Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061006

ARTICLE

DDT-Net: Deep Detail Tracking Network for Image Tampering Detection

Jim Wong[1,2] **and Zhaoxiang Zang**[3,*]

[1]Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, 443002, China

[2]College of Computer and Information Technology, China Three Gorges University, Yichang, 443002, China

[3]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650504, China

*Corresponding Author: Zhaoxiang Zang. Email: zhaoxiang.zang@ctgu.edu.cn

Received: 14 November 2024; Accepted: 03 March 2025; Published: 16 April 2025

ABSTRACT: In the field of image forensics, image tampering detection is a critical and challenging task. Traditional methods based on manually designed feature extraction typically focus on a specific type of tampering operation, which limits their effectiveness in complex scenarios involving multiple forms of tampering. Although deep learning-based methods offer the advantage of automatic feature learning, current approaches still require further improvements in terms of detection accuracy and computational efficiency. To address these challenges, this study applies the U-Net 3+ model to image tampering detection and proposes a hybrid framework, referred to as DDT-Net (Deep Detail Tracking Network), which integrates deep learning with traditional detection techniques. In contrast to traditional additive methods, this approach innovatively applies a multiplicative fusion technique during downsampling, effectively combining the deep learning feature maps at each layer with those generated by the Bayar noise stream. This design enables noise residual features to guide the learning of semantic features more precisely and efficiently, thus facilitating comprehensive feature-level interaction. Furthermore, by leveraging the complementary strengths of deep networks in capturing large-scale semantic manipulations and traditional algorithms' proficiency in detecting fine-grained local traces, the method significantly enhances the accuracy and robustness of tampered region detection. Compared with other approaches, the proposed method achieves an F1 score improvement exceeding 30% on the DEFACTO and DIS25k datasets. In addition, it has been extensively validated on other datasets, including CASIA and DIS25k. Experimental results demonstrate that this method achieves outstanding performance across various types of image tampering detection tasks.

KEYWORDS: Image forensics; image tampering detection; image manipulation detection; noise flow; Bayar

1 Introduction

In today's digital world, images are an integral part of everyday life. With the rise of image editing software, it has become easier for users to manipulate images, often leaving few visible traces. This poses a serious challenge to the authenticity of digital images.

Current methods for detecting image tampering can be generally divided into two major categories: traditional methods and deep learning methods. Each of these methods has its own advantages and disadvantages in feature extraction and detection capability. However, in actual detection applications, there is still a significant research gap: how to ensure the capability to detect large-scale tampering while also sensitively identifying subtle tampering traces, thereby enhancing model interpretability and generalization.

The core advantage of traditional methods lies in their precise capture of low-level physical features. These methods primarily focus on pixel-level and texture-based low-order features, such as image histogram analysis [1] and the Spatial Rich Models (SRM) [2,3]. These features often contain subtle traces left by tampering operations, providing reliable physical evidence for detection. However, traditional methods also suffer from notable limitations: when confronted with complex semantic scenarios, relying solely on low-order features is often insufficient for accurately identifying tampered regions. This issue is particularly pronounced in scenes involving multiple semantic objects, where the detection results are often unsatisfactory. In contrast, deep learning-based methods, owing to their powerful feature learning capabilities, can automatically extract high-level semantic features from images, demonstrating significant advantages in handling large-scale tampering [4]. By leveraging multi-layer neural networks for hierarchical abstraction, these methods can comprehend the semantic content of images, thereby improving the detection of tampering at the semantic level. Nevertheless, deep learning methods also have their inherent drawbacks: they are relatively weak in capturing subtle tampering traces. More importantly, the performance of deep learning models heavily depends on the quality and distribution of the training data. When there is a substantial discrepancy between the distributions of the training and testing datasets, the generalization ability of the model tends to drop significantly. Additionally, the "black-box" characteristics of deep learning models also bring interpretability issues, making it difficult to provide intuitive explanations and reliable evidence for detection results.

In response to the aforementioned research gap, we propose the following research questions: Can integrating traditional noise residual methods with deep learning significantly improve detection accuracy across tampering scenarios of varying scales and complexities? Can the proposed dual-stream design effectively enhance the model's generalization and interpretability, addressing critical limitations of existing methods? How does the dual-stream fusion method perform on datasets with different tampering characteristics? What factors influence its robustness and adaptability?

To address the research questions outlined above, we propose DDT-Net, an innovative dual-stream tampering detection network. This network is based on the U-Net 3+ [5] architecture and incorporates an innovative dual-stream design, enabling the deep integration of noise residual features and semantic features. In our network design, we introduce the Bayar noise stream as a powerful complement to the traditional RGB feature stream. The noise stream focuses on extracting subtle noise residual features in images, while the deep learning network captures high-level semantic information. By employing a multiplicative fusion of feature maps at each layer, the network accurately identifies tampering traces at varying scales, significantly improving overall detection accuracy, particularly in detecting fine details within complex backgrounds. The noise stream aids in capturing subtle local tampering traces and mitigates overfitting, while high-level semantic feature learning ensures effectiveness in large-scale tampering scenarios. The introduction of the noise stream enhances interpretability by making the model's decision-making process more transparent, thereby reducing the "black-box" nature of deep learning models. To validate the effectiveness of our approach, we conducted comprehensive experiments on several datasets with varying tampering characteristics. Experimental results demonstrate that DDT-Net exhibits exceptional robustness and adaptability. The noise stream not only excels in detecting minute tampering but also adapts to various tampering types. With the carefully designed dual-stream multiplication network structure, our method provides stable and accurate detection results across diverse tampering environments.

The goal of this study is to overcome the limitations of current image tampering detection methods by integrating traditional noise detection methods with deep learning frameworks. This integration enables the precise detection of large-scale tampering while maintaining the sensitivity to subtle tampering traces, while also enhancing model interpretability and generalization. In serious application scenarios such as judicial evidence collection and news authenticity verification, simply providing reliable tampering detection

results is insufficient. Professionals and end-users urgently require clear and reliable visual evidence chains to understand how the model reaches specific conclusions. Therefore, one of the key motivations behind this research is to ensure reliable tampering detection while accurately detecting minor local anomalies and providing intuitive and trustworthy explanations for the detection outcomes.

In summary, the research gap, research questions, and other key points of this study are all centered around the critical challenges in image tampering detection technology. The research gap lies in the difficulty of existing methods to simultaneously achieve large-scale tampering detection and identification of subtle tampering traces. To address this issue, we explore the possibility of integrating traditional methods with deep learning and propose the DDT-Net network based on a dual-stream architecture. DDT-Net is the first to closely integrate the Bayar noise stream with the U-Net 3+ model through multiplication, establishing an effective bridge between the dual streams and achieving a $1 + 1 > 2$ effect. In practical applications, DDT-Net's tampering detection technology can not only enhance the authenticity and credibility of digital images while curbing the spread of false information but also play a key role in various fields such as news media and social platforms. Moreover, it enhances image data security and reliability, facilitates transparent information dissemination, and provides strong technical support for public safety and legal affairs. We have shared the source code and related data files of DDT-Net at: https://github.com/Moriartest/DDT-Net (accessed on 2 March 2025).

2 Related Work

Image tampering detection has emerged as a major research focus, with numerous detection algorithms proposed by researchers worldwide. Among the commonly used detection algorithms, traditional methods and deep learning methods each have their own advantages and play an important role in different application scenarios [6].

2.1 Traditional Methods

Traditional image tampering detection methods typically rely on physical consistency analysis, which identifies tampered areas by detecting abnormal traces caused by the tampering process. These methods primarily include those based on imaging content consistency, imaging system fingerprint consistency, post-processing traces, and JPEG re-compression traces.

