A Transformer-Based Image Manipulation Localization Network with Dual-Frequency Attention and Multi-Scale Feature Fusion

Feijiang Chen, Jia Wang*, Qianxi Pan
School of Computer Science and Technology
Xinjiang University
Urumqi, China
107552304030@stu.xju.edu.cn, jw1024@xju.edu.cn, 107552304117@stu.xju.edu.cn
*Corresponding author: Jia Wang

***Abstract*—Image manipulation detection and localization aims to accurately identify and annotate tampered regions within images and represents a critical research direction in multimedia forensics. Although deep learning-based methods have achieved significant progress in image manipulation localization tasks in recent years, existing models still tend to suffer from missed detections, false alarms, and inaccurate boundary localization when dealing with tampered regions of diverse scales, as well as scenarios where manipulation boundaries are highly blended with complex backgrounds and manipulation traces are concealed. To address these issues, this paper proposes an image manipulation localization network based on Dual-Frequency attention and Multi-scale feature fusion (DFM-Net). The proposed method replaces the conventional multi-head self-attention in Transformers with a Dual-Frequency Attention (DFA), where two branches focus on local fine-grained details and global structural information, respectively, thereby enhancing the perception of tampering clues across different scales. Furthermore, a Feature Enhancement Aggregation (FEA) module is designed to apply different forms of enhancement to RGB features and noise features and incorporate a coordinate attention mechanism, enabling more effective utilization of RGB and noise features for precise tampering boundary localization. Experiments on CASIA, NIST16, Columbia, and Coverage demonstrate that DFM-Net achieves the highest average F1 score of 67.6% and AUC of 92.0%, higher than the baseline models by 3.4% and 1.2%, respectively.**

***Index Terms*—image manipulation detection, image manipulation localization, dual frequency attention, multi scale feature fusion, transformer**

I. Introduction

With the rapid advancement of image editing and generation technologies, image tampering has become increasingly prevalent in our daily lives, and such manipulation has grown ever more convenient and covert. At the same time, the widespread application of image editing and generation technologies has also introduced potential security risks. Particularly in news reporting scenarios [1], tampered images may disseminate false information and mislead public opinion. Therefore, accurately identifying and localizing image tampering regions is of significant practical importance. Using sophisticated image processing software, attackers can perform sophisticated image tampering through three common methods [2], namely splicing, copy-move, and removal, while maintaining high visual authenticity.

Although existing image manipulation localization methods have made progress, they still face multiple challenges when handling complex tampered images. For instance, different types of tampering exhibit significant differences in spatial scale and manifestation. Small-scale local tampering often produces subtle artifacts like edge discontinuities or texture anomalies in confined areas [3], while large-scale or semantic-level tampering manifests more as inconsistencies in lighting, color style, or overall structural relationships [4]. This cross-scale feature disparity makes it difficult for a single modeling strategy to simultaneously account for local details and global structural or semantic relationships. Meanwhile, when tampered region boundaries are highly blended with complex backgrounds, they often lack clear gradient discontinuities and exhibit progressive transition characteristics, which poses challenges for precise boundary localization. These factors jointly limit the detection accuracy and generalization ability of existing tampering detection methods in multi-type and multi-scale scenarios, thereby necessitating more effective feature modeling and fusion mechanisms.

Based on the above challenges, this paper proposes an image manipulation localization network based on Dual-Frequency attention and Multi-scale feature fusion (DFM-Net). To improve the perception of tampering clues at different scales, a Dual-Frequency Attention (DFA) module is introduced to enhance the modeling of both local details and global semantic information. In addition, to address the issue of blurred boundaries in image manipulation localization, a Feature Enhancement and Aggregation (FEA) module is designed to strengthen features related to image manipulation and promote more effective feature fusion. By combining the DFA and FEA modules, the proposed method achieves more accurate localization of tampered regions under complex scenarios. In summary, our contributions can be outlined as follows:

- This paper proposes a dual-branch DFA module to replace the traditional multi-head self-attention mechanism.

By employing dual-branch attention, DFA captures both local fine-grained features and global semantic information, enabling more accurate detection of tampered regions across different scales.

