Frequency Mining Empowered by Text Aggregation: A New Perspective on Document Image Tampering Detection

Ziqi Yi*, Guitao Xu*, Shihang Wu, Peirong Zhang, Lianwen Jin†

South China University of Technology

{eezqyi, eegtxu, eeshihang_wu, eeprzhang}@mail.scut.edu.cn, eelwjin@scut.edu.cn

Abstract

Document image tampering detection faces significant challenges due to the subtle and spatially dispersed nature of tampering traces, which are often confined to localized regions within tampered text. While existing methods leverage frequency domain information to reveal hidden artifacts, they fail to fully exploit the rich frequency spectrum and lack effective mechanisms for aggregating scattered tampering evidence across extended text regions. To overcome these limitations, we propose the **T**ext **A**ggregation and multi-**F**requency **E**nhancement **Net**work (**TAFE-Net**). Specifically, to capture more subtle tampering traces, we design a Multi-Frequency Feature Extractor that comprehensively utilizes various proven effective frequency information. In addition, the Visual-Frequency Integration Module and Direction-aware Frequency Decoupling Enhancement module are introduced to aggregate text features in both horizontal and vertical directions within the frequency domain, from coarse to fine granularity, addressing the incomplete detection of tampered text caused by dispersed tampering traces. Experiments on the DocTamper and RTM datasets demonstrate that our approach establishes new state-of-the-art results and maintains superior robustness against various degradations.

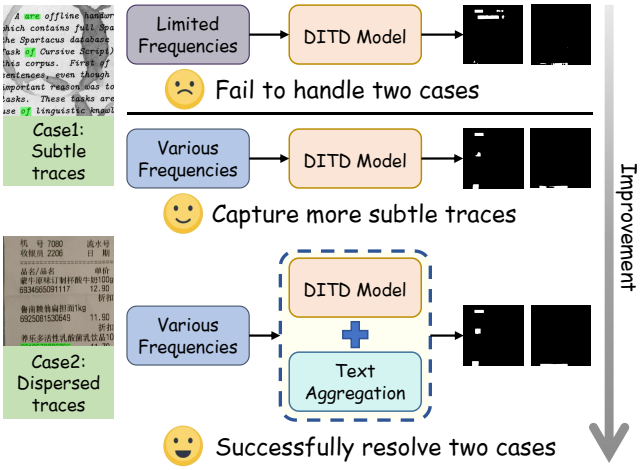

Figure 1: Comparison with existing methods. Our method advances the state of the art in two ways: (1) Introducing rich frequency information to capture more subtle tampering traces. (2) Aggregating text features to address the problem of dispersed forgery artifacts in tampered text.

Introduction

Document images serve as essential information carriers in modern society, supporting digital workflows, legal record-keeping, and paperless practices. However, the swift evolution of image editing technologies (Huang et al. 2025; Zhang et al. 2025) has facilitated the easy creation of visually imperceptible document forgeries, threatening information integrity (Nandanwar et al. 2020; Shao et al. 2024). Consequently, developing robust document image tampering detection (DITD) methods (Xu et al. 2025) has become crucial for defending against malicious manipulations.

In recent years, natural image forgery detection (NIFD) (Dong et al. 2022) has witnessed remarkable progress. However, these methods often fail in DITD, as tampered documents exhibit unique characteristics, including smaller tampering regions and high consistency between the tampered text and background. Although existing DITD methods (Qu et al. 2023) have achieved promising results through utilizing frequency-based modeling approaches to enhance visual features, they have not fully exploited the potential of frequency modeling. As shown in Fig. 1, this potential is primarily manifested in two key aspects: **(1) Insufficient utilization of various frequency information for capturing more subtle tampering traces.** The frequency information of images is rich and contains important traces of tampering, yet existing methods mostly utilize only limited frequency components, such as JPEG DCT (Discrete Cosine Transform) coefficients (Qu et al. 2023), high and low frequencies (Zhu et al. 2025), or wavelet transform frequency enhancement (Chen et al. 2024; Zhang and Jin 2024). Since these various frequency components have all proven useful, exploring their integration into visual features to reveal more subtle tampering artifacts would be significantly valuable. **(2) Absence of effective consideration for aggregating dispersed tampering traces of tampered text in the frequency domain.** Document images commonly contain text with extreme aspect ratios. When text is tampered with,

*These authors contributed equally.

† Corresponding author.

especially long text, tampering traces are typically dispersed and manifest only in localized portions of the tampered text, challenging the detection of the entire tampered content. Existing methods overlook this issue, making it difficult to capture direction-specific (vertical or horizontal) features and textures, such as text features and edge-related details. Therefore, they often perform poorly when confronted with slender text tampering.

Based on this observation, we propose **T**ext **A**ggregation and multi-**F**requency **E**nhancement **Net**work (**TAFE-Net**) to exploit more hidden frequency information and aggregate text features. It builds upon SegFormer (Xie et al. 2021) and introduces three novel modules. To capture more subtle tampering traces, we propose a Multi-Frequency Feature Extractor (MFFE) that extracts and fuses multiple types of frequency information to obtain comprehensive and robust frequency features. In parallel, we design a Visual-Frequency Integration Module (VFIM) and a Direction-aware Frequency Decoupling Enhancement (DFDE) module to address the challenge of typically localized and dispersed tampering traces in tampered text. Specifically, VFIM not only fuses the RGB image with its corresponding frequency components at the model input stage, but also aggregates text features at a coarse granularity, endowing the tampered text with global coherence. Meanwhile, DFDE performs a more refined aggregation of text features for amplification of orientation-specific tampering cues at the decoding stage. It utilizes specifically designed convolutions to process the four subbands decoupled by wavelet transformation, thereby better aggregating detail features of different subbands, especially the text features and edge artifacts residing in vertical (LH) and horizontal (HL) high-frequency subbands.

