DocForgeNet: Dual Cross-Stream Fusion Network for Robust Forgery Detection in Scanned Documents

Nauman Riaz[1,2(✉)], Stefan Agne[1,3], Andreas Dengel[1,2], and Sheraz Ahmed[1,3]

[1] Smart Data and Knowledge Services (SDS), German Research Center for Artificial Intelligence GmbH (DFKI), Trippstadter Straße 122, 67663 Kaiserslautern, Germany
{nauman.riaz,stefan.agne,andreas.dengel,sheraz.ahmed}@dfki.de
[2] Department of Computer Science, RPTU Kaiserslautern-Landau, Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany
[3] DeepReader GmbH, 67663 Kaiserslautern, Germany

Abstract. Document image forgeries, especially text manipulations in scanned documents, pose severe risks to information integrity, impacting domains such as legal, financial, and personal records. These alterations are often subtle, localized, and visually indistinguishable, especially under common compression standards like JPEG. Existing detection methods typically rely on single-model architectures, either convolutional neural networks (CNNs) or transformers, which independently struggle with effectively capturing both local artifacts and global structural inconsistencies. To overcome these limitations, we propose DocForgeNet, a novel dual cross-stream fusion network explicitly designed for robust detection and localization of forged text regions in document images. DocForgeNet simultaneously processes RGB and discrete cosine transform (DCT) features through parallel CNN and transformer streams. The CNN stream excels at identifying local inconsistencies, such as compression artifacts and font irregularities, while the transformer stream leverages self-attention mechanisms to model broader, contextual discrepancies indicative of large-scale manipulations. Cross-linear attention modules facilitate effective feature fusion between the streams without information loss. Extensive evaluations on the DocTamper dataset, which contains forgeries on 170,000 document images of various types, demonstrate that DocForgeNet significantly outperforms state-of-the-art methods, achieving superior precision, recall, and F1-scores across various testing scenarios, including severe JPEG compression. Our approach not only establishes a new benchmark in tampering detection performance but also highlights the effectiveness of integrating complementary local and global representations for enhanced document integrity verification.

Keywords: Document Forgery Detection · Dual Cross-Stream Network · Frequency Perception Head (FPH) · DCT Features · Linear

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-04627-7_19.

X.-C. Yin et al. (Eds.): ICDAR 2025, LNCS 16026, pp. 329–346, 2026.
https://doi.org/10.1007/978-3-032-04627-7_19

Attention · Cross-Linear Attention · Multi-View Iterative Decoder (MID)

1 Introduction

Document images serve as ubiquitous carriers of sensitive information, including personal, legal, and financial data, where the integrity of textual content is paramount. However, advancements in image editing tools have enabled malicious actors to imperceptibly alter or replace text in scanned documents, posing significant threats to security and trust [2,6,30,32,41]. Such forgeries, whether aimed at manipulating contracts, receipts, or identification records, undermine the reliability of digital documentation, necessitating robust methods to detect and localize tampered text [30,33]. While recent research has focused on identifying tampering operations such as splicing, copy-move forgeries, and synthetic text generation, detecting these manipulations remains a formidable challenge due to their subtlety and the complex nature of document layouts.

(a) Forged Images

(b) Ground Truth

Fig. 1. Tampered text in document images typically occupies small areas and provides minimal visual evidence of tampering. This figure presents forged document images alongside their corresponding ground truth forged regions.

Unlike natural image forgeries [22,43], document text manipulations often lack obvious visual artifacts. Tampered regions in documents are often localized and minimal, with forged text carefully crafted to align with the original

in terms of font, size, and background characteristics. This meticulous matching makes the alterations nearly imperceptible to the human eye, as illustrated in Fig. 1. Compounding this challenge, scanned documents are frequently stored in lossy formats like JPEG, where compression artifacts obscure subtle tampering cues [24]. Block boundary noise, quantization effects, and other compression-induced distortions degrade image quality while masking critical forensic evidence. Prior efforts to exploit JPEG-specific patterns for forgery detection have shown limited success, particularly under varying compression levels, leaving room for improvement in robustness and generalization.

Existing detection frameworks often rely on single-model architectures, such as convolutional neural networks (CNNs) or transformers, or their sequential combination, each with inherent limitations [31]. While CNNs excel at capturing local spatial features, such as font irregularities or DCT coefficient discontinuities [6] they struggle to model long-range dependencies critical for detecting forgeries like copy-move spanning document regions. Transformers, conversely, adeptly identify global inconsistencies through self-attention mechanisms but often overlook fine-grained local details essential for localized tampering detection [16,22]. Sequential combinations of these architectures [33], where one model processes features before passing them to the other, risk diluting complementary strengths: early CNN layers may discard global context, while initial transformer stages might aggregate features at the expense of local precision.

To address these limitations, we propose a novel DocForgeNet dual-stream framework that leverages parallel CNN and transformer architectures to simultaneously capture both local and global tampering cues. The main contributions of our paper are as follows:

- **Dual Cross-Stream Processing:** DocForgeNet processes RGB and frequency (DCT) modalities through independent streams, allowing each to retain its unique advantages. The CNN stream is designed to capture localized artifacts, such as block boundary inconsistencies and compression-induced discontinuities, while the transformer stream models long-range dependencies to identify large-scale manipulations like copy-move forgeries. This parallel architecture prevents the information loss common in sequential processing, ensuring both fine-grained details and global structural coherence are preserved.
- **Enhanced Detection and Robustness:** The proposed method not only enhances localization on forged documents but also improves robustness to JPEG compression artifacts. outperforming the state-of-the-art on DocTamper [33] dataset by a significant margin. By leveraging the complementary strengths of CNNs and Transformers, our approach offers a more comprehensive solution for safeguarding document integrity.