- This paper designs a FEA module. This module enhances RGB features by amplifying high-frequency wavelet subbands and reinforcing edge variations with Sobel filters, while simultaneously strengthening noise features. Coordinate attention enables adaptive fusion, thereby obtaining more precise tampering boundaries.

The remainder of this paper is organized as follows. Section II reviews related work on image tampering localization. Section III describes the proposed method in detail. Experimental results and corresponding analyses are presented in Section IV. Finally, Section V concludes this paper.

II. Related Work

With the rapid development of deep learning techniques, data driven image manipulation detection methods have attracted increasing attention. Deep learning based methods can automatically learn hierarchical feature representations from large scale datasets. According to network architecture design, deep learning based image manipulation detection methods can be broadly categorized into single branch and multi branch frameworks [5].

A. Single Branch Frameworks

Single branch frameworks typically take RGB images as input and perform end to end tampering localization through deep neural networks, such as convolutional neural networks and Transformers. Wu et al. [6] proposed a representative single branch framework that models local statistical discrepancies between tampered and authentic regions by learning manipulation related artifacts, enabling unified detection of multiple manipulation types. Zhuang et al. [7] further enhanced feature fusion capability through densely connected fully convolutional architectures, improving localization accuracy to some extent. With the success of Transformer based models in visual tasks, Ma et al. [8] introduced global self attention mechanisms to enhance long range dependency modeling, improving the detection of large scale manipulated regions.

Although single branch deep learning methods reduce dependence on handcrafted features and improve detection accuracy, their reliance on a single RGB feature representation limits modeling flexibility and hinders effective characterization of noise inconsistencies, frequency anomalies, and subtle manipulation cues across diverse manipulation types.

B. Multi Branch Frameworks

To overcome the limitations of single feature modeling, researchers have gradually introduced multi branch deep learning architectures that integrate multi modal or multi view features for improved manipulation localization performance. Dong et al. [9] proposed an early multi branch framework by introducing an edge supervision branch and progressive feature fusion to enhance boundary responses, together with dual attention mechanisms to improve feature selectivity. Guillaro et al. [10] incorporated an auxiliary noise residual branch and adaptive fusion strategy to improve overall detection performance. Lin et al. [11] combined convolutional and Transformer based feature extraction with boundary enhancement and supervision strategies to improve boundary localization accuracy. Shi et al. [12] further integrated semantic and noise features through dual branch architectures and attention mechanisms, improving localization accuracy and boundary clarity. Bai et al. [13] explored manipulation traces from pixel correlation and regional attribute perspectives through inconsistency mining and progressive fusion refinement strategies, improving localization accuracy while reducing semantic interference.

Although multi-branch deep learning methods alleviate the limitations of single-feature modeling by incorporating semantic, noise, and boundary cues, existing approaches often still emphasize a specific type of feature. As a result, their ability to collaboratively model multi-level information remains limited, making it difficult to simultaneously capture fine-grained local details and global structural context. Moreover, insufficient interaction among different branches further constrains the effective utilization of multi-source features.

III. METHOD

A. Overall Architecture

Considering that image manipulation operations not only affect visual appearance but also disrupt local noise consistency, the proposed method adopts a dual-stream architecture consisting of an RGB branch and a noise branch. The original RGB image is fed into the RGB branch, while noise features extracted by a channel-wise high-pass filter are input into the noise branch. Each branch employs SegFormer as the backbone network, with Mix Transformer (MiT-B2) serving as the encoder. Features are extracted through four hierarchical stages to obtain multi-scale representations with rich semantic information. To enhance the capability of Transformers in detecting manipulations at different scales, the standard Transformer blocks in SegFormer are modified by replacing the conventional multi-head self-attention mechanism with the proposed Dual-Frequency Attention (DFA) module, forming new DFA Transformer blocks. In addition, a Feature Enhancement Aggregation (FEA) module is designed to strengthen

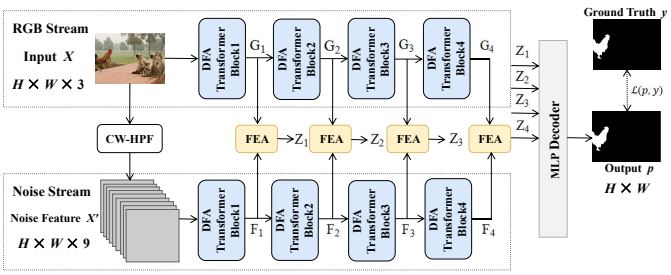

Fig. 1. Overall architecture of the proposed DFM-Net.