In the field of imaging content consistency detection, researchers focus on whether the image content violates the natural characteristics of a real-world scene. The core idea is based on a fundamental assumption: the physical properties (e.g., lighting, shadows, material) in real images should exhibit inherent consistency. For example, Gu et al. [7] in 2024 employed illumination inconsistency for detection, estimating the direction and properties of light sources by analyzing light and shadow contours cast on objects. On the other hand, imaging system fingerprint consistency detection examines the unique "fingerprint" characteristics left by digital imaging devices. These features include color distortions, color filter array patterns [8], and device noise. For instance, Xu et al. [9] in 2023 investigated the impact of sensor physical characteristics on imaging results, focusing on Photo Response Non-Uniformity, offering new research perspectives in device fingerprint analysis. In addition, the analysis of post-processing traces of tampering is also an important direction in traditional methods. These techniques typically rely on detecting texture and edge features to identify traces left by tampering actions, such as geometric transformations, blurring, or median filtering. For instance, Sujin et al. [10] in 2024 employed the Scale-invariant Feature Transform (SIFT) for keypoint matching. To detect traces of geometric transformations, Wang et al. [11] in 2023 applied both frequency-domain and spatial-domain detection methods.

JPEG recompression trace detection, as an important branch of traditional methods, has its theoretical basis in a deep understanding of the image compression principle. Studies show that when an image undergoes multiple JPEG compressions, its Discrete Cosine Transform (DCT) coefficient distribution exhibits specific statistical characteristics. Research in this area can be divided into aligned and non-aligned recompression, addressing different tampering scenarios. Kwon et al. [12] in 2022 made significant progress in aligned recompression detection by proposing a DCT coefficient analysis method that effectively identifies abnormal patterns in compression history. Wang et al. [13] in 2021 contributed to the non-aligned recompression detection field by proposing a feature extraction scheme based on optimized pixel differences, offering new insights into solving tampering detection in complex scenarios.

2.2 Deep Learning Methods

With the rapid progress of deep learning, the field of image tampering detection has experienced major methodological advancements [14]. Compared to traditional approaches, deep learning methods offer distinct advantages: they can automatically learn and extract deep features from images while demonstrating superior generalization ability in complex scenarios.

Convolutional Neural Networks (CNNs) play a crucial role in image tampering detection. By leveraging hierarchical feature extraction through convolutional layers, CNNs effectively capture the spatial structure of images. For instance, Bi et al. [15] in 2019 proposed a novel Ringed Residual U-Net (RRU-Net [15]), which enables precise localization of tampered regions without requiring additional pre-processing or post-processing steps, highlighting the advantages of deep learning in precise tampering region localization. Inspired by this success, subsequent research has explored more sophisticated network architectures. For example, Dong et al. [16] in 2023 introduced MVSS-Net, which innovatively incorporated a dual-path design consisting of a noise stream and an RGB stream, significantly improving the model's ability to identify different types of tampering. Similarly, Kwon et al. [12] in 2022 proposed the CAT-Net architecture using HR-Net, where both a DCT stream and an RGB stream were employed for detection.

The introduction of the Visual Transformer [17] (ViT) marked a pivotal advancement in image tampering detection. ViT is a deep learning model based on the Transformer [18] architecture, capable of capturing the global contextual information of images in visual tasks. By dividing input images into fixed-size patches and treating each patch as a sequence input to the Transformer model, ViT overcomes the limitations of traditional convolutional networks and is better suited for tasks with complex global dependencies. In the field of image tampering detection, ViT's advantage lies in its ability to simultaneously capture the global information of large-scale tampering and the local features of fine-grained tampering. For example, Atak et al. [19] in 2024 proposed a variational autoencoder based on ViT for image tampering localization tasks. This approach not only improved adaptability to complex tampering patterns but also enhanced sensitivity to subtle tampering traces. Similarly, Guillaro et al. [20] in 2023 introduced TruFor, a cross-modal fusion architecture leveraging ViT, which focuses on integrating multi-scale information to enhance robustness and adaptability in complex tampering scenarios. TruFor incorporates a noise-sensitive fingerprint, Noiseprint++, which captures camera characteristics and editing history, significantly improving the generalization capability of image tampering detection models.

As the understanding of tampering detection tasks deepens, researchers have gradually realized that simple classification models may overlook the relative characteristics between forged and original regions, leading to cross-image interference and unstable performance. To address these issues, Wu et al. [21] proposed the FOCAL framework in 2023, which detects tampered regions through contrastive learning and unsupervised clustering. FOCAL employs pixel-level contrastive learning to extract high-level features while dynamically classifying features into forged and original categories using an unsupervised clustering method,

significantly improving detection performance. This approach validates the potential of multimodal feature integration. This idea was further developed in the MMFusion framework proposed by Triaridis et al. [22] in 2024, integrating three filters: Bayar convolution, SRM filter, and Noiseprint++. MMFusion employs an early fusion strategy to combine features output by these filters and incorporates a re-weighted decoder strategy to effectively utilize their complementary advantages in different tampering scenarios.

In summary, current research needs to enhance model robustness and generalization to handle various tampering types and diverse data distributions. One potential improvement direction is to design multi-stream network architectures, where part of the model focuses specifically on noise features in detailed regions, effectively compensating for the limitations of a single feature stream. A dual-stream complementary approach can offer stronger robustness when handling diverse tampering scenarios. To further improve generalization, it may be beneficial to deeply integrate information from different feature layers, thus reducing the model's dependency on specific datasets. This strategy can mitigate overfitting and enhance adaptability and stability across different contexts.

3 Methods

3.1 DDT-Net Model

In this paper, we propose an enhanced dual-stream tampering detection network (as shown in Fig. 1) that integrates deep learning architectures with traditional algorithmic approaches to leverage their complementary strengths. The core methodological innovation of the DDT-Net model is: a deep learning framework based on U-Net 3+ architecture, incorporating the classical Bayar detection algorithm to enhance manipulation detail detection capabilities, implementing multiplicative feature fusion (as shown in Fig. 2) of dual streams prior to skip connections, and ultimately generating tampering masks through full-scale feature fusion.

3.2 Fusion of Dual-Stream Feature Maps

Fig. 2 illustrates the fusion process of the Bayar stream and the RGB stream, where each stream processes input features through independent channels. The Bayar stream primarily captures noise residual features of the image, while the RGB stream focuses on extracting deep semantic features. These two streams undergo a series of feature extraction steps sequentially: first, MaxPooling (2×2) is applied to reduce the spatial dimensions while retaining key information; next, 3×3 convolutional layers (Conv 3×3) are used to further extract and enhance spatial features; subsequently, Batch Normalization is applied to stabilize the training process and accelerate convergence; finally, ReLU nonlinear activation functions are employed to enhance the expressiveness of the features.

Upon completing their respective feature extraction processes, the features from the Bayar and RGB streams are fused through element-wise multiplication. This pixel-level interaction effectively combines noise features and deep semantic features, significantly improving the model's sensitivity to subtle tampered regions. The fused features are then processed again with Batch Normalization and ReLU activation to further enhance the stability and robustness of feature representation. The final fused feature map is output for subsequent image tampering detection tasks.

In the overall network structure, element-wise multiplication serves as a critical step, achieving an effective synergy between fine-grained features and global semantic features. This approach fully leverages the respective advantages of the Bayar and RGB streams.