The remainder of this paper is organized as follows: Sect. 2 reviews related work in document forgery detection and hybrid deep learning architectures. Section 3 details our methodology, including the dual cross-stream framework and fusion strategies. Section 4 presents experimental results, comparisons with state-of-the-art methods and ablation studies followed by a conclusion in Sect. 5.

2 Related Work

2.1 Natural Image Forgery

Early approaches to natural image forgery detection focused on handcrafted features and artifact analysis. Cozzolino *et al.* [11,12] proposed dense-field copy-move detection and blind splicing detection using sensor pattern noise. Fridrich *et al.* [15] introduced rich models for steganalysis through handcrafted noise residuals. While effective for specific artifacts, these methods lacked generalizability to unseen manipulations.

With the advent of deep learning, Zhou *et al.* [44] pioneered CNN-based approaches using RGB and SRM filtered inputs. Bappy *et al.* [4] combined LSTM with encoder-decoder networks for improved localization. Bayar *et al.* [6] proposed constrained CNNs to suppress image content while learning manipulation traces. However, these early deep methods struggled with subtle forgeries and precise localization.

While current state-of-the-art methods have demonstrated improved performance, many rely on single stream processing of RGB and frequency domains [4,6] or late fusion strategies [13], which fail to fully exploit the complementary nature of local and global features across modalities. Single-stream architectures [39] often lack the ability to simultaneously capture fine-grained local details and long-range global context. Additionally, transformer-based methods [25], while effective at modeling global relationships, frequently overlook critical local texture details that are essential for precise localization of forensic artifacts. These limitations highlight the need for a more integrated approach that can effectively leverage both local and global features without sacrificing the strengths of either.

2.2 Document Image Forgery

Early document forgery detection approaches focused on printer classification using texture features [23,29] and DCT analysis [37]. Font-based authentication emerged through statistical analysis [5], typographical features [45], and conditional random fields [8]. Template matching via intrinsic document contents [2] and text-line alignment verification [9] showed promise but required clean scanned documents.

Modern approaches employ deep learning: James *et al.*[19] used graph attention networks to detect manipulated regions, while Abramova *et al.*[1] detected copy-move forgeries through quantization artifacts, though failing under multiple compressions. Wang *et al.*[40] introduced a two-stream Faster R-CNN [34] combining RGB and frequency features, but primarily targets SRNet-generated forgeries [41] rather than careful copy-paste tampering.

The introduction of the recent DocTamper [33] dataset has significantly advanced the field, offering a large-scale, diverse collection of 170,000 document images with various tampering types, including copy-move, splicing, and generation. This dataset, along with its cross-domain testing subsets, provides a robust

benchmark for evaluating the generalization capabilities of tampering detection models. The DocTamper dataset's comprehensive nature and realistic tampering synthesis method make it a valuable resource for developing and testing state-of-the-art document forgery detection systems. Alongside the dataset, the authors proposed the Document Tampering Detector (DTD), a multi-modality Transformer-based model that integrates visual and frequency domain features. The DTD employs a Frequency Perception Head (FPH) to capture tampering clues from DCT coefficients and a Multi-view Iterative Decoder (MID) to leverage multi-scale feature information. While the DTD demonstrates strong performance, particularly in handling inconspicuous tampering traces, it primarily relies on a Swin-Transformer-based encoder for feature fusion. However, the patching mechanism in Swin Transformers, which divides the input image into non-overlapping patches, may lead to the loss of fine-grained local details. This limitation could potentially hinder the model's ability to detect subtle tampering artifacts in document images. For large images, the Swin Transformer requires larger initial patches to mitigate the $\mathcal{O}(n^2)$ computational complexity [21,28,35,36,38] of the attention mechanism, compromising the preservation of local feature details. In contrast, our DocForgeNet, a dual cross-stream CNN and transformer architecture, is specifically designed to address these limitations by simultaneously capturing fine-grained local details and long-range global dependencies.

3 DocForgeNet: The Proposed Approach

We propose DocForgeNet, specifically designed to address the unique challenges in document forgery detection, combining CNNs for local detail extraction with Transformers for global context modeling, as illustrated in Fig. 2. By integrating the complementary strengths of CNNs in capturing fine-grained texture patterns and Transformers in modeling long-range dependencies, this framework robustly detects both subtle pixel-level artifacts and broader layout inconsistencies. The proposed architecture comprises several key modules: a Frequency Perception Head (FPH), a fusion strategy integrated with a projection layer, dual cross-stream blocks, and a Multi-view Iterative Decoder (MID) as a segmentation head.

The motivation for our proposed dual cross-stream approach arises from observed limitations in using a Swin Transformer [26] backbone, as previously employed in [33]. While Swin Transformer is efficient for general image tasks due to its local window focus and periodic shifting for global context aggregation, it potentially weakens local detail capture if window partitions are too coarse or if subtle, fine-grained modifications, such as small text alterations, fall within inadequately represented internal window patterns. Moreover, exclusive reliance on a Transformer pathway might inadequately address local texture variations like pixel-level artifacts, where CNNs traditionally excel. Additionally, homogenously treating RGB and DCT inputs within a single-path architecture risks diluting

the unique frequency artifacts captured by DCT and the visual pattern cues inherent in RGB data, thus failing to fully exploit their distinct characteristics.

