics, demonstrating the combined efficacy of the VEM and WFE in elevating the model's tampering detection capabilities. The subsequent sections delve deeper into the contributions and interactions of each module.

Ablation of Visual Enhancement Module. In our ablation study, we scrutinize the design of our Visual Enhancement Module, which is crafted to introduce frequency information into spatial features. This module is crucial for improving the model's ability to detect subtle tampering without compromising the strengths of RGB-based detection. Table 3 presents the results of various adjustments to the fusion process and their effects on the model's performance. We considered the following modifications: (a) and (b) involve substituting the mechanism that enhances frequency information F_d^{enh}, with conventional methods such as CBAM [38] and Efficient Attention [27], respectively. These alterations lead to a decline in performance. (c) Moreover, we experimented with initializing the convolutional layers that process F_d^{enh} with random weights instead of zero weights. This change results in the RGB information losing its dominance in the model, which in turn adversely affects the benefits derived from pre-trained models, culminating in a decrease in performance. (d) Using the max function for fusion, instead of addition, was less effective. This method potentially discards valuable combined information, limiting the model's detection capabilities. These outcomes emphasize the importance of a carefully calibrated fusion pro-

Table 5: Model performance comparison of different backbones on three datasets.

Name	Backbone	DocTamper-T			DocTamper-FCD			DocTamper-SCD			Param.
		P	R	F	P	R	F	P	R	F	
ConvNextV2-Uper	ConvNextv2-B	0.82	0.79	0.81	0.79	0.75	0.77	0.74	0.76	0.75	121M
+DCT	ConvNextv2-B	0.84	0.81	0.83	0.91	0.89	0.90	0.77	0.79	0.78	126M
+Ours	ConvNextv2-B	0.87	0.84	0.86	0.93	0.91	0.92	0.81	0.82	0.81	140M
Swin-Uper	Swin-S	0.75	0.72	0.73	0.80	0.70	0.75	0.66	0.68	0.67	81M
+DCT	Swin-S	0.81	0.77	0.79	0.87	0.81	0.84	0.73	0.76	0.74	85M
+Ours	Swin-S	0.84	0.81	0.82	0.90	0.86	0.88	0.77	0.77	0.77	100M

cess. They validate our original approach to VEM, which successfully leverages the complete range of information to enhance tamper detection.

Ablation of Wavelet-like Frequency Enhancement. Our study delves into WFE module's role in accentuating high and low-frequency details for tampering detection. Table 4 outlines the performance impact of various WFE configurations: (a) Initially, we utilized the direct outputs of the deep wavelet decomposition, F_i^{LL} and F_i^{HH}, as features without any further enhancement through spatial or channel attention. (b) We employed upsampling and 1x1 convolution to fuse the multi-scale frequency features instead of using FFM. (c) and (d) We selectively enhanced either the high or low-frequency features within FFM. The results showed that removing the enhancement of high and low-frequency features or replacing the FFM led to a performance drop. Enhancing only the high-frequency information yielded good results, highlighting the importance of high-frequency information in document tampering detection. Further inclusion of low-frequency information led to additional performance improvement. These insights confirm the value of our WFE approach in improving tampering detection through meticulous frequency detail enhancement.

Dependency on Backbone. Our experimental analysis evaluates the adaptability of our method across different backbone architectures, as detailed in Table 5. We tested the FFDN not only with the ConvNextV2-B [37] but also with the Swin Transformer-S [18], which has fewer parameters and is used by the DTD [22].Our experiments reveal that the integration of an FPH to introduce frequency components results in significant performance gains without a substantial increase in parameter count. Further improvements are achieved with our proposed methods, leading to a moderate parameter growth while maintaining efficiency. These consistent performance gains, even with the addition of DCT components, highlight the FFDN's flexibility and effectiveness across various backbone architectures.

4.4 Robustness and Generalization Analysis

Our robustness analysis involved evaluating our model's performance under various JPEG compression levels. As shown in Table 6, our model demonstrated robustness across a wide range of JPEG compression quality settings, while DTD dropped significantly in performance due to high reliance on DCT information.

Table 6: The performance of various models under different JPEG compression quality on DocTamper-T dataset. The '*' marks instances where we have reimplemented the training code and retrained the models under identical settings.