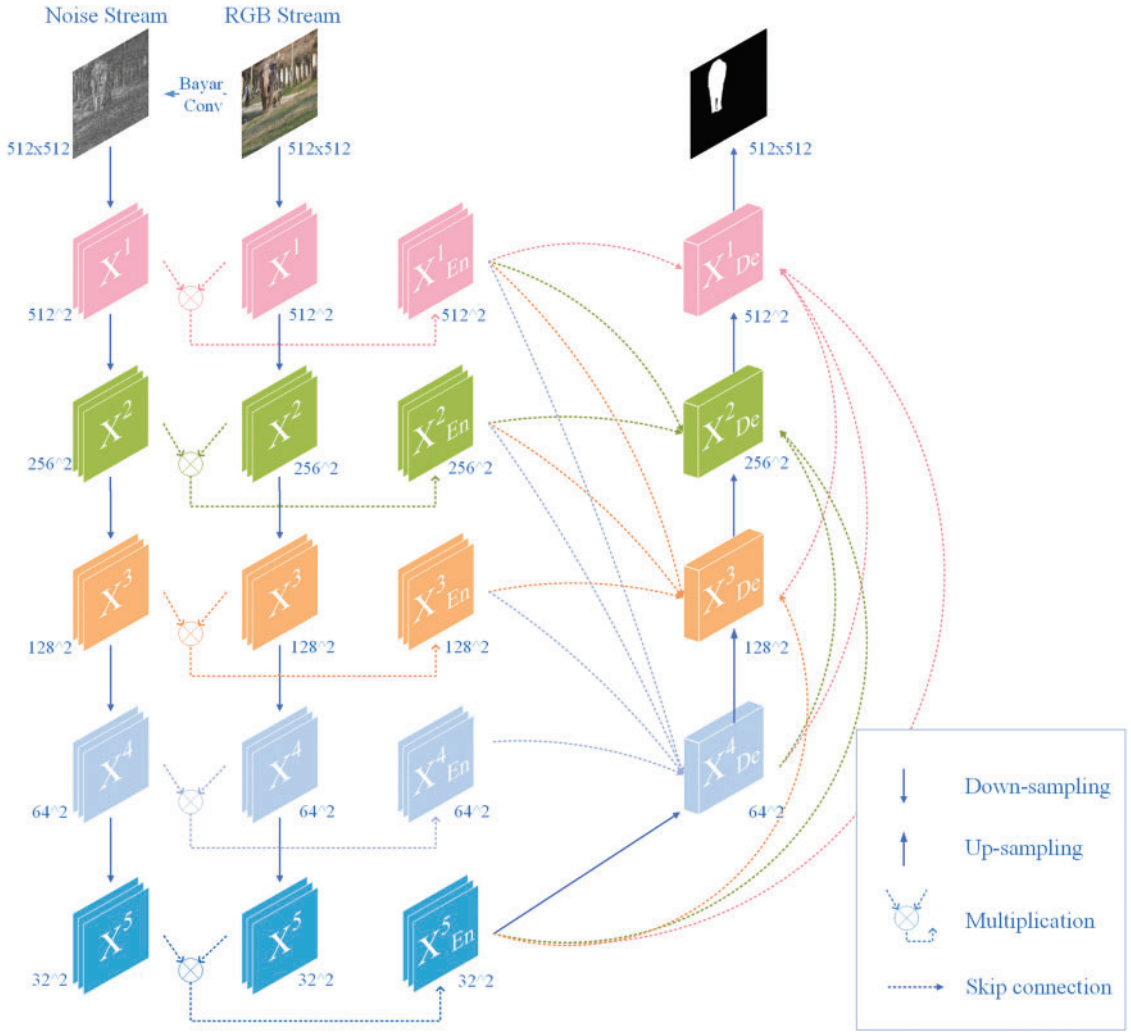

Figure 1: DDT-Net network architecture

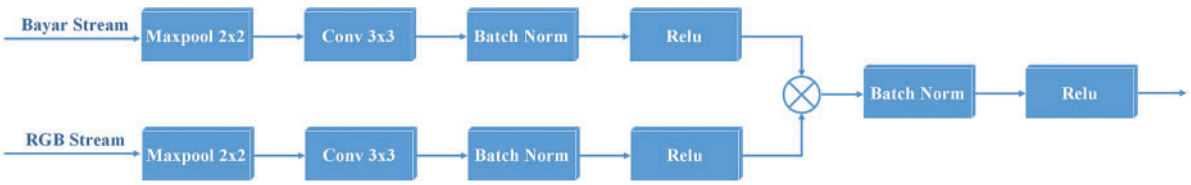

Figure 2: The process of fusing two feature maps of RGB stream and Bayar stream by multiplication

3.3 Full-Scale Skip Connection Module

The connections between encoder and decoder feature maps, implemented through dashed lines, are termed skip connections. These connections not only assist in restoring spatial resolution but also ensure that detailed information from the encoder can be utilized by the decoder, preventing excessive information loss. The utilization of full-scale features effectively combines feature information at different spatial resolutions, balancing global semantic information with local details, which is crucial in numerous computer vision

tasks [23]. The construction of feature maps of decoder X^2_{De} is shown in Fig. 3, to more clearly illustrate the operational mechanism of the full-scale skip connection module.

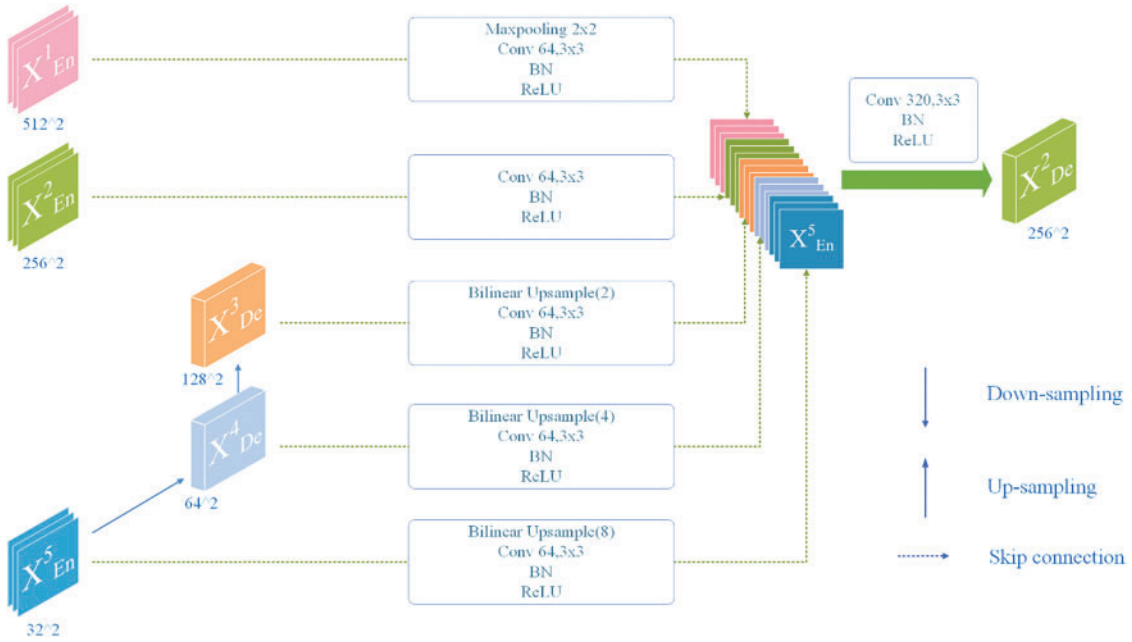

Figure 3: X^2_{De} feature map construction process in the decoder

First, the model employs 64 convolution kernels, along with regularization and activation functions, to capture local features and standardize the channel count of feature maps across different layers.