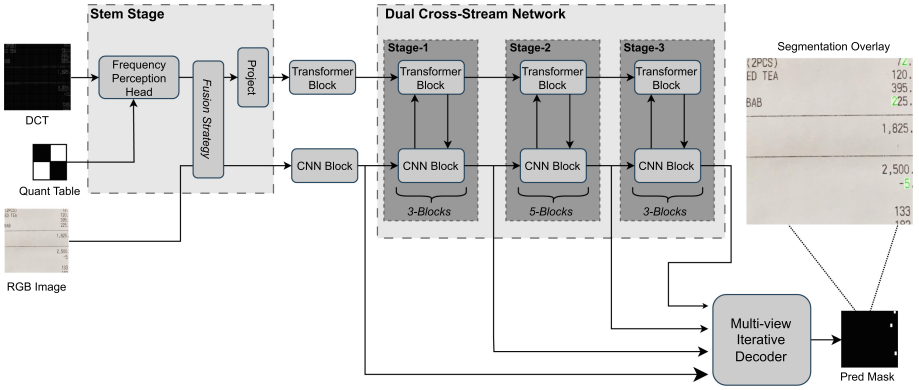

Fig. 2. Overview of DocForgeNet. We employ a dual cross-stream framework that processes both DCT and RGB images. Initially, the two modalities pass through the stem stage, where the assignment of each modality to a stream is configurable. This assignment is determined based on a fusion strategy, which serves as a hyperparameter for ablation studies. Subsequently, the features are processed through the Dual Cross-Stream Network, and multi-scale spatial features are fed into a Multi-View Iterative Decoder for final prediction.

3.1 Stem Stage

In the stem stage, we employ an FPH module similar to [33], which follows a dual-head design. It first extracts the DCT coefficient map from the Y channel and embeds it using convolution layers. Simultaneously, it processes the quantization table, expands it, and embeds it with learnable parameters. The embeddings are fused through element-wise multiplication and concatenated before being down-sampled to align with the input's BAG [24] structure. Finally, position embeddings and MobileConv [18] layers are applied to refine the frequency feature representation. Given a DCT coefficient map of size $H \times W$, we obtain the feature representation $F \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$, ensuring that each spatial unit in F corresponds to an 8×8 block in the original DCT domain. Next, the Fusion Strategy Module is primarily designed for ablation studies, where we tested three different combinations: (1) DCT features passed through the Transformer stream and RGB image passed through the Conv stream, (2) the reverse setup, and (3) both features fused via concatenation and passed through both streams. The best results were obtained when DCT features were processed through the Transformer stream, while the RGB image was passed through the Conv stream, as discussed in Sect. 4.4. The Projection Layer is necessary to divide the features

into fixed window-sized patches and flatten them. Given the DCT feature map $F \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$, it is transformed into $F' \in \mathbb{R}^{S \times C}$, where $S = \frac{H}{8} \times \frac{W}{8}$ represents the total number of patches. Since the FPH module already divides the features into 8×8 windows, we directly use each feature pixel as an individual patch. This stem stage is illustrated in Fig. 2.

3.2 Dual Cross-Stream Network

Before employing the dual cross-stream module, we pass the DCT features through a Linear Transformer [20] block and the RGB input image via a Convolution block without the cross-stream fusion (Linear Cross-Attention). The linear transformer block is similar to a naive transformer [38] encoder block with the difference of Multi-Head Linear Self Attention(MHLSA). This can be seen in Fig. 3.

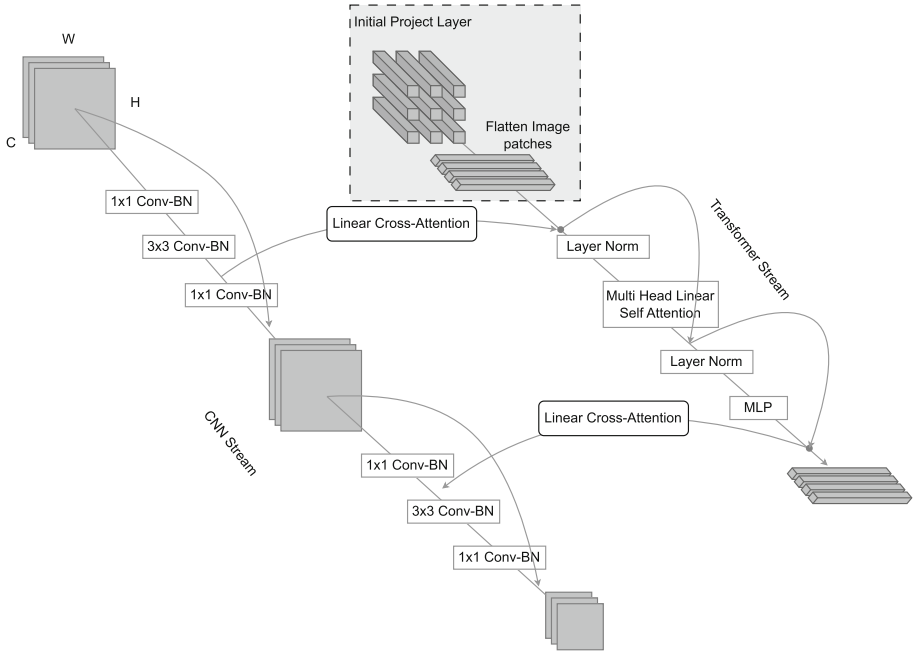

Fig. 3. Network architecture of the proposed DocForgeNet's dual cross-stream module, highlighting the implementation details of the CNN block, Transformer block, and Cross-Linear Attention mechanism used for cross-stream fusion.