edge and texture information, improve boundary clarity, and enhance the complementarity and correlation between RGB and noise features. At each scale, features from the two branches are enhanced and fused, followed by coordinate attention to guide effective collaboration across branches. Finally, the fused multi-scale features are fed into an MLP decoder for upsampling to generate the final manipulation prediction mask. During training, Dice Loss and Focal Loss are jointly used for optimization. The overall architecture of the proposed DFM-Net is illustrated in Fig. 1.

B. Dual-Frequency Attention Module

The multi-head self-attention mechanism in Transformers typically applies the same processing to all attention heads, causing the focus ranges and priorities of different heads to gradually converge. This paper rationally divides the number of attention heads: one portion is proportionally allocated to the high-frequency branch for detecting local high-frequency features; the remaining attention heads are assigned to the low-frequency branch to capture global low-frequency semantic information.

Specifically, given an input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times D}$, where D denotes the channel dimension of each token, the number of attention heads is denoted as N_h, and the dimension of each attention head is $d_h = D/N_h$. Query, Key, and Value matrices are generated through linear projections.The attention heads are divided according to a predefined ratio β. Specifically, βN_h heads are assigned to the high-frequency branch, while $(1-\beta)N_h$ heads are allocated to the low-frequency branch. The architecture of the DFA module is illustrated in Fig. 2.

The high-frequency branch adopts local-window self-attention [14]. Instead of performing global attention on the entire feature map, the input feature is partitioned into non-overlapping windows of size $s \times s$. Given the input feature $\mathbf{X} \in \mathbb{R}^{H\times W\times D}$, the number of windows is $M = \frac{H}{s} \times \frac{W}{s}$.Let $\mathbf{X}_i \in \mathbb{R}^{s^2\times D}$ denote the token set of the i-th window.For the high-frequency branch, the corresponding Query, Key, and Value for the h_i-th attention head in the i-th window are denoted as: $\mathbf{Q}_{h_i}^{(i)}$, $\mathbf{K}_{h_i}^{(i)}$, and $\mathbf{V}_{h_i}^{(i)}$, respectively. Scaled dot-product attention [15] is computed within each window as

$$\mathrm{Attn}_{h_i}^{(i)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{h_i}^{(i)}\left(\mathbf{K}_{h_i}^{(i)}\right)^{\top}}{\sqrt{d_h}}\right)\mathbf{V}_{h_i}^{(i)} \tag{1}$$

After computing all heads, the outputs are concatenated along the channel dimension to form the high-frequency window output. Then , the outputs of all windows are rearranged to their original spatial positions to obtain the high-frequency branch output.

The low-frequency branch performs window-wise average pooling on the input features to preserve global semantics. To reduce computational cost and improve efficiency, the proposed pooling strategy uses pooled Key tokens and Value tokens, while Query tokens remain at the original resolution

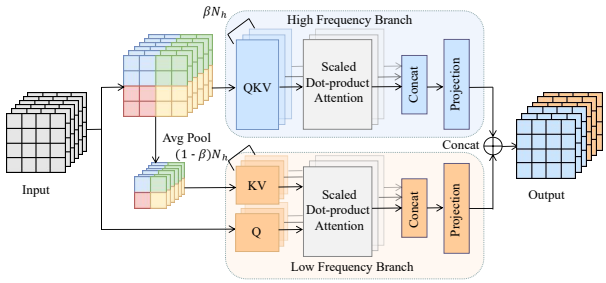

Fig. 2. Architecture of the proposed DFA module.

to ensure spatial alignment between the output and the input. Global attention is then computed to capture large-range structural and semantic information. Specifically, given the input feature, we first apply average pooling to Key and Value within each window to reduce the spatial resolution. Then, $\mathrm{Attn}_{lo}^{(i)}$ is obtained through scaled dot-product attention.Similar to the high-frequency branch, the output is obtained through concatenation and linear mapping.