Second, the model needs to unify feature map resolutions:

When encountering resolutions higher than X^2_{De}, encoder X^1_{En} first undergoes max pooling operations to reduce spatial resolution (from 512 × 512 to 256 × 256) and extract significant features, followed by channel number unification;

When encountering resolutions identical to X^2_{De} (256 × 256), X^2_{En} only undergoes channel number unification without additional operations;

When encountering resolutions lower than X^2_{De}, decoders X^3_{De}, X^4_{De}, and X^5_{En} first employ bilinear interpolation for upsampling to restore feature maps to higher resolution (from 128 × 128, 64 × 64, and 32 × 32 to 256 × 256), followed by channel number unification.

At this point, the full-scale skip connection operation is complete, forming the initial decoder X^2_{De} feature map. However, to further capture correlations between local pixels and enhance feature representation, the model employs 320 convolution kernels and other sequential steps to construct the final decoder X^2_{De} feature map.

3.4 Bayar Algorithm

The convolutional kernel template weights in the Bayar algorithm are represented as shown in Eq. (1), where the superscript (l) denotes the l-th CNN layer, and subscript k indicates the k-th convolutional kernel within the layer. The kernel's central value is denoted by spatial index (0, 0), with the central weight fixed at −1 and the remaining weights at 1. The Bayar algorithm functions primarily as a noise extractor, significantly

reducing the impact of image content on subtle tampering traces through constrained convolutional layers, enabling the network to adaptively extract tampering trace features from images. The algorithm generates prediction error images capable of detecting suspicious manipulation locations within the image.

$$\begin{cases} \omega_k^{(\iota)}(0,0) = -1 \\ \sum_{m,n\neq 0} \omega_k^{(\iota)}(m,n) = 1 \end{cases} \tag{1}$$

In general, deep learning models excel at capturing large-scale semantic tampering features holistically, while the Bayar algorithm specializes in detecting localized subtle trace features, demonstrating superior detection capabilities for small-scale manipulations. The synergistic integration of these approaches enables comprehensive coverage of diverse types and scales of image tampering, thereby significantly enhancing detection accuracy and robustness.

4 Experiments

4.1 Experimental Setup

The experiments in this study were conducted using the PyTorch framework, with accelerated training on two NVIDIA GeForce RTX 3090 GPUs. The input size was set to 256 × 256. The model was trained from scratch, with all parameters learned directly on the target dataset without using any pre-trained weights.

For data processing, random cropping was employed as a data augmentation technique to enhance data diversity and improve the model's adaptability to complex scenarios. Training was conducted in mini-batches, with each GPU processing 12 images, achieving a balance between memory efficiency and training stability.

Regarding hyperparameter configuration, it was observed that training became unstable with oscillations when the learning rate exceeded 0.1. After extensive experimentation, the Adam optimizer was selected, with an initial learning rate set to 0.0003. The learning rate was reduced to 10% of its current value every 20 epochs during training to facilitate dynamic adjustment. Binary Cross-Entropy Loss was chosen as the loss function [24]. To enhance model robustness and mitigate overfitting, an early stopping strategy was employed, halting training if validation loss failed to improve for five consecutive iterations.

4.2 Datasets

In this study, several commonly used public datasets in the field of image tampering detection were selected to conduct comprehensive and in-depth experiments and evaluations (as shown in Table 1) [25]. We randomly selected over 100,000 images from the dataset for training and approximately 12,000 images for testing (as shown in Table 2).

Table 1: Information about the dataset

Dataset name	Splicing	Copy move	Removal	Tampered image	Real image
DEFACTO	√	√	√	149,000	0
DIS25k	√			24,965	0
IMD2020	√	√	√	2010	2010
MS-COCO				0	118,287
CASIAv1	√	√	√	800	921

(Continued)

Table 1 (continued)

Dataset name	Splicing	Copy move	Removal	Tampered image	Real image
Columbia	√			180	183
NIST16	√	√	√	564	560

Table 2: Composition of the dataset

Dataset	Training				Testing						
	DEFACTO	DIS25k	IMD2020	MS-COCO	DEFACTO	DIS25k	IMD2020	MS-COCO	CASIAv1	Columbia	NIST16
Positive	70,336	16,478	1186		2100	2500	267		918	180	144
Negative				20,000			267	4600	797	180	144

Training Datasets: The primary training datasets in this study comprise tampered datasets DEFACTO, IMD, and DIS25k, along with the untampered MS-COCO dataset. These four training sets, totaling 100,000 selected images, are intended to ensure that the proposed model is trained on diverse images, thereby enhancing its generalization capability.

Testing Datasets: To evaluate the model's generalization capability, our test dataset includes four test datasets corresponding to the training datasets (with no overlap in training images), supplemented with three additional test datasets: CASIAv1, Columbia, and NIST16. This setup enables a multi-faceted assessment of the model's performance, enhancing its ability to detect splicing tampering.

In summary, as shown in Table 2, this study uses seven datasets, with a total of over 120,000 images, encompassing mainstream datasets in the image tampering detection field, thereby fully validating the proposed algorithm's effectiveness and robustness.

4.3 Evaluation Metrics

This study employed multiple evaluation metrics to comprehensively assess the performance of the proposed algorithm. These evaluation metrics include image-level F1 score, recall, specificity, and AUC.

The image-level F1 score is commonly used for binary classification tasks to evaluate whether an image has been correctly classified. F1 calculation involves sensitivity (Sen, also known as recall) and precision, where sensitivity represents the proportion of true positives correctly identified by the model among all positive samples, while precision reflects the accuracy of the model's classification.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = Sen = \frac{TP}{TP + FN} = TPR \tag{3}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Specificity (Spe) is the proportion of true negative samples among all samples predicted by the model to be negative [26].

$$Spe = \frac{TN}{FP + TN} \tag{5}$$

AUC, the area under the receiver operating characteristic (ROC) curve, measures the area formed between the ROC curve and the coordinate axis, which can mitigate the impact of class imbalance [27]. The ROC curve plots recall on the x-axis and the false positive rate (FPR) on the y-axis [28].

Each metric used in this study highlights different performance aspects. For example, the F1 score provides a balanced evaluation of precision and recall in cases of class imbalance; sensitivity and specificity measure the model's handling of positive and negative samples, respectively; and AUC offers an intuitive, threshold-independent metric. By combining these metrics, the model's performance can be evaluated more comprehensively, objectively, and accurately from different perspectives.

4.4 Comparative and Ablation Experiments

4.4.1 Comparative Experiments

To comprehensively evaluate the performance of the proposed algorithm, we selected the deep learning-only methods FOCAL [21], PSCC-Net [29] and RRU-Net [15], as well as hybrid approaches combining deep learning and traditional methods, such as MMFusion [22], MVSS-Net [16] and NEDB-Net [30], for comparison. The number of parameters of each model is shown in Table 3.

Table 3: Comparison of the number of model parameters

NEDB-Net	RRU-Net	MVSS-Net	FOCAL	MMFusion	PSCC-Net	DDT-Net
45.08 M	4.10 M	146.88 M	340.88 M	0.56 M	3.67 M	45.83 M

Some examples of comparative experimental results are shown in Fig. 4. MVSS-Net and PSCC-Net demonstrated relative insensitivity to tampered regions during testing, while MMFusion, FOCAL, NEDB-Net and RRU-Net exhibited substantial omission and misdetection. In contrast, the proposed DDT-Net model achieved relatively better performance in actual tampering detection tests.

The relative superiority of DDT-Net in tampering detection can be attributed to several factors:

On one hand, PSCC-Net primarily relies on local spatial features for image tampering detection, excelling in identifying spatial structural changes in images. However, it largely relies on local spatial features, which may not effectively capture the subtle noise introduced by tampering. When tampering involves minor changes in the frequency domain of the image (such as JPEG compression artifacts), the detection performance significantly declines, making it difficult to recognize these concealed tampering traces.