In summary, our contributions are as follows:

- We propose the TAFE-Net model for DITD, which successfully aggregates text features to address the problem of incomplete detection of slender text tampering, and integrates numerous frequency information to explore more subtle tampering traces.
- We are the first to observe that tampering traces are dispersed yet restricted to localized regions of the tampered text. Building on this insight, we design two dedicated modules, VFIM and DFDE, to aggregate text features and directional details from coarse to fine in the frequency domain, thereby boosting detection performance.
- Through extensive experiments, we demonstrate that our method significantly outperforms previous state-of-the-art approaches while exhibiting strong robustness.
- We find that Transformer architectures offer distinct advantages in more challenging DITD scenarios, delivering markedly stronger robustness across domains.

Related Work

Natural Image Forgery Detection

Recently, many deep learning-based NIFD methods (Dong et al. 2022; Qu et al. 2024) have demonstrated powerful performance. These methods primarily leverage CNN or Transformer backbones to extract local and global tampering features, respectively. For example, ManTra-Net (Wu, AbdAlmageed, and Natarajan 2019) and CAT-Net (Kwon et al. 2022) employed VGG (Simonyan and Zisserman 2015) and HRNet (Sun et al. 2019), respectively, to effectively extract local anomalies in the noise and frequency domains. In contrast, works such as IML-ViT (Ma et al. 2024), TruFor (Guillaro et al. 2023), and SparserViT (Su et al. 2025) utilized the powerful contextual awareness capabilities of Transformer architectures to capture global inconsistencies in tampered images. Moreover, some works (Wang et al. 2022a; Zhu et al. 2025) adopted hybrid CNN-Transformer models to combine the advantages of both architectures. However, these methods often face significant challenges when extended to the DITD task, as forged documents typically feature smaller tampered regions and high consistency between text and background elements.

Document Image Tampering Detection

In the context of DITD, researchers (Wang et al. 2022b; Qu et al. 2023; Chen et al. 2024) have attempted to utilize frequency information to enhance visual features for detecting visually imperceptible tampered text. Qu et al. (2023) proposed DTD, which leveraged DCT coefficients to identify inconsistencies in image compression artifacts, thereby enhancing model performance. Building upon this work, FFDN (Chen et al. 2024) further incorporated Wavelet-like Frequency Enhancement during the decoding process to explicitly retain high-frequency details. Although these approaches yield notable improvements, they still fail to fully leverage the comprehensive frequency information available, such as frequencies obtained through various transforms (Zhu et al. 2025) or directional (horizontal/vertical) frequency components. In addition, other methods (Luo et al. 2025) have explored more transformation domains (e.g., noise, error level analysis) to track tampering artifacts from different aspects. However, existing methods rarely consider the special structural characteristics of document images, particularly how text in documents typically exhibits strong directionality (e.g., horizontally aligned text lines or vertically arranged text columns). Such text displays greater aspect ratios, resulting in a scattered distribution of tampering traces, which makes tampering features difficult to aggregate. Therefore, in this paper, we advance the state-of-the-art in DITD through dual innovations: comprehensive utilization of various frequency information and enhanced aggregation of tampering features in slender text.

Method

The core insight driving our approach is that tampering traces in document images exhibit two distinct characteristics: they are both subtle (requiring comprehensive frequency analysis) and spatially dispersed (requiring systematic aggregation mechanisms). To address these challenges systematically, we propose the **T**ext **A**ggregation and multi-**F**requency **E**nhancement **Net**work (**TAFE-Net**) for DITD, as presented in Fig. 2. Building upon the SegFormer baseline (Xie et al. 2021), we introduce three novel modules: (1) Visual-Frequency Integration Module to preprocess and

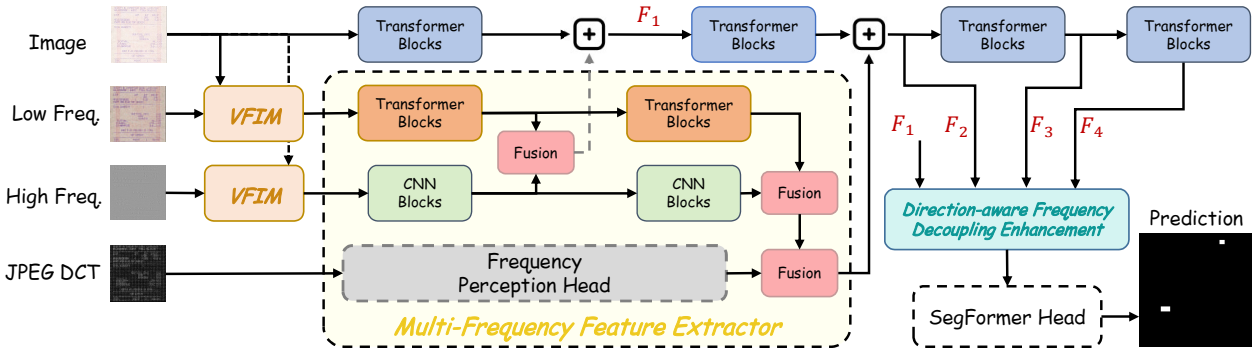

Figure 2: The overall architecture of the proposed TAFE-Net. It is built upon the SegFormer baseline, introducing three novel modules to enhance frequency features. VFIM stands for Visual-Frequency Integration Module.

enhance both high-frequency and low-frequency representations, (2) Multi-Frequency Feature Extractor to integrate various frequency features, and (3) Direction-aware Frequency Decoupling Enhancement module to decouple and amplify orientation-specific frequency information.