MHLSA. The Multi-Head Linear Self-Attention (MHLSA) block employs a computationally efficient attention mechanism, reducing the quadratic complexity [21,36,38] of standard self-attention to a linear form. Given an input feature

map $X \in \mathbb{R}^{S \times C}$, where S is the sequence length and C is the embedding dimension, the standard self-attention is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are the query, key, and value projections with learnable weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{C \times d_k}$. To achieve linear complexity, Linear Attention replaces the softmax operation with a kernel function $\phi(\cdot)$:

$$A = \phi(Q)(\phi(K)^T V)$$

where $\phi(\cdot)$ is a feature transformation such as ReLU or exponential mapping. The computation is performed as follows:

1. Projection of Inputs:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

2. Feature Transformation:

$$\tilde{Q} = \phi(Q), \quad \tilde{K} = \phi(K)$$

3. Efficient Attention Calculation:

$$Z = \tilde{K}^T V \in \mathbb{R}^{d_k \times d_k}$$

$$A = \tilde{Q} Z$$

This formulation reduces complexity from $\mathcal{O}(S^2)$ to $\mathcal{O}(S)$, making it memory-efficient and scalable for high-resolution vision tasks.

Conv Block. The Convolutional Block follows a ResNet-like design, focusing on extracting local details from images. According to the definition in ResNet [17], a bottleneck consists of a 1×1 down-projection convolution, a 3×3 spatial convolution, and a 1×1 up-projection convolution. Additionally, a residual connection is introduced between the input and output of the bottleneck to enhance gradient flow and improve feature propagation as shown in Fig. 3. Visual Transformers [14,26] map image patches into feature vectors in a single step, which can lead to a loss of local details. In contrast, CNNs utilize sliding convolutional kernels that process feature maps with overlapping regions, allowing them to retain fine-grained local features. As a result, the CNN branch effectively preserves and continuously supplies local feature information to the transformer branch, enhancing its ability to capture spatial details.

Cross-Stream Fusion via Cross Linear Attention. We employ the Transformer and Convolution blocks in a dual-stream manner with Cross Linear Attention for cross-stream fusion, facilitating both global and local feature understanding. The dual cross-stream architecture is divided into three stages (refer to Fig. 2). After each stage, the spatial resolution of the features from the Convolution stream is reduced by half. Within each stage, Cross Linear Attention is applied between the Transformer and Convolution features. Specifically, when cross-attending from the Convolution stream to the Transformer stream, the **keys** and **values** originate from the Convolution stream, while the **queries** come from the Transformer stream. The reverse is applied when cross-attending in the opposite direction, as illustrated in Fig. 3.

When cross-fusion between the Convolution stream and Transformer stream occurs, the Convolution stream features of shape $F_{\text{conv}} \in \mathbb{R}^{H \times W \times C}$ are first reshaped into a sequence form:

$$F'_{\text{conv}} \in \mathbb{R}^{S \times C}, \quad S = H \times W$$

where S represents the total number of spatial locations after flattening.

Since we employ Linear Attention, the large spatial resolution $H \times W$ does not significantly impact computational complexity. The linear attention mechanism is formulated as:

$$A = \phi(Q)(\phi(K)^T V)$$

1. Cross-Attention from Convolution Stream to Transformer Stream:

$$Q = F_{\text{trans}} W_Q, \quad K = F'_{\text{conv}} W_K, \quad V = F'_{\text{conv}} W_V$$

2. Cross-Attention from Transformer Stream to Convolution Stream:

$$Q = F'_{\text{conv}} W_Q, \quad K = F_{\text{trans}} W_K, \quad V = F_{\text{trans}} W_V$$

Here, $W_Q, W_K, W_V \in \mathbb{R}^{C \times d_k}$ are learnable projection matrices, and $\phi(\cdot)$ is a kernel transformation ensuring computational efficiency.

Due to the linear complexity $O(S)$ of Linear Attention, the large spatial resolution of the Convolution stream does not introduce significant overhead. This enables efficient cross-stream information exchange, enhancing both global and local feature representations.

3.3 Multi-view Iterative Decoder

We employ a Multi-View Iterative Decoder (MID) similar to [33] to refine multi-scale features for more accurate predictions. Given the encoder's output features, MID iteratively processes them through cascaded operations, generating hierarchical decoder features. These features are then concatenated and projected via a convolution layer to obtain the final prediction. This iterative multi-view approach mimics human perception by integrating information across different

scales. We train our model combining Cross-Entropy Loss and Lovász Loss [7] to optimize segmentation accuracy. Given a prediction mask $\hat{y}$ for an input image x with ground-truth mask y, the total loss is formulated as:

$$L = L_{\text{ce}}(\hat{y}, y) + L_{\text{lov}}(\hat{y}, y)$$

where L_{ce} represents Cross-Entropy Loss, and L_{lov} is the Lovász Loss for better boundary optimization.