On the other hand, MMFusion and MVSS-Net rely on simple feature concatenation to merge RGB and noise streams. While this strategy allows for the integration of features from different sources to some extent, it may lead to weak feature expression in complex scenarios. In contrast, DDT-Net's fusion strategy multiplies the downsampling results of each RGB and Bayar noise stream layer, achieving deeper feature fusion. This approach not only preserves the independent information of each stream but also effectively removes irrelevant features and enhances critical ones, thus highlighting tampered regions. This strategy, especially in multi-scale feature extraction and fusion, improves the model's robustness to complex tampering scenarios.

Lastly, NEDB-Net uses SRM as its noise stream and therefore performs better than RRU-Net, which lacks any traditional method incorporation. However, the Bayar noise stream outperformed SRM, which may explain why both models underperform compared to DDT-Net. FOCAL presents a special case. Although it utilizes HRNet and ViT as its backbones, it tends to perform pixel-level segmentation for each object during detection, which results in the generation of numerous false detection regions.

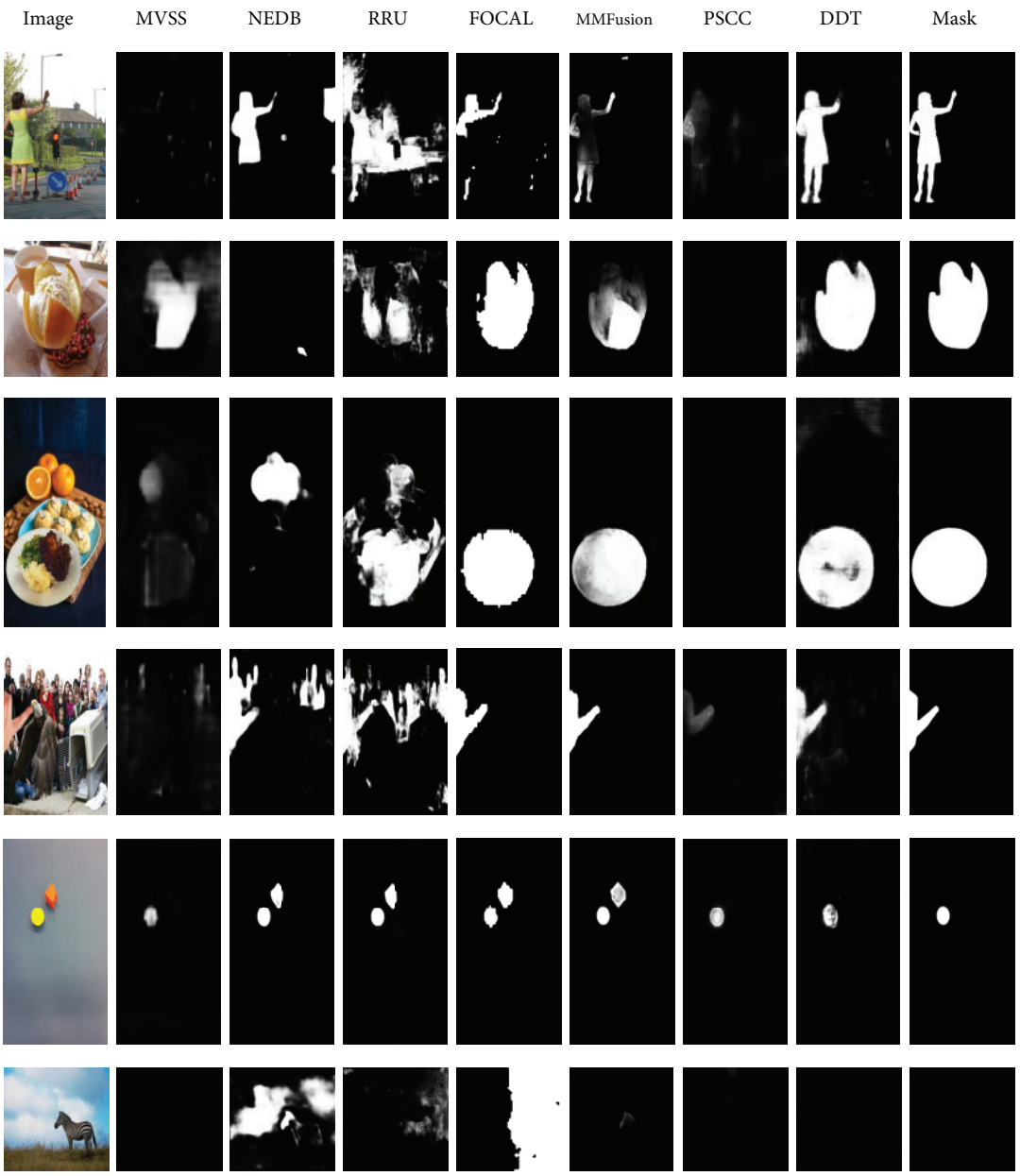

Figure 4: Some results of comparative experiments

In terms of image-level F1 scores, the experimental results in Table 4 indicate that the proposed method achieves higher F1 values than other algorithms across four public datasets: CASIAv1, DEFACTO, IMD, and DIS25k. Specifically, DDT-Net attains the highest F1 score (89) on the DEFACTO dataset, outperforming

the second-ranked by 30.76%, showcasing its robust capability in handling complex tampered images. Additionally, DDT-Net performs excellently on the DIS25k dataset, achieving an F1 score of 93.55, outperforming the second-ranked by 35.54%, further validating its robustness on large-scale datasets. On the Columbia and NIST16 datasets, DDT-Net achieved F1 scores of 66.3, only 2.2% and 1.95% below the second-ranked, respectively, indicating potential for improvement in handling large-object tampering patterns.

Table 4: Comparative experiment results. F1 decision threshold: 0.5

Method	DEFACTO				DIS25k				CASIAv1			
	F1	**AUC**	**Sen.**	**Spe.**	**F1**	**AUC**	**Sen.**	**Spe.**	**F1**	**AUC**	**Sen.**	**Spe.**
MVSS-Net [16]	35.62	0.62	85.14	22.52	34.14	0.69	83.72	21.44	75.34	0.84	61.55	97.11
NEDB-Net [30]	66.67	0.53	100	0	66.67	0.54	100	0	69.71	0.77	99.78	0.25
RRU-Net [15]	66.67	0.95	100	0	66.67	0.95	100	0	70.13	0.92	91.83	19.22
FOCAL [21]	66.67	0.5	100	0	66.67	0.5	100	0	66.67	0.5	100	0
PSCC-Net [29]	64.29	0.63	63	67	22.3	0.17	17.04	64.2	73.84	0.67	66.88	83.54
MMFusion [22]	68.2	0.98	98	10	69.02	1	100	10	67.25	1	100	3
DDT-Net (Ours)	**89.18**	0.89	91.24	86.62	**93.55**	0.99	98.92	87.44	**75.92**	0.88	87.79	50.13
Method	**IMD**				**Columbia**				**NIST16**			
	F1	**AUC**	**Sen.**	**Spe.**	**F1**	**AUC**	**Sen.**	**Spe.**	**F1**	**AUC**	**Sen.**	**Spe.**
MVSS-Net [16]	40.92	0.67	88.76	26.59	50	0.98	100	33	17.72	0.67	100	9.72
NEDB-Net [30]	66.67	0.58	100	0	66.67	0.62	100	0	66.67	0.52	100	0
RRU-Net [15]	66.67	0.93	100	0	66.67	0.92	100	0	66.67	0.9	100	0
FOCAL [21]	69.73	0.5	100	0	66.67	0.5	100	0	66.67	0.5	100	0
PSCC-Net [29]	56.19	0.47	47.57	78.28	66.79	0.75	100	0.56	**67.65**	0.9	95.14	13.89
MMFusion [22]	75.08	0.98	98	27.2	**67.8**	1	100	5	67.46	0.99	100	10
DDT-Net (Ours)	**67.83**	0.95	94.76	15.36	66.3	0.53	100	0	66.33	0.9	90.28	18.06

In terms of AUC, DDT-Net achieved high AUC values across various datasets, indicating its strong classification capability. For example, on the DEFACTO dataset, DDT-Net's AUC reached 0.89, while other methods mostly scored below 0.70, suggesting that DDT-Net can more accurately distinguish tampered from untampered regions.