Model Input

Tampered text is often subtle and consistent with authentic content and background, making it visually elusive. Prior studies (Qu et al. 2023; Zhu et al. 2025; Zhang, Ding, and Jin 2025) confirm that different frequency features contain evidence for detecting forgeries. To uncover more subtle tampering traces, we attempt to introduce more frequency features and fuse them with visual features, constructing more comprehensive tampering representations.

Given an RGB image $I_v \in \mathbb{R}^{H\times W\times 3}$, we convert it to YCrCb color space and compute the JPEG DCT coefficient map $D \in \mathbb{R}^{H\times W}$ and the quantization table $T \in \mathbb{R}^{8\times 8}$ from its Y channel. In parallel, we employ the discrete cosine transform (DCT) (Gonzalez 2009) to separate it into high-frequency view $I_{hf} \in \mathbb{R}^{H\times W\times 3}$ and low-frequency view $I_{lf} \in \mathbb{R}^{H\times W\times 3}$.

Visual-Frequency Integration Module

We concatenate both high-frequency and low-frequency views with the original image to enhance representations:

$$I_h = \mathrm{Cat}(I_v, I_{hf}), \quad I_h \in \mathbb{R}^{H\times W\times 6}, \tag{1}$$

$$I_l = \mathrm{Cat}(I_v, I_{lf}), \quad I_l \in \mathbb{R}^{H\times W\times 6}, \tag{2}$$

where $\mathrm{Cat}(.)$ means concatenate operation.

Then we design a Visual-Frequency Integration Module (VFIM) to convert I_h and I_l to three-channel representations to fully leverage the knowledge of the pre-trained backbone, as shown in Fig. 3. Specifically, VFIM executes dimensionality reduction operations through two complementary branches: the Direction-Sensitive Branch (DSB) and the Direction-Agnostic Branch (DAB).

Direction-Sensitive Branch. Horizontal text features can identify forgery traces such as font variations, abnormal character spacing, and inconsistent in-line styles, while vertical text features effectively detect changes in line spacing,

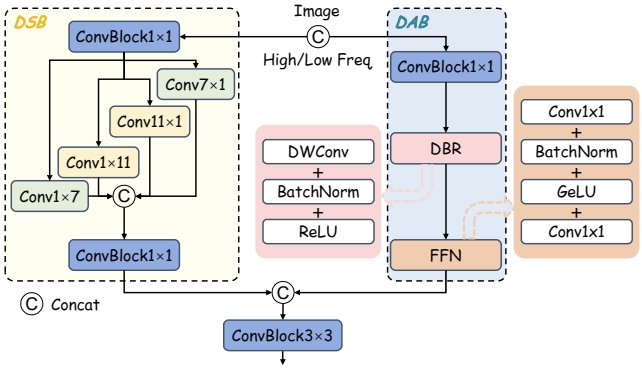

Figure 3: The proposed Visual-Frequency Integration Module. DSB and DAB denote the Direction-Sensitive Branch and Direction-Agnostic Branch, respectively. Conv$k_1 \times k_2$ denotes a convolution layer with kernel size $k_1 \times k_2$ and ConvBlock$k_1 \times k_2$ dentoes a $k_1 \times k_2$ convolution layer with batch normalization and ReLU activation.

vertical alignment anomalies, and paragraph structure inconsistencies. Therefore, we introduce the DSB module to aggregate text features at a coarse granularity from visual and frequency domains in both horizontal and vertical directions. Initially, a 1×1 ConvBlock performs dimensionality reduction on I_h and I_l separately. The features then pass through four asymmetric convolution layers to aggregate text features. Finally, the concatenated features are processed by another 1×1 ConvBlock to generate the final feature map.

Direction-Agnostic Branch. As a complement, the DAB module is used to extract direction-agnostic features in the document, such as local blurring and frequency domain anomalies. It consists of a 1×1 ConvBlock, a 3×3 depthwise separable convolution block, and a Feed-Forward Network.

Finally, feature maps from both branches are concatenated and fed into a 3×3 ConvBlock, producing new representations $\hat{I}_h$ and $\hat{I}_l$ with dimensions of $H \times W \times 3$. The process can be formulated as:

$$\hat{I}_h = \mathrm{ConvBlock}(\mathrm{Cat}(\mathrm{DSB}(I_h), \mathrm{DAB}(I_h))), \tag{3}$$

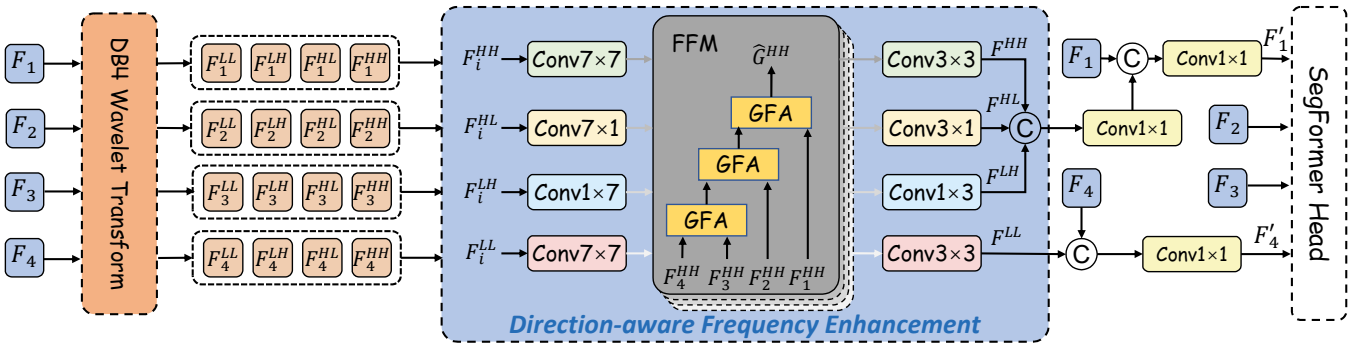

Figure 4: The structure of the Direction-aware Frequency Decoupling Enhancement (DFDE) module.