Finally, the features computed from the high-frequency and low-frequency branches are concatenated along the channel dimension, ultimately yielding features that integrate both local and global information.

C. Feature Enhancement Aggregation Module

To address the issue of blurred boundaries in image tampering localization, this paper designed an FEA module. The FEA module enhances the RGB and noise features separately, then uses coordinate attention for adaptive fusion. The architecture of the FEA module is illustrated in Fig. 3.

For RGB features, this paper first introduces a layer of Discrete Wavelet Transform (DWT) to decompose the input features into four subbands: LL, LH, HL, and HH [16]. This process can be formulated as:

$$\mathrm{DWT}(\mathbf{X}_{\mathrm{RGB}}) = \{\mathbf{X}_{LL}, \mathbf{X}_{LH}, \mathbf{X}_{HL}, \mathbf{X}_{HH}\}. \tag{2}$$

Wavelet decomposition separates structural information of different orientations and frequencies while preserving spatial correspondences, laying the foundation for subsequent directional enhancement. In tamper detection scenarios, the LH and HL subbands correspond to high-frequency responses in the vertical and horizontal directions, respectively. Since tampering operations often introduce discontinuous horizontal or vertical structural changes at target boundaries, the LH and HL subbands are the most sensitive frequency bands to tampering boundaries. Leveraging this characteristic, this paper applies High-Frequency Subband Enhancement (HFSE) to the LH and HL subbands, applying a linear enhancement coefficient e to amplify boundary inconsistencies in different directions. Specifically, the HFSE operation is formulated as:

$$\mathbf{X}'_{LH} = e \cdot \mathbf{X}_{LH}, \quad \mathbf{X}'_{HL} = e \cdot \mathbf{X}_{HL}. \tag{3}$$

After completing the high-frequency subband enhancement, the features are reconstructed back into the spatial domain via

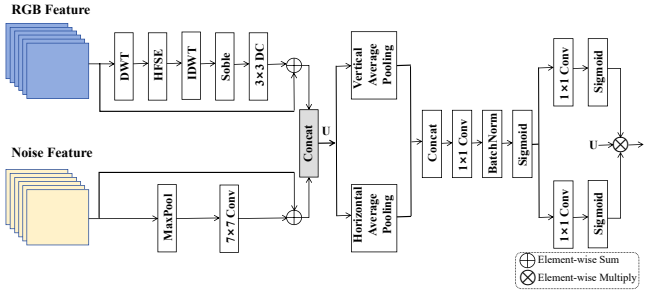

Fig. 3. Architecture of the proposed FEA module. DWT denotes Discrete Wavelet Transform, HFSE denotes High-Frequency Subband Enhancement, IDWT denotes Inverse Discrete Wavelet Transform, and 3×3 DC denotes Dynamic Convolution.

Inverse Discrete Wavelet Transform (IDWT). Subsequently, a Sobel filter is introduced to further enhance gradient information. The Sobel filter directly responds to local grayscale variations, effectively highlighting gradient jumps at edges. Positioning the Sobel filter after wavelet enhancement allows for a secondary reinforcement of boundary separability on top of the already amplified directional high-frequency structures, making tampered edges more prominent in the feature space. The processed features are then fed into a 3×3 dynamic convolution. Finally, residual connections aggregate the processed features with the input RGB features.

For noise features, this paper employs a combined structure of max pooling and large convolutional kernels. Max pooling preserves more pronounced anomalous responses, highlighting the statistical inconsistencies in noise introduced by tampering operations. Subsequently, a 7×7 convolution is applied to expand the receptive field, capturing broader contextual information. Finally, the processed features are aggregated with the input noise features via residual connections.

After that, the processed RGB features and noise features are concatenated to obtain the fused feature U. Next, a coordinate attention mechanism is introduced to model the long-range dependencies among features along the height and width directions, generating directional attention weights to adaptively enhance the fused features. In this way, the module can highlight key cues such as boundary responses and noise anomalies while preserving positional information, and the enhanced output feature is obtained by $Z = U \otimes A_h \otimes A_w$, thereby achieving clearer and more accurate localization of tampered regions.