In terms of Sen, DDT-Net displayed robust performance across most datasets, achieving a maximum score of 98.92% on DIS25k and exceeding 90% on other datasets, including DEFACTO, IMD, and Columbia, indicating a strong ability to detect positive samples.

In terms of Spe, DDT-Net performs unevenly. While it scored high on DEFACTO (86.62%) and DIS25k (87.44%), its performance was relatively weaker on IMD and Columbia, suggesting potential for improvement in the recognition of negative samples on certain datasets.

Overall, DDT-Net's superior performance in F1, AUC, and Sen demonstrates its notable advantages in detecting tampered regions and its overall classification capability. Although there is room for improvement in Spe and the number of model parameters, DDT-Net already outperforms most existing methods in general. In conclusion, despite minor fluctuations in specific metrics, DDT-Net maintains a well-balanced performance across all metrics, affirming its reliability as a detection tool.

4.4.2 Ablation Studies

To quantitatively assess the impact of the added noise stream in the U-Net 3+ model, ablation experiments were conducted, comparing F1 scores across six datasets with and without the noise stream.

As shown in Table 5, the inclusion of the Bayar noise stream increased F1 scores on most datasets, with the most significant improvement seen on the DIS25K dataset, where tampering edges are less pronounced. This indicates that the noise stream allows for a more profound capture of tampering traces, with the F1 score increasing by 39% when the noise stream is included.

Table 5: Ablation experiment results. F1 decision threshold: 0.5

Method	DEFACTO				DIS25k				CASIAv1			
	Img-F1	**AUC**	**Sen.**	**Spe.**	**Img-F1**	**AUC**	**Sen.**	**Spe.**	**Img-F1**	**AUC**	**Sen.**	**Spe.**
U-Net 3+	67.12	0.99	99	3.8	67.31	0.99	99	40	69.73	1	1	0
DDT-Net (U-Net 3++Bayar)	**89.18**	0.89	91	86.6	**93.55**	0.99	99	87	**75.92**	0.88	88	50.1
Method	**IMD**				**IMD**				**IMD**			
	Img-F1	**AUC**	**Sen.**	**Spe.**	**Img-F1**	**AUC**	**Sen.**	**Spe.**	**Img-F1**	**AUC**	**Sen.**	**Spe.**
U-Net 3+	66.84	0.99	99	40	**67.67**	1	1	40	65.98	1	1	10
DDT-Net (U-Net 3++Bayar)	**67.83**	0.95	95	15	66.3	0.53	100	0	**66.33**	0.9	9	18

In conclusion, extensive experimental results demonstrate the excellent performance of the proposed image tampering detection algorithm, which integrates deep learning with traditional methods, in terms of both detection accuracy and robustness. The results also validate the effectiveness of the proposed innovative mechanisms, indicating that this algorithm is a reliable and efficient solution to current image tampering detection challenges.

4.4.3 Robustness Evaluation

To evaluate robustness, we applied two common image processing methods, JPEG compression and Gaussian blur, to assess the performance of DDT-Net as well as other tampering detection models such as RRU-Net, NEDB-Net, PSCC-Net, and MVSS-Net.

1) JPEG Compression Robustness Analysis

Images often undergo multiple compressions during transmission, making robustness assessment essential for model stability. In the JPEG compression test (as shown in Fig. 5), DDT-Net demonstrated high robustness, with its F1 score declining only slightly from around 76% to approximately 70% as compression quality dropped from 100 to 50, showing a relatively gradual decrease. FOCAL, MMFusion, NEDB-Net, and RRU-Net also showed stability, with F1 scores staying around 70% across the full compression quality range, suggesting moderate robustness to compression. However, PSCC-Net was more sensitive, with a rapid drop in performance when the compression quality fell below 70, suggesting that it may lack sufficient robustness to compression artifacts, possibly because compression-induced quality loss was not fully considered in its design. MVSS-Net performed the worst in the compression test, with a sharp decline from 75% to 41% in F1 score as compression quality decreased, indicating that MVSS-Net struggles with low-quality compressed images, likely due to its heavy reliance on high-quality image features.

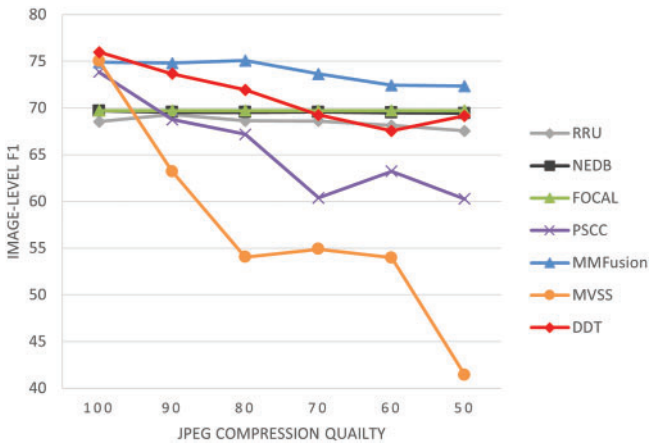

Figure 5: Robustness evaluation: JPEG compression on the CASIAv1 test dataset

2) Gaussian Blur Robustness Analysis

In the Gaussian blur test (as shown in Fig. 6), DDT-Net initially performed the best (with an F1 score of 75.92%). Although its performance declined with increasing kernel size, it generally maintained a high level. Under mild blurring (kernel size $\leq 5 \times 5$), DDT-Net's performance was almost unaffected, indicating its ability to maintain efficient detection under light blurring in case of daily occurrence. Under moderate blurring (kernel size 11×11), DDT-Net's performance even increased slightly (69.47%), reflecting good adaptability to moderate blurring. When the kernel size reached its maximum (29×29), DDT-Net's performance dropped to 54.29%, suggesting there is still room for improvement under extreme blurring. However, such extreme conditions are rare in practical applications, as image blur in daily life usually comes from slight motion, focus deviation, or low-quality capture, generally retaining most edges and details. Extreme blurring exceeds common practical requirements and is unlikely to significantly impact real-world performance.

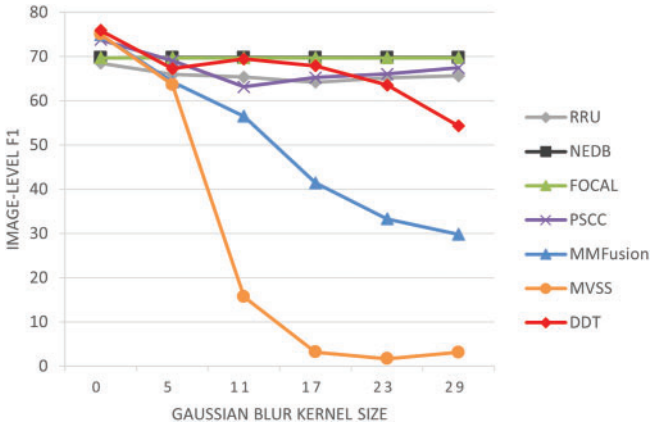

Figure 6: Robustness evaluation: gaussian blurring on the CASIAv1 test dataset

FOCAL and NEDB-Net demonstrated relatively stable performance across the entire blurring range, with F1 scores remaining around 70%. This stability is noteworthy and may indicate that they were specifically designed with robustness to blurring. RRU-Net and PSCC-Net performed well under mild blurring, but their performance declined with increased blurring. At a kernel size of 29, their F1 scores dropped to around

65% and 67%, respectively, indicating moderate adaptability to blurring but room for improvement under severe blurring. MMFusion and MVSS-Net was the most sensitive model in the blurring test; its performance declined sharply with kernel sizes over 11, reaching an F1 score close to 2% at a kernel size of 29 × 29. This suggests that they may be overly reliant on high-frequency image features, making it difficult to capture sufficient tampering traces under heavy blurring, leading to a performance collapse.