$$\hat{I}_l = \text{ConvBlock}(\text{Cat}(\text{DSB}(I_l), \text{DAB}(I_l))). \quad (4)$$

Multi-Frequency Feature Extractor

To effectively integrate different frequency information, we propose a Multi-Frequency Feature Extractor (MFFE), as illustrated in Fig. 2. The MFFE module employs a three-branch architecture to process and combine features across multiple frequency domains.

Frequency Encoding. First, we utilize the Frequency Perception Head (FPH) from DTD (Qu et al. 2023) to encode the JPEG DCT coefficient, enabling the identification and extraction of concealed spectral anomaly features within 8×8 transformation blocks. Taking the DCT coefficient map D and the quantization table T as input, the FPH outputs the DCT frequency representation $F_d \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_d}$ as follows:

$$F_d = \text{FPH}(D, T). \quad (5)$$

Drawing from established principles that high-frequency features contain fine-grained texture details while low-frequency features provide coarse-grained contextual information (Zhu et al. 2025), we employ architecture-specific encoders to maximize the extraction of relevant information from each domain. Specifically, we separately employ a CNN backbone and a Transformer backbone to encode $\hat{I}_h$ and $\hat{I}_l$, respectively, to fully exploit their complementary advantages in feature extraction:

$$\{F_{h1}, F_{h2}\} = \text{CNNBlocks}(\hat{I}_h), \quad F_{hi} \in \mathbb{R}^{\frac{H}{2^{(i+1)}}\times\frac{W}{2^{(i+1)}}\times C_{hi}}, \quad (6)$$

$$\{F_{l1}, F_{l2}\} = \text{TransformerBlocks}(\hat{I}_l), \quad F_{li} \in \mathbb{R}^{\frac{H}{2^{(i+1)}}\times\frac{W}{2^{(i+1)}}\times C_{li}}, \quad (7)$$

where C_{hi} and C_{li} represent the number of output channels for high-frequency and low-frequency features at each scale i, respectively.

Frequency Fusion. In the feature fusion stage, we concatenate the low-frequency and high-frequency features of corresponding resolutions and fuse them through scSE (Roy, Navab, and Wachinger 2018) modules to obtain the integrated representations $F_{hl1} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_1}$ and $F_{hl2} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$:

$$F_{hli} = \text{scSE}(\text{Cat}(F_{hi}, F_{li})), \quad i \in 1, 2. \quad (8)$$

Subsequently, the DCT frequency representation F_d undergoes the same fusion operation with F_{hl2}, yielding the comprehensive frequency feature F_{cf2}.

Visual Encoding and Fusion

Turning to visual encoding, we adopt SegFormer (Xie et al. 2021) as our baseline model. In the first stage of SegFormer, the RGB image I_v is encoded to extract the initial visual feature $F_v \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_1}$. Then, after adding F_v with F_{hl1}, the combined feature F_1 is fed into the second stage of SegFormer, resulting in the multi-modality feature $F_m \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$. Finally, the element-wise addition F_2 of F_m and F_{cf2} is input to the last two stages of SegFormer to extract two higher level features $F_3 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_3}$ and $F_4 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C_4}$.

Direction-aware Frequency Decoupling Enhancement

Existing methods (Chen et al. 2024) only utilize the low-frequency approximation and diagonal high-frequency detail components for frequency enhancement in the wavelet domain, neglecting high-frequency information in horizontal and vertical directions. However, these discarded subbands notably contain crucial features needed for document forgery detection, such as text characteristics and stripe artifacts. Therefore, we propose the Direction-aware Frequency Decoupling Enhancement (DFDE) module to fully utilize all frequency subbands, maximizing the extraction of directional features across the complete wavelet spectrum. The structure of DFDE is depicted in Fig. 4.

Frequency Decoupling. We first apply the Daubechies-4 (DB4) wavelet transform to each multi-scale feature $F_i, i \in \{1, 2, 3, 4\}$, obtaining four subbands: a low-frequency approximation F_i^{LL} that preserves global structure, and three high-frequency detail bands that capture vertical (F_i^{LH}), horizontal (F_i^{HL}), and diagonal (F_i^{HH}) information, respectively. All subbands are downsampled by a factor of 2 compared to the original feature F_i.