4 Experiments

We adopt an evaluation procedure similar to the one proposed in [33], ensuring consistency and comparability with prior work on the DocTamper dataset. Our evaluation follows the same metrics and benchmarking strategies, allowing for a direct assessment of our model's performance against existing approaches. The DocTamper dataset is evaluated using Precision, Recall, F1-score, and Intersection over Union (IoU), as the tampering detection task is formulated as a binary semantic segmentation problem. These metrics provide a comprehensive assessment of the model's performance in detecting manipulated regions.

4.1 DocTamper Dataset

The DocTamper [33] dataset is a comprehensive, large-scale collection specifically designed for tampered text detection in document images. It consists of 170,000 images, encompassing diverse document types such as contracts, invoices, and receipts, presented in both English and Chinese. The dataset addresses three prevalent tampering methods: copy-move, splicing, and generation, with each method evenly represented across the dataset to simulate real-world document manipulation scenarios accurately.

DocTamper includes a primary training set of 120,000 images, a general testing set of 30,000 images, and two specialized cross-domain testing subsets, DocTamper-FCD and DocTamper-SCD with 2,000 and 18,000 images, respectively. These cross-domain subsets facilitate robust evaluation of model generalization across varying image sources and document styles. Unlike previous datasets that typically contain fewer than 1,000 images and limited document types, DocTamper significantly expands scale and diversity, thus providing a more realistic and challenging benchmark for advancing research in document image tampering detection.

4.2 Implementation Details

We configure the RGB input image to have dimensions $512 \times 512 \times 3$ with its corresponding DCT representation set to 512×512. As the image progresses through the stem stage and early transformer block, the DCT features are downsampled to $128 \times 128 \times 64$, while the Conv block processes the RGB image into feature

maps of size $256 \times 256 \times 64$ (more details in ablation studies). For optimization, we employ AdamW [27] with an initial learning rate of 3×10^{-4}, which gradually decreases to 1×10^{-5} following a cosine decay schedule. Training is conducted on four A100-40GB GPUs, each handling a batch size of 8, resulting in a total batch size of 32. To align with the Doctamper testing setup, we introduce dynamic JPEG compression during training. The JPEG quality factor is randomly sampled between 75 and 100, while the number of compression iterations varies between 1 and 3. The final predictions are binarized using a threshold of 0.5.

4.3 Comparison with State-of-the-Art Methods

To the best of our knowledge, we compare our DocForgeNet against state-of-the-art image manipulation detection approaches [13,22,25,42], including the baseline model proposed in [33], as well as advanced semantic segmentation models [3,26]. Our experimental results demonstrate that our model achieves a significant performance improvement over existing approaches, as detailed in Table 1, highlighting the importance of integrating local and global representations for robust document image forgery detection. By employing a dual cross-stream architecture combining CNNs and transformers, our method effectively captures fine-grained structural inconsistencies and long-range contextual relationships, significantly enhancing tampering detection performance. Moreover, our approach establishes a new benchmark on the DocTamper dataset, improving upon the DTD model by a substantial margin. These results validate the effectiveness of our fusion strategy in handling document forgery detection under various compression and degradation scenarios, further reinforcing the role of multi-scale feature integration in this domain. We also conduct a visual comparison of our DocForgeNet dual cross-stream architecture with the state-of-the-art method DTD on the DocTamper dataset, as shown in Fig. 4, where the results indicate that DTD struggles with detecting forgeries in very small areas due to its Swin Transformer-based architecture, which depends on initial window patches that may lose fine-grained local details. Note that we did not include comparisons with [10] as their experimental setup deviates from the original DocTamper protocol, as clarified in their official GitHub repository, making a direct comparison under identical conditions infeasible. Following the DocTamper dataset settings, our model achieves an inference time of approximately 0.6678 s per image, outperforming the DTD model's 0.9863 s on CPU. Notably, our linear attention transformer operates with per-pixel embeddings and global attention, in contrast to DTD's Swin Transformer which relies on patch-based windowed attention—yet our approach remains more efficient.

4.4 Ablation Study

We conducted ablation study to explore three distinct fusion strategies for document forgery detection, each leveraging both Discrete Cosine Transform (DCT) and RGB features in different ways:

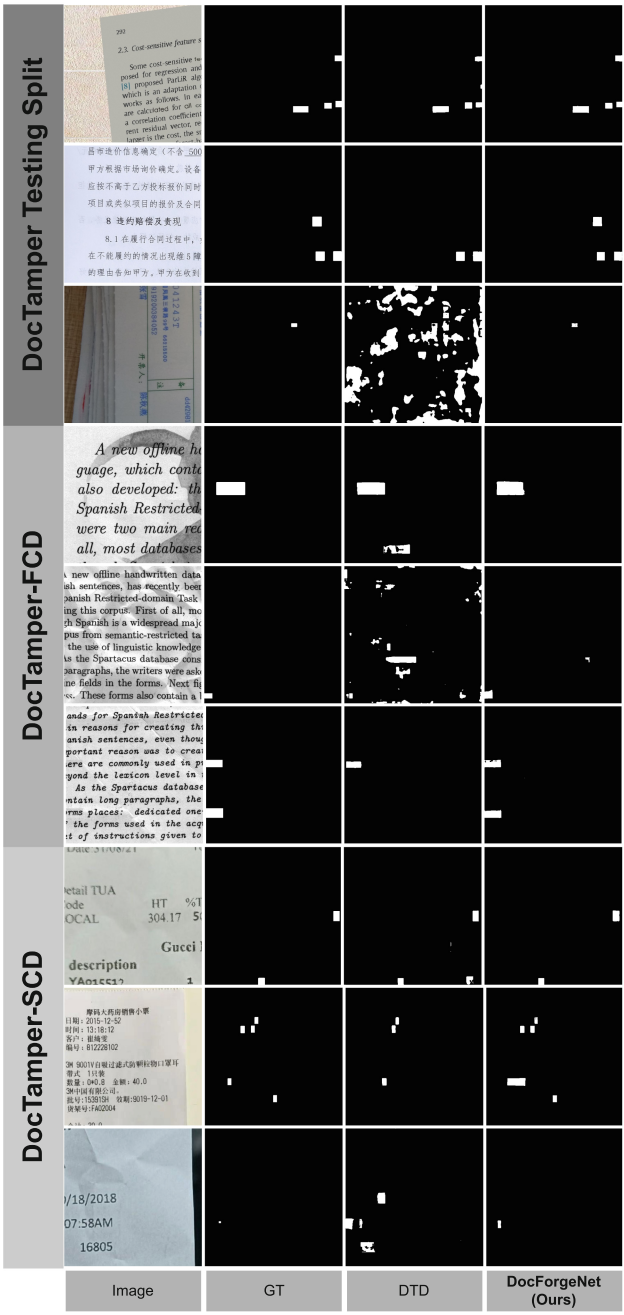

Fig. 4. Qualitative results on the DocTamper dataset [33], evaluated using the three testing splits proposed by the authors. Our proposed DocForgeNet architecture is compared against the state-of-the-art DTD method [33]. GT denotes the ground-truth segmentation mask.

Table 1. Comparison on the DocTamper dataset. All test images undergo random compression between one and three times with quality factors of at least 75, following the random seed specified by the authors in [33]. "D-TestingSet" refers to the primary DocTamper testing split, alongside the FCD and SCD splits proposed by the authors. "P" denotes precision, "R" recall, "F" the F-score, and "Params" indicates the number of parameters for each model.

Method	D-TestingSet			DocTamper-FCD			DocTamper-SCD			Params
	P	R	F	P	R	F	P	R	F	
Mantra-Net [42]	0.123	0.204	0.153	0.175	0.261	0.209	0.124	0.218	0.157	**4M**
MVSS-Net [13]	0.494	0.383	0.431	0.480	0.381	0.424	0.478	0.366	0.414	143M
PSCC-Net [25]	0.309	0.506	0.384	0.330	0.580	0.420	0.286	0.540	0.374	**4M**
BEiT-Uper [3]	0.564	0.451	0.501	0.550	0.436	0.487	0.408	0.395	0.402	120M
Swin-Uper [26]	0.671	0.608	0.638	0.642	0.475	0.546	0.541	0.612	0.574	121M
CAT-Net [22]	0.737	0.666	0.700	0.644	0.484	0.553	0.645	0.618	0.631	114M
DTD [33]	0.814	0.771	0.792	0.849	0.786	0.816	**0.745**	0.762	**0.754**	66M
DocForgeNet	**0.856**	**0.814**	**0.820**	**0.915**	**0.884**	**0.891**	0.711	**0.769**	0.712	32M

1. **DCT-TF**: DCT features were processed through a transformer-based stream, while RGB features were processed through a CNN-based stream.
2. **DCT-Conv**: The reverse of DCT-TF, where the DCT stream used a CNN and the RGB stream used a transformer.
3. **DCT+RGB Fusion**: The DCT and RGB features were concatenated, then passed jointly through both transformer and CNN streams.

We conducted our experiments on the DocTamper [33] dataset, following the same splits and protocols proposed by the original authors: the standard testing split, 'D-FCD' (DocTamper-FCD), and 'D-SCD' (DocTamper-SCD). To simulate different levels of compression-induced degradation, we subjected each image to one to three random JPEG compression cycles, with quality factors ranging from 75 to 100. We further tested lower-bound compression quality settings of 75100 and 90100, where 'Q' denotes the lowest applied compression factor. Evaluation was carried out using four key metrics—Intersection over Union (IoU), Precision (P), Recall (R), and F1-score (F)—across all splits. By adhering closely to the authors' original experimental setup and examining model performance under varied compression scenarios, we aimed to provide a direct and comprehensive comparison of these fusion strategies for detecting tampered text regions. Detailed results under these settings are presented in Tables 2 and 3.

Table 2. Ablation study results on the DocTamper dataset ("D-TestingSet", including FCD and SCD subsets). Each test image undergoes random JPEG compression (13 times, quality 75100) using the seed from [32]. "P", "R", and "F" denote precision, recall, and F-score, respectively. "Fusion" combines DCT and RGB features for both streams; "DCT-Conv" inputs DCT features into convolution and RGB into transformer; "DCT-TF" uses the opposite feature arrangement.

Method	D-TestingSet				DocTamper-FCD				DocTamper-SCD			
	IoU	P	R	F	IoU	P	R	F	IoU	P	R	F
Fusion	0.830	0.816	0.748	0.763	0.811	0.896	0.851	0.861	0.656	0.728	0.723	0.690
DCT-Conv	0.845	0.835	0.783	0.794	0.827	0.906	0.856	0.871	0.584	0.701	0.662	0.646
DTD [33]	0.828	0.814	0.771	0.792	0.749	0.849	0.786	0.816	**0.691**	**0.745**	0.762	**0.754**
DCT-TF (Ours)	**0.866**	**0.856**	**0.814**	**0.820**	**0.843**	**0.915**	**0.884**	**0.891**	0.657	0.711	**0.769**	0.712

Table 3. Ablation study on the DocTamper dataset under varying compression qualities, following the approach in [33]. All experiments use the IoU metric. 'Q' denotes the lowest compression quality factor. 'D-TestingSet' refers to the main DocTamper testing set, 'D-FCD' to DocTamper-FCD, and 'D-SCD' to DocTamper-SCD. The "Fusion" method refers to fusing DCT and RGB features before feeding them into both streams. "DCT-Conv" indicates DCT features passed through the convolution stream and RGB features through the transformer stream, while "DCT-TF" indicates the opposite arrangement.