	Q100				**Q90**				**Q75**				**Avg. Drop**	
	IoU	P	R	F	IoU	P	R	F	IoU	P	R	F	Q90	Q75
Swin-Uper	0.89	0.89	0.90	0.90	0.82	0.80	0.77	0.78	0.72	0.68	0.62	0.65	11.4%	25.4%
ConvNextV2	0.92	0.91	0.91	0.91	0.89	0.87	0.84	0.85	0.79	0.77	0.70	0.74	5.5%	**17.8%**
DTD*	0.93	0.94	0.96	0.95	0.90	0.89	0.86	0.87	0.75	0.69	0.61	0.65	6.8%	28.5%
Ours	**0.95**	**0.96**	**0.97**	**0.96**	**0.93**	**0.92**	**0.90**	**0.91**	**0.82**	**0.78**	**0.71**	**0.75**	**4.6%**	20.2%

Table 7: Robustness evaluation on DocTamper-T Dataset(F1). All distortions compressed at 90% quality except random JPEG.

Model	Gaussian Noise	Gaussian Blur	Resize 1.5X	Resize 0.75X	Color Jitter
DTD	0.71	0.69	0.78	0.69	0.71
Ours	0.84	0.79	0.85	0.78	0.81

Table 8: Comparison public on T-SROIE dataset. 'P', 'R', and 'F' denote precision, recall, and F1-score, respectively.

Method	P	R	F
EAST [47]	0.919	0.896	0.908
ATRR [34]	0.947	0.925	0.936
Wang *et al.* [36]	0.961	0.976	0.968
Ours	0.993	0.993	0.993

Additionally, as shown in Table 7, we tested various perturbation methods, such as Gaussian Noise and Gaussian Blur, and our model also exhibited better performance.

Further, to verify our model's generalization ability, we retrained and tested it on the T-SROIE dataset. The results, detailed in Table 8, demonstrate its effectiveness. To accommodate the higher resolution images in T-SROIE, we utilized a sliding window approach during testing, which allowed for segment processing and subsequent fusion into a cohesive output. These findings confirm our model's strong generalization, proving its efficacy in detecting tampered text across various document types and conditions.

5 Conclusion

In this paper, we introduce a novel approach for Document Image Tampering Detection (DITD), the Feature Fusion and Decomposition Network (FFDN). Our method integrates frequency and RGB features and enhances high-frequency details, addressing key DITD challenges. The Visual Enhancement Module leads the detection process, introducing the frequency features to the spatial domain through attention while maintaining the integrity of the original RGB detection capabilities. The Wavelet-like Frequency Enhancement module explicitly decomposes features into high- and low-frequency components to fully exploit high-frequency features with subtle tampering traces. Extensive experiments on the DocTamper dataset validate our approach, significantly outperforming current methods and advancing DITD.

Acknowledgements

This work was supported by National Science and Technology Major Project (No. 2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of Chine a (No. U21B2037, No. U22B2051, No. U23A20383, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2022J06001).

References

1. Anderson, J.C., Closen, M.L.: Document authentication in electronic commerce: the misleading notary public analog for the digital signature certification authority. J. Marshall J. Computer & Info. L. **17**, 833 (1998)
2. Bappy, J.H., Simons, C., Nataraj, L., Manjunath, B., Roy-Chowdhury, A.K.: Hybrid lstm and encoder–decoder architecture for detection of image forgeries. IEEE Transactions on Image Processing **28**(7), 3286–3300 (2019)
3. Castro-Bleda, M.J., Espana-Boquera, S., Pastor-Pellicer, J., Zamora-Martínez, F.: The noisyoffice database: A corpus to train supervised machine learning filters for image processing. The Computer Journal **63**(11), 1658–1667 (2020)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
5. Cloud, H.: Huawei cloud visual information extraction competition (2021)
6. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation` (2020)
7. Cruz, F., Sidere, N., Coustaty, M., d'Andecy, V.P., Ogier, J.M.: Local binary patterns for document forgery detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1223–1228. IEEE (2017)
8. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3539–3553 (2022)
9. Guo, Z., Yang, G., Chen, J., Sun, X.: Exposing deepfake face forgeries with guided residuals. IEEE Transactions on Multimedia **25**, 8458–8470 (2023)
10. Hao, J., Zhang, Z., Yang, S., Xie, D., Pu, S.: Transforensics: image forgery localization with dense self-attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15055–15064 (2021)
11. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22046–22055 (June 2023)
12. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019)