Method	Venue	Test				FCD				SCD				RTM			
		IoU	P	R	F1	IoU	P	R	F1	IoU	P	R	F1	IoU	P	R	F1
UPerNet	ECCV'18	70.0	66.0	60.0	62.0	30.0	57.0	35.0	43.0	48.0	57.0	58.0	57.0	8.26	32.5	10.0	15.3
Swin-UPer	ICCV'21	82.7	79.7	77.0	78.3	65.9	78.2	77.5	77.8	62.3	69.8	73.7	71.7	17.0	41.8	22.3	29.1
ConvNeXt	CVPR'22	83.0	78.7	76.5	77.6	62.5	77.3	68.7	72.7	61.5	71.0	70.8	70.9	16.2	35.8	22.8	27.9
SegFormer	NeurIPS'21	81.0	77.0	74.0	75.0	69.0	82.0	74.0	78.0	61.0	68.0	70.0	69.0	19.4	56.5	22.8	32.5
Mask2Former	NeurIPS'22	84.0	82.0	83.0	82.0	66.0	81.0	75.0	78.0	59.0	70.0	79.0	74.0	12.4	19.1	25.9	22.0
PSCC-Net	TCSVT'22	17.0	25.0	83.0	39.0	13.0	19.0	82.0	30.0	11.0	15.0	83.0	25.0	3.31	3.59	30.3	6.41
CAT-Net V2	IJCV'22	78.0	75.0	69.0	72.0	66.0	85.0	70.0	76.0	58.0	65.0	65.0	65.0	13.7	25.8	22.6	24.1
MVSS-Net ++	TPAMI'22	21.0	34.4	34.7	34.5	16.4	42.4	28.2	33.9	20.6	30.8	40.9	35.1	4.27	17.1	5.39	8.20
Mesorch	AAAI'25	35.4	50.4	45.8	48.0	50.1	76.9	53.3	63.0	24.3	38.7	26.1	31.2	12.1	19.9	23.6	21.6
DTD	CVPR'23	84.0	81.0	77.0	79.0	79.0	88.0	82.0	85.0	68.0	75.0	76.0	75.0	6.51	11.9	12.5	12.2
FFDN	ECCV'24	89.3	87.4	84.0	85.6	87.6	91.8	90.9	91.3	75.2	80.8	82.1	81.4	10.9	16.9	23.5	19.7
ASC-Former	PR'25	81.2	77.8	72.0	74.8	69.9	82.5	73.9	78.0	61.2	69.2	68.4	68.8	19.7	50.4	24.4	32.9
TAFE-Net (Ours)	**This Work**	89.4	87.1	84.7	85.9	**89.3**	**92.9**	**91.7**	**92.3**	74.0	78.6	82.9	80.7	**23.1**	**60.6**	27.2	**37.5**
TAFE-Net* (Ours)	**This Work**	**90.3**	**88.5**	**85.3**	**86.8**	88.6	92.4	91.5	91.9	**75.9**	**80.9**	**83.4**	**82.1**	17.8	29.9	**30.6**	30.2

Table 1: Comparison on the DocTamper dataset and the RTM dataset. Training and evaluation are conducted within the same dataset only. Models marked with * denote that we employ ConvNeXt V2 (Woo et al. 2023) as the baseline instead of SegFormer (Xie et al. 2021). Each image in the DocTamper test set has been compressed with a quality factor specified by the public repository of DTD (Qu et al. 2023). 'P', 'R', and 'F1' denote precision, recall, and F1-score, respectively. The best results are highlighted in **bold** and the second-best results are underlined.

Direction-aware Frequency Enhancement. Then we construct a Direction-aware Frequency Enhancement (DFE) module designed to effectively aggregate multi-scale frequency subband features. Specifically, we use different convolution kernels for various frequency subbands: standard square convolution layers for low-frequency subband F_i^{LL} and diagonal high-frequency subband F_i^{HH}, while directional asymmetric convolution layers for vertical high-frequency subband F_i^{LH} and horizontal high-frequency subband F_i^{HL}. This design offers significant advantages as the directional convolution layers align perfectly with the single-orientation characteristics of LH and HL subbands, thereby enhancing text feature aggregation and more precisely capturing the corresponding edge artifacts.

For each type of subband, we utilize the Frequency Fusion Module (FFM) from FFDN (Chen et al. 2024), which consists of three cascaded Guidance-based Feature Aggregation (GFA) modules (He et al. 2023), to effectively consolidate features from four different scales into a more robust unified representation. This fused feature is subsequently processed by specifically designed convolution layers. The entire process can be described as follows:

$$F^{LL} = \text{Conv}_{(3,3)}\left(\text{FFM}\left(\{\text{Conv}_{(7,7)}(F_i^{LL})\}_{i=1}^4\right)\right), \quad (9)$$

$$F^{LH} = \text{Conv}_{(1,3)}\left(\text{FFM}\left(\{\text{Conv}_{(1,7)}(F_i^{LH})\}_{i=1}^4\right)\right), \quad (10)$$

$$F^{HL} = \text{Conv}_{(3,1)}\left(\text{FFM}\left(\{\text{Conv}_{(7,1)}(F_i^{HL})\}_{i=1}^4\right)\right), \quad (11)$$

$$F^{HH} = \text{Conv}_{(3,3)}\left(\text{FFM}\left(\{\text{Conv}_{(7,7)}(F_i^{HH})\}_{i=1}^4\right)\right), \quad (12)$$

where $\text{Conv}_{(k_1\times k_2)}$ denotes a convolution layer with kernel size $k_1 \times k_2$.

After obtaining the representations F^{LL}, F^{LH}, F^{HL}, and F^{HH} for different frequency subbands, we first concatenate the three high-frequency subband features (F^{LH}, F^{HL}, F^{HH}) and pass them through a 1×1 convolution layer to derive the final high-frequency feature F^H:

$$F^H = \text{Conv}_{(1,1)}(\text{Cat}(F^{LH}, F^{HL}, F^{HH})). \quad (13)$$

The high-frequency feature F^H contains rich fine-grained structures such as edges and contours, which are crucial for capturing tampering traces. Therefore, we fuse it with the low-level feature F_1 to enhance detail perception. In contrast, the low-frequency feature F^{LL} carries more comprehensive global semantic information, significantly improving the model's robustness. Consequently, we combine it with the high-level feature F_4 to strengthen the model's resistance against complex interferences. The fusion operations are expressed as:

$$F_1' = \text{Conv}_{(1,1)}(\text{Cat}(F^H, F_1)), \quad (14)$$

$$F_4' = \text{Conv}_{(1,1)}(\text{Cat}(F^{LL}, F_4)). \quad (15)$$

The enriched multi-scale feature set $\{F_1', F_2, F_3, F_4'\}$ is subsequently decoded by a SegFormerHead to predict the final result $M_p \in \mathbb{R}^{H\times W}$, that is

$$M_p = \text{SegFormerHead}(F_1', F_2, F_3, F_4'). \quad (16)$$

Loss Function

During the training phase, we employ the Cross-Entropy Loss and the Lovasz Loss as the optimization function:

$$\mathcal{L} = \mathcal{L}_{ce}(M_p, M_g) + \mathcal{L}_{lov}(M_p, M_g), \quad (17)$$

where M_g represents the ground-truth mask.