Although DDT-Net demonstrates high robustness under common degradation conditions such as JPEG compression and Gaussian blur, there is still significant room for improvement in model performance under more extreme degradation scenarios. Based on the findings of this study, we propose two key directions to enhance the robustness of image tampering detection models under extreme degradation conditions: data augmentation and transfer learning.

For data augmentation, introducing diverse extreme degradation operations during training—such as high-intensity JPEG compression, Gaussian blur, noise addition, and resolution reduction—can effectively improve the model's generalization capability. This approach allows the model to learn richer feature representations, enabling it to maintain stable performance when encountering unseen degradation conditions.

For transfer learning, fine-tuning models pre-trained on high-quality images to adapt to low-quality image environments can effectively leverage existing knowledge while accommodating the characteristics of new domains. Additionally, domain adaptation techniques can help the model maintain consistent performance across varying degrees of degradation, ensuring high detection accuracy even under extreme conditions.

In summary, this study's comprehensive evaluation of various image tampering detection models under different degradation conditions reveals that DDT-Net, FOCAL, and NEDB-Net exhibit high robustness. DDT-Net demonstrated strong adaptability to both Gaussian blurring and JPEG compression, highlighting its potential advantages in practical applications. Similarly, NEDB-Net showed consistent performance under both blurring and compression. In contrast, RRU-Net and PSCC-Net performed well under mild degradation but struggled under extreme conditions, while MMFusion and MVSS-Net displayed high sensitivity to image degradation, with significant performance declines. Future research should focus on exploring more effective data augmentation strategies and transfer learning methods to further enhance model robustness under extreme degradation conditions.

4.4.4 Analysis of False Detections and Missed Detections

False detections and missed detections are also key metrics for evaluating the performance of tampering detection models. Heatmaps provide an intuitive visualization of the model's focus on tampered regions, where brighter color gradients indicate a stronger tendency for the model to classify the area as tampered. As shown in the heatmap results in Fig. 7, DDT-Net significantly outperforms other methods in detecting tampered regions. Compared to actual tampered areas, the heatmaps generated by DDT-Net not only accurately cover the tampered regions but also excel in boundary recognition and detail preservation.

In contrast, other methods often exhibit missed detections or false detections across various scenarios, particularly in small-scale tampering and complex background settings, where their detection performance is notably inadequate. DDT-Net, however, demonstrates exceptional sensitivity to these details while maintaining minimal false labeling of non-tampered areas, fully showcasing its robust performance.

Notably, DDT-Net excels in complex scenarios. For example, in multi-object backgrounds (e.g., the human scene in the third row) and complex textured images (e.g., the food scene in the fourth row), DDT-Net accurately identifies tampered regions, effectively suppresses background interference, and produces clear

and precise detection results. This capability is markedly superior to other comparative methods, which often suffer from significant missed detections or blurry detection results in similar scenarios.

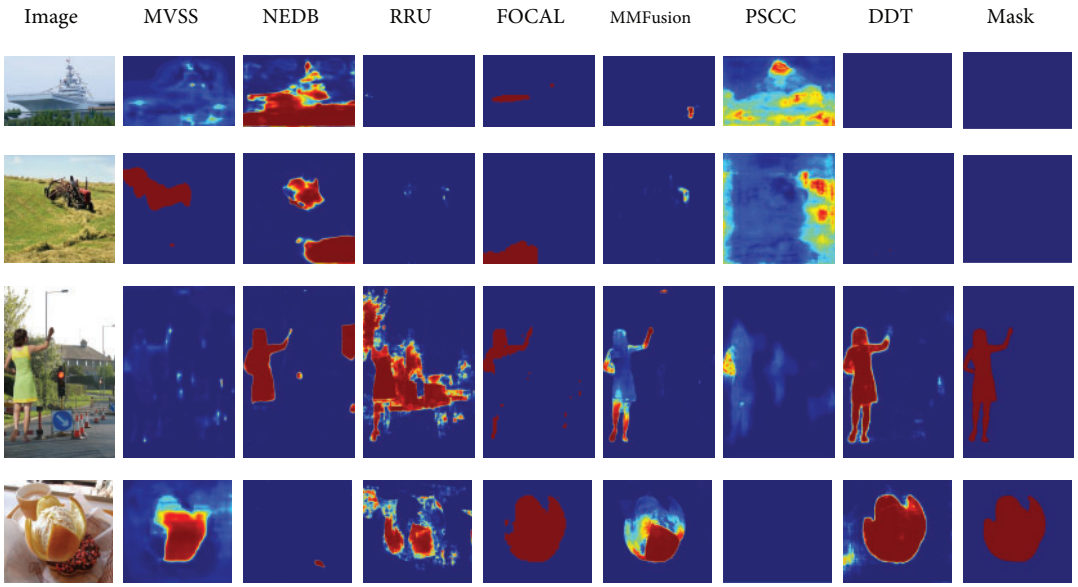

Figure 7: Heatmap of comparative experiments

5 Conclusions

In this paper, we propose an innovative image tampering detection method, DDT-Net, which integrates an enhanced U-Net 3+ architecture with the traditional Bayar noise stream. This approach effectively leverages deep learning models' capabilities in high-level semantic feature extraction while harnessing the Bayar noise stream's sensitivity to image noise characteristics, thereby achieving an organic fusion of multiple strategies. Compared with traditional multi-scale feature fusion methods, DDT-Net demonstrates significant advantages in two aspects:

1) Enhanced Scale Feature Fusion: Conventional U-Net architectures primarily utilize RGB streams for multi-scale feature fusion, implementing layer-wise downsampling and upsampling for feature integration across scales. However, RGB features alone prove insufficient for capturing subtle noise characteristics in tampered images. The incorporated Bayar noise stream in DDT-Net specifically addresses image noise processing, establishing a novel approach for minute tampering trace detection, particularly enhancing detection capability in regions with significant noise features.
2) Dual-Stream Complementarity: DDT-Net implements a synergistic fusion of structural and noise features through the integration of RGB and Bayar noise streams. The RGB stream processes color and structural information, while the Bayar noise stream specializes in noise pattern detection. Performing feature multiplication after downsampling allows comprehensive tampering detection that relies on both RGB stream and noise features, thereby improving detection accuracy for visually subtle manipulations.

In conclusion, the multi-stream architecture of DDT-Net significantly enhances image tampering detection accuracy and robustness by leveraging multi-scale feature fusion, enhanced edge and detail

processing, and precise sensitivity to minor tampering traces. Experimental results demonstrate that DDT-Net achieves superior performance over current state-of-the-art methods across multiple public datasets, validating the effectiveness of the multi-strategy fusion approach. This organic integration of deep learning with traditional noise detection methods not only holds broad application prospects in image tampering detection but also provides new directions and insights for other visual tasks, such as video analysis and object detection.