IV. Experiments

A. Dataset

We conduct experiments on four benchmark datasets, including CASIA [17], NIST16 [18], Columbia [19], and Coverage [20]. The CASIA dataset includes two versions: CASIA v2.0 contains 5,163 tampered images and CASIA v1.0 contains 920 tampered images. To evaluate generalization capability, we use CASIA v2.0 for training and CASIA v1.0 for testing. To ensure fair comparisons, all experiments on each dataset follow the same training and testing splits. The dataset statistics are summarized in Table I.

TABLE I
Dataset statistics.

Dataset	Total	Involved manipulation technique		
		Copy-move	Splicing	Removal
CASIA	6043	✓	✓	×
NIST16	564	✓	✓	✓
Columbia	180	×	✓	×
Coverage	100	✓	×	×

B. Implementation Details

All input images are resized to 512×512 and preprocessed using data augmentation strategies, including random flipping, scaling, and cropping. The model is trained for 100 epochs with a batch size of 8. The initial learning rate is set to 5×10^{-3} and gradually decayed to 5×10^{-4} using a cosine decay strategy. The AdamW optimizer is employed with a momentum coefficient of 0.9. The parameter β is set to 0.6, 0.5, 0.4, and 0.3 for the four stages of the backbone network, respectively. The window size s is set to 2, and the linear enhancement coefficient e is set to 1.8. For the loss function, the hyperparameters are empirically set to $\alpha = 0.5$ and $\gamma = 2$.

C. Evaluation Metrics

This study employs pixel-level F1 score and AUC as evaluation metrics for image tampering localization tasks. The F1 score comprehensively measures both accuracy and completeness of detection results, making it suitable for identifying tampered pixels under imbalanced data conditions [21]. AUC, meanwhile, reflects the model's overall ability to distinguish between tampered and non-tampered pixels across different thresholds [22].

D. Experimental Results and Analysis

1) Comparison Experiments: To comprehensively evaluate the performance of the proposed image manipulation localization model DFM-Net, we conduct comparative experiments on four widely used public benchmark datasets, including CASIA, NIST16, Columbia, and Coverage. The proposed method is compared with several state-of-the-art image manipulation localization approaches. The comparison methods can be broadly divided into two categories: single-branch networks and multi-branch networks. Single-branch networks include ManTra-Net [6] and IML-ViT [8]. Multi-branch methods include MVSS-Net [9], EMT-Net [11], TruFor [10], DAE-Net [12], and PIM-Net [13]. Quantitative comparison results in terms of pixel-level F1 score are reported in Table II.The AUC (%) comparison results for image tampering localization are shown in Table III.

For single-branch networks, ManTra-Net primarily focuses on modeling local statistical anomalies, which can lead to false positives or inaccurate localization in large-scale tampering scenarios. IML-ViT enhances long-range dependency modeling through a global self-attention mechanism, but its ability

TABLE II
COMPARISON OF PIXEL-LEVEL F1 SCORE (%) ON DIFFERENT DATASETS.

Method	CASIA	NIST16	Columbia	Coverage	Average
ManTra-Net	22.3	46.2	65.4	28.3	40.6
MVSS-Net	39.0	82.7	70.3	28.4	55.1
EMT-Net	45.9	82.5	56.1	35.3	55.0
TruFor	44.5	36.0	56.6	31.9	42.3
IML-ViT	46.5	82.8	80.1	41.4	62.7
DAE-Net	49.4	87.1	87.2	33.0	64.2
PIM-Net	51.2	84.2	79.2	40.2	63.7
Ours	**51.8**	**87.9**	**88.3**	**42.5**	**67.6**

TABLE III
COMPARISON OF PIXEL-LEVEL AUC (%) ON DIFFERENT DATASETS.

Method	CASIA	NIST16	Columbia	Coverage	Average
ManTra-Net	64.8	80.5	82.4	77.7	76.4
MVSS-Net	84.7	98.1	86.9	80.8	87.6
EMT-Net	85.6	96.7	83.2	81.2	86.7
TruFor	84.4	80.6	86.9	83.7	83.9
IML-ViT	78.4	85.3	88.4	85.1	84.3
DAE-Net	80.5	98.4	97.3	74.1	87.6
PIM-Net	89.2	98.1	93.5	82.2	90.8
Ours	**90.4**	**98.4**	**97.5**	**85.7**	**92.0**

to detect small-scale tampered regions remains insufficient. By introducing a dual-branch architecture that jointly exploits RGB features and noise features, and performing feature fusion across multiple scales, DFM-Net enables the simultaneous utilization of visual information and noise anomaly cues, thereby improving adaptability to complex tampering scenarios. Experimental results show that DFM-Net consistently outperforms representative single-branch methods across all four datasets.