Experiments

Experimental Setup

Datasets. We evaluate our model on the large-scale synthetic dataset Doctamper (Qu et al. 2023), which provides one in-domain test set (*Test*) and two cross-domain test sets (*FCD* and *SCD*), and the handcrafted dataset RTM (Luo et al. 2025).

#Line	VFIM	MFFE	DFDE	Test				FCD				SCD				AVG			
				IoU	P	R	F1	IoU	P	R	F1	IoU	P	R	F1	IoU	P	R	F1
1				79.3	75.1	71.7	73.4	69.8	81.1	75.4	78.2	57.4	64.6	66.2	65.4	68.8	73.6	71.1	72.3
2		✓	✓	87.6	85.3	82.1	83.7	87.9	91.6	91.4	91.5	72.4	78.1	79.8	79.0	82.6	85.0	84.4	84.7
3			✓	83.7	79.3	76.5	77.9	73.1	82.0	79.2	80.6	61.9	70.5	72.8	71.6	72.9	77.3	76.2	76.7
4	✓	✓		85.1	82.5	80.2	81.3	85.9	90.3	90.6	90.4	69.3	73.3	78.0	75.6	80.1	82.0	82.9	82.4
5	✓	✓	✓	**89.4**	**87.1**	**84.7**	**85.9**	**89.3**	**92.9**	**91.7**	**92.3**	**74.0**	**78.6**	**82.9**	**80.7**	**84.2**	**86.2**	**86.4**	**86.3**

Table 2: Ablation study of the proposed modules on the DocTamper dataset. VFIM, MFFE, and DFDE denote the Visual-Frequency Integration Module, Multi-Frequency Feature Extractor, and Direction-aware Frequency Decoupling Enhancement module, respectively. *AVG* denotes the mean performance computed across the three DocTamper subsets.

Implementation Details. While training, we resize or crop the images to 512×512. For optimization, we implement the AdamW optimizer with a learning rate of 0.0001, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay coefficient of 0.05. The model is trained for a total of $100k$ iterations with a batch size of 16, employing a poly learning rate policy with a power of 0.9.

Metrics. Following previous works (Chen et al. 2024; Luo et al. 2025), we employ Precision, Recall, F1-score, and Intersection over Union (IoU) as evaluation metrics.

Comparison with State-of-the-Art Methods

We compare our method with five general semantic segmentation methods (Xiao et al. 2018; Liu et al. 2021; Xie et al. 2021; Liu et al. 2022b; Cheng et al. 2022), four natural image forgery detection methods (Liu et al. 2022a; Kwon et al. 2022; Dong et al. 2022; Zhu et al. 2025), and three document image tampering detection methods (Qu et al. 2023; Chen et al. 2024; Luo et al. 2025), as shown in Table 1. The results show that our proposed TAFE-Net achieves the best performance on three test sets except for DcTamper-SCD, with notable improvements of 3.4 and 4.6 points in IoU and F1-score metrics, respectively, on the RTM dataset. Moreover, we observe that purely frequency-based models (e.g., Mesorch) perform poorly across multiple datasets, while purely visual models (e.g., Swin-UPer and ConvNeXt) show insufficient generalization ability in cross-domain testing. This indicates that frequency and visual information are complementary and both essential for DITD. Finally, the qualitative visual comparison of tampering localization results is presented in Fig. 5. Our model significantly outperforms other models, particularly when handling slender text tampering and subtle forgery traces.

We further implement TAFE-Net*, a CNN variant that replaces the SegFormer baseline with ConvNeXt V2 (Woo et al. 2023), and uncover an important trend: CNN-based models (e.g., ConvNeXt, FFDN, and our TAFE-Net*) generally excel on DocTamper but underperform on RTM, whereas Transformer-based models (e.g., Swin-UPer, SegFormer, and our TAFE-Net) exhibit the inverse pattern. Notably, the latter display higher robustness, evidenced by the smaller performance drop from DocTamper-Test to DocTamper-FCD. This indicates that the global modeling capability of Transformers is particularly crucial in more challenging DITD scenarios, providing important insights for subsequent researchers in framework design.

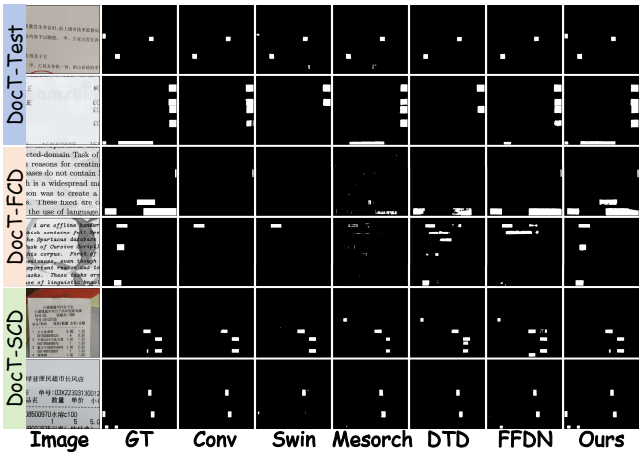

Figure 5: Qualitative results on DocTamper of comparing our model with state-of-the-art methods. 'Conv' and 'Swin' refer to ConvNeXt and Swin-UPer, respectively.