Method	D-TestingSet		D-FCD		D-SCD	
	Q75	Q90	Q75	Q90	Q75	Q90
Fusion	0.830	0.901	0.811	0.850	0.656	0.784
DCT-Conv	0.845	0.901	0.827	0.859	0.584	0.699
DTD [33]	0.83	0.89	0.75	0.83	**0.69**	**0.78**
DCT-TF (Ours)	**0.866**	**0.919**	**0.843**	**0.869**	0.657	0.765

The DCT-TF configuration achieved the best results, demonstrating the advantage of combining global and local feature learning. DCT features, rich in global structure and frequency information, benefit from transformer-based long-range dependency modeling, aiding anomaly detection. RGB features, capturing fine-grained spatial details, are well-suited for CNNs, which excel at detecting tampering traces and texture inconsistencies. This complementary feature processing allows DCT-TF to outperform DCT-Conv (which lacks effective global modeling via CNN) and DCT+RGB Fusion (which introduces redundant information). The results highlight the importance of a well-structured multi-stream approach in forgery detection. Thus, the DCT-TF strategy provides the optimal fusion, leveraging both global and local feature representations for enhanced forgery detection in compressed document images.

5 Conclusion

The proposed DocForgeNet is tailored for document image forgery detection, addressing the unique challenges of identifying subtle manipulations in structured, text-rich content. By leveraging DCT features through a Linear Transformer with Multi-Head Linear Self-Attention (MHLSA), the model captures global spectral anomalies—such as inconsistencies in JPEG compression artifacts or tampering traces—while the Conv Block extracts localized spatial irregularities, including altered text regions, or inconsistent texture patterns. Cross Linear Attention facilitates bidirectional fusion, correlating global spectral discrepancies with localized spatial anomalies to pinpoint forgery regions with high precision.

The dual-stream design effectively bridges the complementary strengths of CNNs and Transformers: CNNs excel at capturing fine-grained local details, while Transformers model long-range dependencies and global context. This synergy overcomes the limitations of standalone CNNs, which may struggle with global patterns, and standalone Transformers, which can lose local detail due to patch-based tokenization. The linear complexity of MHLSA ensures scalability to high-resolution document scans, where preserving fine details (e.g., font edges, signature strokes) is critical. By retaining spatial resolution in early stages of the Conv stream, the framework maintains sensitivity to small-scale forgeries, while progressive downsampling balances computational efficiency.

The proposed architecture's efficiency and interpretability position it as a promising solution for real-world document forensics. Future work will focus on enhancing its capabilities by incorporating explainability mechanisms, such as cross-attention maps, to aid analysts in validating detected forgeries and understanding the model's decision-making process.

References

1. Abramova, S., Böhme, R.: Detecting copy-move forgeries in scanned text documents. In: Media Watermarking, Security, and Forensics (2016). https://api.semanticscholar.org/CorpusID:4468868
2. Ahmed, A.G.H., Shafait, F.: Forgery detection based on intrinsic document contents. In: Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS), pp. 252–256. IEEE (2014)
3. Bao, H., Dong, L., Wei, F.: Beit: BERT pre-training of image transformers. CoRR abs/2106.08254 (2021). https://arxiv.org/abs/2106.08254
4. Bappy, J.H., Simons, C., Nataraj, L., Manjunath, B.S., Roy-Chowdhury, A.K.: Hybrid LSTM and encoder-decoder architecture for detection of image forgeries. IEEE Trans. Image Process. **28**(7), 3286–3300 (2019)
5. Bataineh, B., Abdullah, S.N.H.S., Omar, K.: A statistical global feature extraction method for optical font recognition. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011. LNCS (LNAI), vol. 6591, pp. 257–267. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20039-7_26
6. Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection. IEEE Trans. Inf. Forensics Secur. **13**(11), 2691–2706 (2018)