For multi-branch networks, MVSS-Net mainly focuses on spatial and edge cues, but lacks explicit modeling of frequency-domain inconsistencies, making it less effective for detecting subtle or small-scale manipulations. EMT-Net and DAE-Net both mainly focus on boundary artifact enhancement; however, their detection performance may decline when the authentic background is highly blended with the tampered regions. TruFor integrates multiple forensic cues through noise residual modeling, yet its feature fusion is mainly performed at high semantic levels, which may limit its sensitivity to fine-grained or small-scale tampering regions. PIM-Net exploits progressive inconsistency mining to improve localization accuracy, but it may suffer from misclassification in large-area tampering scenarios. Compared to existing multi-branch image tampering detection methods, DFM Net achieves better detection performance in multi-scale and multi type tampering scenarios by proposing DFA and FEA modules. In particular, experiments on the CASIA dataset indicate that DFM-Net exhibits stronger generalization capability.

2) Ablation Experiments: To verify the effectiveness of the proposed DFA and FEA modules, we conduct systematic ablation studies. The baseline model consists of the RGB branch, the noise branch, and the corresponding connection modules. Then, the DFA and FEA modules are gradually added to the baseline for comparison. All experiments are performed under the same datasets and training settings. The ablation results are reported in Table IV.

TABLE IV
ABLATION STUDY RESULTS ON FOUR DATASETS.

Methods	CASIA		NIST16		Columbia		Coverage	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Baseline	48.2	87.7	83.8	95.2	87.1	95.8	38.2	81.2
Baseline + DFA	51.5	89.8	86.4	98.1	87.2	97.1	39.1	83.1
Baseline + FEA	50.4	90.1	87.1	96.5	87.8	96.2	42.2	84.9
Baseline + DFA + FEA	**51.8**	**90.4**	**87.9**	**98.4**	**88.3**	**97.5**	**42.5**	**85.7**

After incorporating the DFA module into the baseline model, both F1 score and AUC exhibit consistent improvements. Notably, the F1 score shows significant gains on the CASIA and NIST datasets, which contain a large number of small-scale tampering samples. This indicates that the DFA module effectively enhances the model's sensitivity to locally anomalous regions, thereby alleviating missed detections in small tampered areas.

The introduction of the FEA module to the baseline model also leads to considerable performance improvements. Results on the NIST and Coverage datasets, which are characterized by complex boundaries and prominent noise cues, demonstrate that the FEA module effectively improves boundary localization accuracy and reduces false positives in tampered regions.

When both DFA and FEA modules are integrated, the model achieves the best performance across all four datasets. This result highlights their complementary roles: DFA strengthens the representation of multi-scale tampering clues during feature encoding, while FEA further enhances boundary perception and cross-modal feature coordination during feature fusion. Their combined effect significantly improves the overall accuracy of tampered region localization.

3) Visualization Analysis: To provide an intuitive comparison of the localization performance of DFM-Net, Fig. 4 presents qualitative results on representative manipulated images selected from four benchmark datasets. The ground truth annotations, the prediction results of representative methods including MVSS-Net and IML-ViT, and the results produced by the proposed DFM-Net are displayed for side-by-side comparison. As illustrated in the figure, DFM-Net consistently

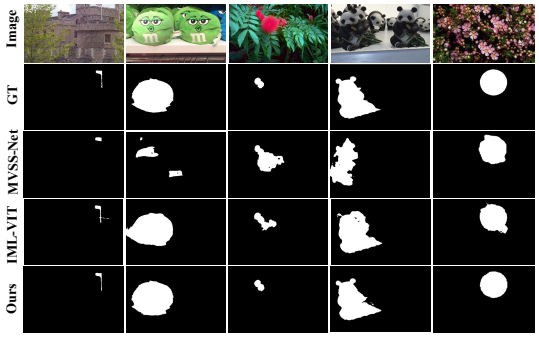

Fig. 4. Qualitative comparison of image manipulation localization results on representative samples.

achieves more accurate localization results across different datasets. Whether dealing with small-scale or large-scale manipulated regions, the proposed method is able to precisely identify tampered areas while maintaining clear and well-aligned boundaries. In addition, even in challenging scenarios where tampered regions are highly blended with complex backgrounds and boundary cues are deliberately weakened, DFM-Net still produces reliable localization results. These qualitative comparisons further verify the effectiveness of the proposed method in handling diverse and complex image manipulation scenarios.