Ablation Study

To evaluate the contribution of our proposed components, we conduct comprehensive ablation studies on the Doctamper dataset, with results detailed in Table 2.

Visual-Frequency Integration Module. Comparing lines 2 and 5 in Table 2, we observe a clear performance gain after VFIM is incorporated into TAFE-Net: the average IoU and F1-score each increase by 1.6 points. This improvement can be attributed to two key design choices in VFIM: (1) Channel dimensionality reduction enables the full utilization of pre-trained knowledge from the selected backbone. (2) The DSB module is capable of aggregating horizontal and vertical text features, which are vital for detecting tampered text in document images.

Multi-Frequency Feature Extractor. As demonstrated in line 3 of Table 2, removing the entire MFFE module leads to substantial degradation in both the model's effectiveness and its ability to generalize. To assess the impact of different frequency components on our model, we further perform ablation studies by selectively removing them. The results are reported in Table 3, which reveal that eliminating any single frequency component (high-frequency, low-frequency, or

High Freq.	Low Freq.	DCT	Test	FCD	SCD	AVG
			83.7	73.1	61.9	72.9
	✓	✓	88.1	88.1	72.6	82.9
✓		✓	88.1	88.4	72.9	83.1
✓	✓		84.3	76.6	67.3	76.1
		✓	87.3	87.8	71.6	82.2
✓	✓	✓	**89.4**	**89.3**	**74.0**	**84.2**

Table 3: Ablation study about different frequency components on the DocTamper dataset. *Freq.* denotes Frequency. All reported scores are evaluated with the IoU metric.

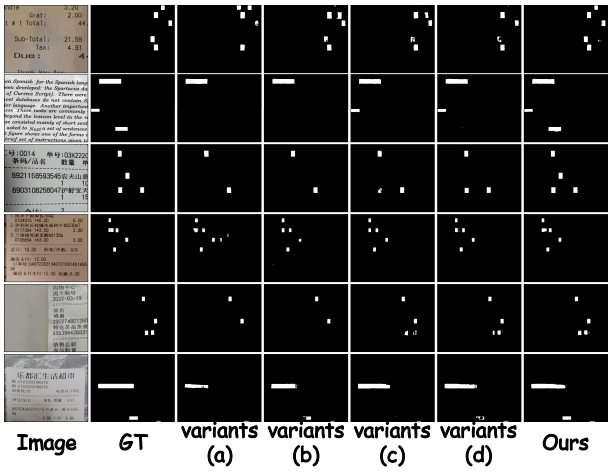

Figure 6: We visually compare DFDE and its variants. DFDE shows major advantages in detecting slender tampered text and handling easily misclassified authentic text.

DCT coefficients) leads to performance degradation. Among these, DCT coefficients prove to be the most critical component, while high-frequency and low-frequency components have roughly equal influence.

Direction-aware Frequency Decoupling Enhancement. The results in lines 4 and 5 of Table 2 show that introducing DFDE yields a substantial performance boost, raising the average IoU and F1-score by 4.1 and 3.9 points, respectively. To determine the specific factors responsible for this gain, we design comparative experiments with DFDE and its variants. Table 4 presents four variants, specifically: (a) TAFE-Net without DFDE. (b) Replacing DFDE with the Wavelet-like Frequency Enhancement (WFE) module of FFDN (Chen et al. 2024). (c) DFDE without vertical (LH) and horizontal (HL) frequency subbands. (d) Changing asymmetric convolutions (AsymConv) with standard square convolutions (StdConv). Relative to variant (b), DFDE improves the average IoU on the three DocTamper subsets by 2.3 points. Furthermore, the performance degradation observed in variant (c) confirms that vertical and horizontal frequency subbands contain significant forensic information, encompassing critical tampering indicators such as text patterns and edge-related artifacts. Finally, the drop in variant (d) demonstrates that asymmetric convolutions are essential for aggregating directional features, such as long-text features and fractured boundary artifacts. To further verify the effectiveness of DFDE, we also visualise the prediction maps of several TAFE-Net variants in Fig. 6. As can be seen, our model yields more complete and precise tampering masks, especially on long-text regions, whereas other variants leave noticeable holes or false positives.

Variants	Test	FCD	SCD	AVG
(a) TAFE-Net w/o DFDE	85.1	85.9	69.3	80.1
(b) DFDE → WFE	87.2	86.5	71.9	81.9
(c) DFDE w/o LH and HL	87.4	87.3	71.9	82.2
(d) AsymConv → StdConv	88.3	87.6	73.8	83.2
TAFE-Net w/ DFDE (Ours)	**89.4**	**89.3**	**74.0**	**84.2**

Table 4: Ablation study of the proposed DFDE module and its variants on the DocTamper dataset.

Model	Gaussian Noise	Gaussian Blur	Resize 1.5X	Resize 0.75X	JPEG 75	JPEG 50
FFDN	72.7	60.6	46.4	30.0	45.0	39.8
Ours	**76.0**	**67.3**	**52.3**	**41.6**	**49.6**	**43.3**

Table 5: Robustness evaluation on DocTamper-FCD dataset. Each image is first subjected to the designated degradation and is then JPEG-compressed once more using the quality factor provided by the public DTD repository.