6 Future Research Directions and Improvement Suggestions

In future research, DDT-Net can be extended to video tampering detection, enabling precise localization of tampered regions through temporal feature analysis. We can also explore multimodal data fusion strategies to integrate information from audio, text, and other dimensions, thereby enhancing its cross-domain application capabilities. On the model optimization side, lightweight design can be prioritized by optimizing the network structure, reducing parameter counts, and minimizing computational overhead, ensuring efficient deployment on resource-constrained devices. Moreover, to address extreme degradation conditions, introducing diverse data augmentation techniques—such as high-intensity JPEG compression, Gaussian blur, noise addition, and resolution reduction—during training can significantly enhance the model's generalization and robustness. Additionally, transfer learning can help the model maintain consistent performance across various levels of degradation, offering robust solutions for tampering detection in challenging environments.

In terms of model design, the noise stream's feature extraction capabilities can be improved by integrating multiple noise detection methods to enrich feature diversity. Advanced attention modules can be incorporated into the existing U-Net 3+ architecture to enhance the model's ability to capture and focus on critical features. Furthermore, combining adversarial training with augmentation techniques like color transformation can further improve the model's adaptability and resilience in complex tampering scenarios. Through these optimizations and expansions, DDT-Net is expected to achieve greater technological breakthroughs in the fields of image and video tampering detection, while also providing valuable research insights and technical support for related areas such as computer vision and image security.

Acknowledgement: The authors gratefully acknowledge the computational support provided by the Advanced Computing Center of China Three Gorges University.

Funding Statement: This work was supported by National Natural Science Foundation of China (No. 61502274).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Jim Wong, Zhaoxiang Zang; data collection: Jim Wong; analysis and interpretation of results: Jim Wong; draft manuscript preparation: Jim Wong, Zhaoxiang Zang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Bayar B, Stamm MC. Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection. IEEE Trans Inf Forensics Secur. 2018;13(11):2691–706. doi:10.1109/TIFS.2018.2825953.

2. Shi X, Li P, Wu H, Chen Q, Zhu H. A lightweight image splicing tampering localization method based on MobileNetV2 and SRM. IET Image Process. 2023;17(6):1883–92. doi:10.1049/ipr2.12763.
3. Zhou P, Han X, Morariu VI, Davis LS. Learning rich features for image manipulation detection. In: Proceeding IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 1053–61. doi:10.1109/CVPR.2018.00116.
4. Zhang YX, Zhao XF, Cao Y. A review on blind detection of digital image tampering. J Inf Secur. 2023;2022(12):100–10. (In Chinese). doi:10.19363/J.cnki.cn10-1380/tn.2022.05.05.
5. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. UNet 3+: a full-scale connected UNet for medical image segmentation. In: Proceeding IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); May 2020; Barcelona, Spain. p. 1055–9. doi:10.1109/ICASSP40776.2020.9053405.
6. Al-Shamasneh AR, Ibrahim RW. Image splicing forgery detection using feature-based of sonine functions and deep features. Comput Mater Contin. 2024;78(1):795–810. doi:10.32604/cmc.2023.042755.
7. Gu F, Dai Y, Fei J, Chen X. Deepfake detection and localisation based on illumination inconsistency. Int J Auton Adapt Commun Syst. 2024;17(4):352–68. doi:10.1504/IJAACS.2024.139383.
8. González Fernández E, Sandoval Orozco AL, García Villalba LJ. A multi-channel approach for detecting tampering in colour filter images. Expert Syst Appl. 2023;230:120498. doi:10.1016/j.eswa.2023.120498.
9. Xu Y, Irfan M, Fang A, Zheng J. Multiscale attention network for detection and localization of image splicing forgery. IEEE Trans Instrum Meas. 2023;72:5026315. doi:10.1109/TIM.2023.3300434.
10. Sujin JS, Sophia S. High-performance image forgery detection via adaptive SIFT feature extraction for low-contrast or small or smooth copy-move region images. Soft Comput. 2024;28:437–45. doi:10.1007/s00500-023-08209-6.
11. Wang K, Xia X, Zhang Z, Gao T. Hashing-based remote sensing image tamper detection system. Digit Signal Process. 2023;140:104101. doi:10.1016/j.dsp.2023.104101.
12. Kwon MJ, Nam SH, Yu IJ, Lee HK, Kim C. Learning JPEG compression artifacts for image manipulation detection and localization. Int J Comput Vis. 2022;130(8):1875–95. doi:10.1007/s11263-022-01617-5.
13. Wang J, Huang W, Luo X, Shi YQ, Jha SK. Detecting non-aligned double JPEG compression based on amplitude-angle feature. ACM Trans Multimedia Comput Commun Appl. 2021;17(4):138. doi:10.1145/3464388.
14. Mehrjardi FZ, Latif AM, Zarchi MS, Sheikhpour R. A survey on deep learning-based image forgery detection. Pattern Recognit. 2023;144(3):109778. doi:10.1016/j.patcog.2023.109778.
15. Bi X, Wei Y, Xiao B, Li W. RRU-Net: the ringed residual U-Net for image splicing forgery detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun; Long Beach, CA, USA. p. 30–9. doi:10.1109/CVPRW.2019.00010.
16. 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. 2023;45(3):3539–53. doi:10.1109/TPAMI.2022.3180556.
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv: 1706.03762. 2017.
19. Atak IG, Yasar A. Image forgery detection by combining visual transformer with variational autoencoder network. Appl Soft Comput. 2024;165:112068. doi:10.1016/j.asoc.2024.112068.
20. Guillaro F, Cozzolino D, Sud A, Dufour N, Verdoliva L. Trufor: leveraging all-round clues for trustworthy image forgery detection and localization. In: Proceeding IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Vancouver, BC, Canada. p. 20606–15.
21. Wu H, Chen Y, Zhou J. Rethinking image forgery detection via contrastive learning and unsupervised clustering. arXiv:2308.09307. 2023.
22. Triaridis K, Mezaris V. Exploring multi-modal fusion for image manipulation detection and localization. In: Proceeding 30th International Conference on Multimedia Modeling (MMM 2024); 2024 Jan–Feb; Cham, Switzerland. p. 198–211.
23. Xiong L, Xu J, Yang CN, Zhang X. CMCF-Net: an end-to-end context multiscale cross-fusion network for robust copy-move forgery detection. IEEE Trans Multimedia. 2024 Dec;26:6090–101. doi:10.1109/TMM.2023.3345160.

24. Ding H, Chen L, Tao Q, Fu Z, Dong L, Cui X. DCU-Net: a dual-channel U-shaped network for image splicing forgery detection. Neural Comput Appl. 2023;35(7):5015–31. doi:10.1007/s00521-021-06329-4.
25. Barglazan A-A, Brad R, Constantinescu C. Image inpainting forgery detection: a review. J Imaging. 2024;10(2):42. doi:10.3390/jimaging10020042.
26. Baumy A, Algarni AD, Abdalla M, El-Shafai W, El-Samie A, Fathi E, et al. Efficient forgery detection approaches for digital color images. Comput Mater Contin. 2022;71(2):3257–76. doi:10.32604/cmc.2022.021047.
27. Xu Y, Zheng J, Fang A, Irfan M. Feature enhancement and supervised contrastive learning for image splicing forgery detection. Digit Signal Process. 2023;136(5):104005. doi:10.1016/j.dsp.2023.104005.
28. El Biach FZ, Iala I, Laanaya H, Minaoui K. Encoder-decoder based convolutional neural networks for image forgery detection. Multimedia Tools Appl. 2022;81(16):22611–28. doi:10.1007/s11042-020-10158-3.
29. 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. 2022;32(11):7505–17. doi:10.1109/TCSVT.2022.3189545.
30. Zhang Z, Qian Y, Zhao Y, Zhang X, Zhu L, Wang J, et al. Noise and edge based dual branch image manipulation detection. In: Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things; 2023; Xiamen, China: Association for Computing Machinery. p. 963–8. doi:10.1145/3603781.3604221.