13. Joren, H., Gupta, O., Raviv, D.: Learning document graphs with attention for image manipulation detection. In: International Conference on Pattern Recognition and Artificial Intelligence. pp. 263–274. Springer (2022)
14. Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C.: Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision **130**(8), 1875–1895 (2022)
15. Kwon, M.J., Yu, I.J., Nam, S.H., Lee, H.K.: Cat-net: Compression artifact tracing network for detection and localization of image splicing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 375–384 (2021)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
17. Liu, X., Liu, Y., Chen, J., Liu, X.: Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 7505–7517 (2022)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
20. Nandanwar, L., Shivakumara, P., Mondal, P., Raghunandan, K.S., Pal, U., Lu, T., Lopresti, D.: Forged text detection in video, scene, and document images. IET Image Processing **14**(17), 4744–4755 (2020)
21. Nguyen, V., Blumenstein, M.: An application of the 2d gaussian filter for enhancing feature extraction in off-line signature verification. In: 2011 International Conference on Document Analysis and Recognition. pp. 339–343. IEEE (2011)
22. Qu, C., Liu, C., Liu, Y., Chen, X., Peng, D., Guo, F., Jin, L.: Towards robust tampered text detection in document image: New dataset and new solution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5937–5946 (2023)
23. Rodriguez, M.X.B., Gruson, A., Polania, L., Fujieda, S., Prieto, F., Takayama, K., Hachisuka, T.: Deep adaptive wavelet network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3111–3119 (2020)
24. Roy, A.G., Navab, N., Wachinger, C.: Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE transactions on medical imaging **38**(2), 540–549 (2018)
25. Roy, P., Bag, S.: Detection of handwritten document forgery by analyzing writers' handwritings. In: International conference on pattern recognition and machine intelligence. pp. 596–605. Springer (2019)
26. Shao, H., Huang, K., Wang, W., Huang, X., Wang, Q.: Progressive supervision for tampering localization in document images. In: International Conference on Neural Information Processing. pp. 140–151. Springer (2023)
27. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3531–3539 (2021)
28. Sirajudeen, M., Anitha, R.: Forgery document detection in information management system using cognitive techniques. Journal of Intelligent & Fuzzy Systems **39**(6), 8057–8068 (2020)

29. Van Beusekom, J., Shafait, F., Breuel, T.M.: Text-line examination for document forgery detection. International Journal on Document Analysis and Recognition (IJDAR) **16**, 189–207 (2013)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Verdoliva, L.: Media forensics and deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing **14**(5), 910–932 (2020)
32. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2364–2373 (2022)
33. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14408–14419 (2023)
34. Wang, X., Jiang, Y., Luo, Z., Liu, C.L., Choi, H., Kim, S.: Arbitrary shape scene text detection with adaptive text region representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6449–6458 (2019)
35. Wang, Y., Xie, H., Xing, M., Wang, J., Zhu, S., Zhang, Y.: Detecting tampered scene text in the wild. In: European Conference on Computer Vision. pp. 215–232. Springer (2022)
36. Wang, Y., Zhang, B., Xie, H., Zhang, Y.: Tampered text detection via rgb and frequency relationship modeling. Chinese Journal of Network and Information Security **8**(3), 29–40 (2023)
37. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
38. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
39. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
40. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
41. Xu, W., Luo, J., Zhu, C., Lu, W., Zeng, J., Shi, S., Lin, C.: Document images forgery localization using a two-stream network. International Journal of Intelligent Systems **37**(8), 5272–5289 (2022)
42. Yang, C., Wang, Z., Shen, H., Li, H., Jiang, B.: Multi-modality image manipulation detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
43. Yanikoglu, B., Kholmatov, A.: Online signature verification using fourier descriptors. EURASIP Journal on Advances in Signal Processing **2009**, 1–13 (2009)
44. Yul, H., ZHANG, T., ZHU, W., ZHANG, L., et al.: High-resolution noise artifact tracking network for image splicing forgery detection. Journal of Information Science & Engineering **39**(4) (2023)

45. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
46. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1053–1061 (2018)
47. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 5551–5560 (2017)
48. Zhuang, P., Li, H., Tan, S., Li, B., Huang, J.: Image tampering localization using a dense fully convolutional network. IEEE Transactions on Information Forensics and Security **16**, 2986–2999 (2021)