Robustness Analysis

Real-world applications often involve degraded images, making robustness evaluation essential for practical deployment. For example, social media images routinely undergo heavy degradation, which weakens tampering traces and complicates detection. Therefore, we evaluate our model's robustness against various degradations (Gaussian noise, Gaussian blur, image scaling, and JPEG compression) and compare it with FFDN, the second-best performing model on the DocTamper-FCD dataset. The results illustrated in Table 5 indicate that our model consistently outperforms the previous state-of-the-art across different degradation types.

Conclusion

Forgery traces in document images are often not spread throughout the entire tampered text, but discretely hidden within a few characters or local regions. Motivated by this critical observation, we introduce the Text Aggregation and multi-Frequency Enhancement Network, termed as TAFE-Net. Our method is grounded in two core innovations: (1) Comprehensive multi-frequency feature integration that exploits the full frequency spectrum information to reveal subtle tampering artifacts. (2) Visual-frequency integration and direction-aware frequency decoupling enhancement mechanisms to aggregate text features and enhance details in the frequency domain to amplify local forgery traces, thereby improving the completeness of tampered text detection. Extensive experiments demonstrate that TAFE-Net achieves SOTA performance and exhibits strong robustness.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (Grant No.:62476093).

References

Chen, Z.; Chen, S.; Yao, T.; Sun, K.; Ding, S.; Lin, X.; Cao, L.; and Ji, R. 2024. Enhancing Tampered Text Detection Through Frequency Feature Fusion and Decomposition. In *Proceedings of the European Conference on Computer Vision*, 200–217.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-Attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Gonzalez, R. C. 2009. *Digital Image Processing*. Pearson education india.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and Localization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 20606–20615.

He, C.; Li, K.; Zhang, Y.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2023. Camouflaged Object Detection With Feature Decomposition and Edge Reconstruction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 22046–22055.

Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Cao, L.; and Chen, S. 2025. Diffusion Model-Based Image Editing: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kwon, M.-J.; Nam, S.-H.; Yu, I.-J.; Lee, H.-K.; and Kim, C. 2022. Learning JPEG Compression Artifacts for Image Manipulation Detection and Localization. *International Journal of Computer Vision*, 130(8): 1875–1895.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022a. 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.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A ConvNet for the 2020s. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Luo, D.; Liu, Y.; Yang, R.; Liu, X.; Zeng, J.; Zhou, Y.; and Bai, X. 2025. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157: 110828.

Ma, X.; Du, B.; Jiang, Z.; Hammadi, A. Y. A.; and Zhou, J. 2024. IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Nandanwar, L.; Shivakumara, P.; Mondal, P.; Raghunandan, K. S.; Pal, U.; Lu, T.; and Lopresti, D. 2020. Forged Text Detection in Video, Scene, and Document Images. *IET Image Processing*, 14(17): 4744–4755.

Qu, C.; Liu, C.; Liu, Y.; Chen, X.; Peng, D.; Guo, F.; and Jin, L. 2023. Towards Robust Tampered Text Detection in Document Image: New Dataset and New Solution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 5937–5946.

Qu, C.; Zhong, Y.; Liu, C.; Xu, G.; Peng, D.; Guo, F.; and Jin, L. 2024. Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Roy, A. G.; Navab, N.; and Wachinger, C. 2018. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*, 421–429.

Shao, H.; Qian, Z.; Huang, K.; Wang, W.; Huang, X.; and Wang, Q. 2024. Delving into Adversarial Robustness on Document Tampering Localization. In *Proceedings of the European Conference on Computer Vision*, 290–306.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations*.

Su, L.; Ma, X.; Zhu, X.; Niu, C.; Lei, Z.; and Zhou, J.-Z. 2025. Can We Get Rid of Handcrafted Feature Extractors? SparseViT: Nonsemantics-Centered, Parameter-Efficient Image Manipulation Localization Through Spare-Coding Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7024–7032.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022a. ObjectFormer for Image Manipulation Detection and Localization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2364–2373.

Wang, Y.; Xie, H.; Xing, M.; Wang, J.; Zhu, S.; and Zhang, Y. 2022b. Detecting Tampered Scene Text in the Wild. In *Proceedings of the European Conference on Computer Vision*, 215–232.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 16133–16142.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified Perceptual Parsing for Scene Understanding. In *Proceedings of the European Conference on Computer Vision*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*, 12077–12090.

Xu, G.; Yi, Z.; Zhang, P.; Cao, J.; Wu, S.; and Jin, L. 2025. From Pixels to Semantics: A Novel MLLM-Driven Approach for Explainable Tampered Text Detection. In *Proceedings of the ACM International Conference on Multimedia*, 757–766.

Zhang, P.; Ding, K.; and Jin, L. 2025. Capturing More: Learning Multi-Domain Representations for Robust Online Handwriting Verification. In *Proceedings of the ACM International Conference on Multimedia*, 1471–1479.

Zhang, P.; and Jin, L. 2024. Online Writer Retrieval With Chinese Handwritten Phrases: A Synergistic Temporal-Frequency Representation Learning Approach. *IEEE Transactions on Information Forensics and Security*, 19: 10387–10399.

Zhang, P.; Xu, H.; Zhang, J.; Xu, G.; Zheng, X.; Yang, Z.; Liu, J.; Zhang, Y.; and Jin, L. 2025. Aesthetics is Cheap, Show me the Text: An Empirical Evaluation of State-of-the-Art Generative Models for OCR. arXiv:2507.15085.

Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; and Zhou, J.-Z. 2025. Mesoscopic Insights: Orchestrating Multi-Scale & Hybrid Architecture for Image Manipulation Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11022–11030.