7. Berman, M., Blaschko, M.B.: Optimization of the Jaccard index for image segmentation with the Lovász hinge. CoRR abs/1705.08790 (2017). http://arxiv.org/abs/1705.08790
8. Bertrand, R., Terrades, O.R., Gomez-Krämer, P., Franco, P., Ogier, J.M.: A conditional random field model for font forgery detection. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 576–580 (2015). https://api.semanticscholar.org/CorpusID:22313406
9. van Beusekom, J., Shafait, F., Breuel, T.M.: Text-line examination for document forgery detection. Int. J. Document Anal. Recogn. (IJDAR) **16**, 189–207 (2012). https://api.semanticscholar.org/CorpusID:254113860
10. Chen, Z., et al.: Enhancing tampered text detection through frequency feature fusion and decomposition. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024, pp. 200–217. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-73414-4_12
11. Cozzolino, D., Poggi, G., Verdoliva, L.: Efficient dense-field copy–move forgery detection. IEEE Trans. Inf. Forensics Secur. **10**(11), 2284–2297 (2015). https://doi.org/10.1109/TIFS.2015.2455334
12. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: a new blind image splicing detector. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2015). https://doi.org/10.1109/WIFS.2015.7368565
13. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: MVSS-net: multi-view multi-scale supervised networks for image manipulation detection. IEEE Trans. Pattern Anal. Mach. Intell. (2022). https://doi.org/10.1109/TPAMI.2022.3180556
14. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. CoRR abs/2010.11929 (2020). https://arxiv.org/abs/2010.11929
15. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **7**(3), 868–882 (2012). https://doi.org/10.1109/TIFS.2012.2190402
16. 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 (ICCV), pp. 15035–15044 (2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
18. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861 (2017). http://arxiv.org/abs/1704.04861
19. Joren, H., Gupta, O., Raviv, D.: Learning document graphs with attention for image manipulation detection. In: El Yacoubi, M., Granger, E., Yuen, P.C., Pal, U., Vincent, N. (eds.) Pattern Recognition and Artificial Intelligence, pp. 263–274. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-09037-0_22
20. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: fast autoregressive transformers with linear attention. CoRR abs/2006.16236 (2020). https://arxiv.org/abs/2006.16236
21. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: the efficient transformer. CoRR abs/2001.04451 (2020). https://arxiv.org/abs/2001.04451
22. 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)

23. Lampert, C.H., Mei, L., Breuel, T.M.: Printing technique classification for document counterfeit detection. In: 2006 International Conference on Computational Intelligence and Security, vol. 1, pp. 639–644 (2006). https://doi.org/10.1109/ICCIAS.2006.294214
24. Li, W., Yuan, Y., Yu, N.: Passive detection of doctored JPEG image via block artifact grid extraction. Signal Process. **89**(9), 1821–1829 (2009)
25. Liu, X., Liu, Y., Chen, J., Liu, X.: PSCC-net: progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Trans. Circuits Syst. Video Technol. **32**, 1–1 (2022). https://doi.org/10.1109/TCSVT.2022.3189545
26. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. CoRR abs/2103.14030 (2021). https://arxiv.org/abs/2103.14030
27. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR abs/1711.05101 (2017). http://arxiv.org/abs/1711.05101
28. Maqsood, A., Riaz, N., Ul-Hasan, A., Shafait, F.: A unified architecture for Urdu printed and handwritten text recognition. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) ICDAR 2023, pp. 116–130. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-41685-9_8
29. Mikkilineni, A.K., Chiang, P.J., Ali, G.N., Chiu, G.T.C., Allebach, J.P., Delp, E.J.: Printer identification based on graylevel co-occurrence features for security and forensic applications. In: IS&T/SPIE Electronic Imaging (2005). https://api.semanticscholar.org/CorpusID:13441570
30. Nandanwar, L., et al.: A new method for detecting altered text in document images. Int. J. Pattern Recogn. Artif. Intell. **35**, 2160010 (2021). https://doi.org/10.1142/S0218001421600107
31. Peng, Z., et al.: Conformer: local features coupling global representations for visual recognition (2021). https://arxiv.org/abs/2105.03889
32. Pun, A.K., Javed, M., Doermann, D.S.: A survey on change detection techniques in document images (2023). https://arxiv.org/abs/2307.07691
33. Qu, C., et al.: 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 (CVPR), pp. 5937–5946 (2023)
34. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR abs/1506.01497 (2015). http://arxiv.org/abs/1506.01497
35. Riaz, N., Arbab, H., Maqsood, A., Nasir, K., Ul-Hasan, A., Shafait, F.: Convtransformer architecture for unconstrained off-line Urdu handwriting recognition. Int. J. Doc. Anal. Recogn. **25**(4), 373–384 (2022). https://doi.org/10.1007/s10032-022-00416-5
36. Riaz, N., Latif, S., Latif, R.: From transformers to reformers. In: 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), pp. 1–6 (2021). https://doi.org/10.1109/ICoDT252288.2021.9441516
37. Schulze, C., Schreyer, M., Stahl, A., Breuel, T.: Using DCT features for printing technique and copy detection. In: Peterson, G., Shenoi, S. (eds.) DigitalForensics 2009. IAICT, vol. 306, pp. 95–106. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04155-6_7
38. Vaswani, A., et al.: Attention is all you need (2023). https://arxiv.org/abs/1706.03762
39. Wang, J., et al.: Objectformer for image manipulation detection and localization (2022). https://arxiv.org/abs/2203.14681

40. Wang, Y., Zhang, B., Xie, H., Zhang, Y.: Tampered text detection via RGB and frequency relationship modeling. Chin. J. Netw. Inf. Secur. **8**(3), 29 (2022). https://doi.org/10.11959/j.issn.2096-109x.2022035. https://www.infocomm-journal.com/cjnis/EN/abstract/article_172502.shtml
41. Wu, L., et al.: Editing text in the wild. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1500–1508. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3343031.3350929
42. Wu, Y., AbdAlmageed, W., Natarajan, P.: Mantra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
43. Yan, C., Li, S., Li, H.: TransU2-Net: a hybrid transformer architecture for image splicing forgery detection. IEEE Access **11**, 33313–33323 (2023)
44. 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 (CVPR) (2018)
45. Zramdini, A., Ingold, R.: Optical font recognition using typographical features. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 877–882 (1998). https://doi.org/10.1109/34.709616