V. Conclusion

This paper addresses the challenges of image manipulation localization in complex real-world scenarios, including diverse tampering scales, blurred boundaries, and covert manipulations, by proposing DFM-Net. Experimental results on four public datasets demonstrate that DFM-Net outperforms multiple mainstream methods in terms of both F1 score and AUC. Comparative and ablation experiments validate the effectiveness and synergy of the proposed modules, indicating that the method achieves strong performance and generalization capability in multi-type and multi-scale image manipulation detection tasks.Although favorable experimental results are obtained, future work will further explore the adaptability of DFM-Net to emerging generative manipulation scenarios.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62363032), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No.2023D01C20), the National Science and Technology Major Project (No.2022ZD0115803).

References

[1] F. Z. Mehrjardi, A. M. Latif, M. S. Zarchi, et al., "A survey on deep learning-based image forgery detection," *Pattern Recognit.*, vol. 138, p. 109778, 2023.

[2] C. Qu, C. Liu, Y. Liu, et al., "Towards robust tampered text detection in document images: New dataset and new solution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5937–5946.

[3] S. S. Gornale, G. Patil, and R. Benne, "Document image forgery detection using RGB color channel," *Trans. Eng. Comput. Sci.*, vol. 10, no. 5, pp. 1–14, 2022.

[4] S. Li, W. Ma, J. Guo, et al., "UnionFormer: Unified-learning transformer with multi-view representation for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 12523–12533.

[5] C. Shi, C. Wang, X. Zhou, et al., "DAE-Net: Dual attention mechanism and edge supervision network for image manipulation detection and localization," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.

[6] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9535–9544.

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

[8] X. Ma, B. Du, Z. Jiang, et al., "IML-ViT: Benchmarking image manipulation localization by vision transformer," arXiv preprint arXiv:2307.14863, 2023.

[9] C. Dong, X. Chen, R. Hu, et al., "MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3539–3553, Mar. 2023.

[10] F. Guillaro, D. Cozzolino, A. Sud, et al., "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 20606–20615.

[11] X. Lin, S. Wang, J. Deng, et al., "Image manipulation detection by multiple tampering traces and edge artifact enhancement," *Pattern Recognit.*, vol. 133, p. 109026, Jan. 2023.

[12] C. Shi, C. Wang, X. Zhou, et al., "DAE-Net: Dual attention mechanism and edge supervision network for image manipulation detection and localization," *IEEE Trans. Instrum. Meas.*, 2024.

[13] N. Bai, X. Wang, R. Han, et al., "PIM-Net: Progressive inconsistency mining network for image manipulation localization," *Pattern Recognit.*, vol. 159, p. 111136, 2025.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[16] J. Fridrich, "Digital image forensics using sensor noise," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 26–37, Mar. 2009.

[17] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Beijing, China, Jul. 2013, pp. 422–426.

[18] H. Guan, et al., "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, Jan. 2019, pp. 63–72.

[19] Y.-F. Hsu and S.-F. Chang, "Detecting image splicing using geometry invariants and camera characteristics consistency," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Toronto, ON, Canada, Jul. 2006, pp. 549–552.

[20] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "Coverage—A novel database for copy-move forgery detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 161–165.

[21] C. Kong, A. Luo, S. Wang, H. Li, A. Rocha, and A. C. Kot, "Pixel-inconsistency modeling for image manipulation localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 6, pp. 4455–4472, Jun. 2025.

[22] D. Li, et al., "Learnable frequency decomposition for image forgery detection and localization," in Proc. Int. Joint Conf. Artif. Intell (IJCAI),2025.