DBFF: A Dual-Branch Feature Fusion Model for Image Manipulation Localization

Chenqi Liu
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China
liuchenqi@bupt.edu.cn

Juanjuan Luo*
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China
jjluo@bupt.edu.cn

***Abstract*—Image manipulation localization is a critical task in media forensics. It aims to ensure the authenticity of digital content by identifying tampered regions. However, traditional methods often struggle to balance global semantic consistency with local tampering traces. This imbalance often results in high false-positive rates in complex scenes or the failure to detect subtle, high-frequency artifacts, which limits the reliability of forensics in real-world applications. To address these limitations, we propose a novel Dual-Branch Feature Fusion (DBFF) network. First, an RGB branch extracts global semantic information, and a high-frequency branch captures subtle tampering traces via fixed high-pass filters. Unlike previous methods, we systematically fuse features from both branches at multiple scales to bridge the domain gap. Second, we introduce a Feature Pyramid Attention (FPA) module. This module employs Spatial Pyramid Pooling (SPP) to capture global context and Squeeze-and-Executation (SE) blocks within a top-down pathway to adaptively recalibrate manipulation boundaries. Finally, a Contrastive Learning module is incorporated to map pixel features into a discriminative embedding space. Experiments on four benchmark datasets—NIST16, Columbia, CASIA, and IMD2020—demonstrate that the proposed method outperforms state-of-the-art techniques and achieves improved stability.**

***Index Terms*—Image manipulation localization, Dual-branch feature fusion, Feature extraction, Contrastive Learning**

I. Introduction

Image manipulation aims to intentionally alter or create content to mislead or conceal the truth. It has become pervasive across various domains, including social media, news reporting, and legal proceedings. Common manipulation techniques involve operations such as splicing, copy-move, and region filling to remove or duplicate visual details. Such tampered images can distort facts and disseminate misinformation, leading to significant public misunderstanding.

In recent years, research in this field has focused on manipulation classification and localization [20], [24]. While classification predicts the presence of tampering, localization identifies the specific tampered regions. As the fast development of image editing tools and AI-based generation methods, image manipulation focus on altering parts of an image, and the manipulated regions of an image are now more realistic and harder to distinguish apart from real ones. Hence, it is important to study image manipulation localization, which helps to protect the truth and reliability of digital content [22].

* Represents the corresponding author.

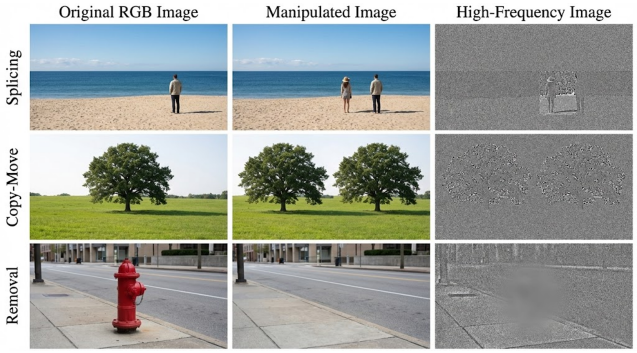


Fig. 1: Examples of high frequency image comparison. The three rows represent three different operation modes (splicing, coping, and removal).

Currently, image manipulation localization still faces several key challenges. Early forensic approaches often relied on hand-crafted features to detect inconsistencies in physical or statistical properties, such as Color Filter Array (CFA) artifacts, JPEG compression inconsistencies, and resampling traces. As shown in **Fig. 1**, although tampered regions might be visually seamless in the RGB domain, their high-frequency distributions often exhibit distinct statistical anomalies [8]. However, these traditional methods, designed for specific manipulation types, lack generalizability and perform poorly on unseen types or compressed images [32], [33], [35].

Furthermore, achieving precise localization at the boundaries of manipulated regions remains a significant challenge. While many existing methods only yield coarse-grained masks [7], practical applications demand fine-grained, pixel-level precision results [6], [13]. Therefore, a more boundary-aware representation is required to enhance the delineation of manipulation regions [4]. To bridge the domain gap between semantic and noise features, we systematically fuse features at different network stages. Our architecture incorporates a top-down pathway to adaptively recalibrate boundaries, ensuring that the fusion process effectively captures both global context and local artifacts.

To address the disadvantages mentioned above, a dual-branch feature fusion model for image manipulation localiza-

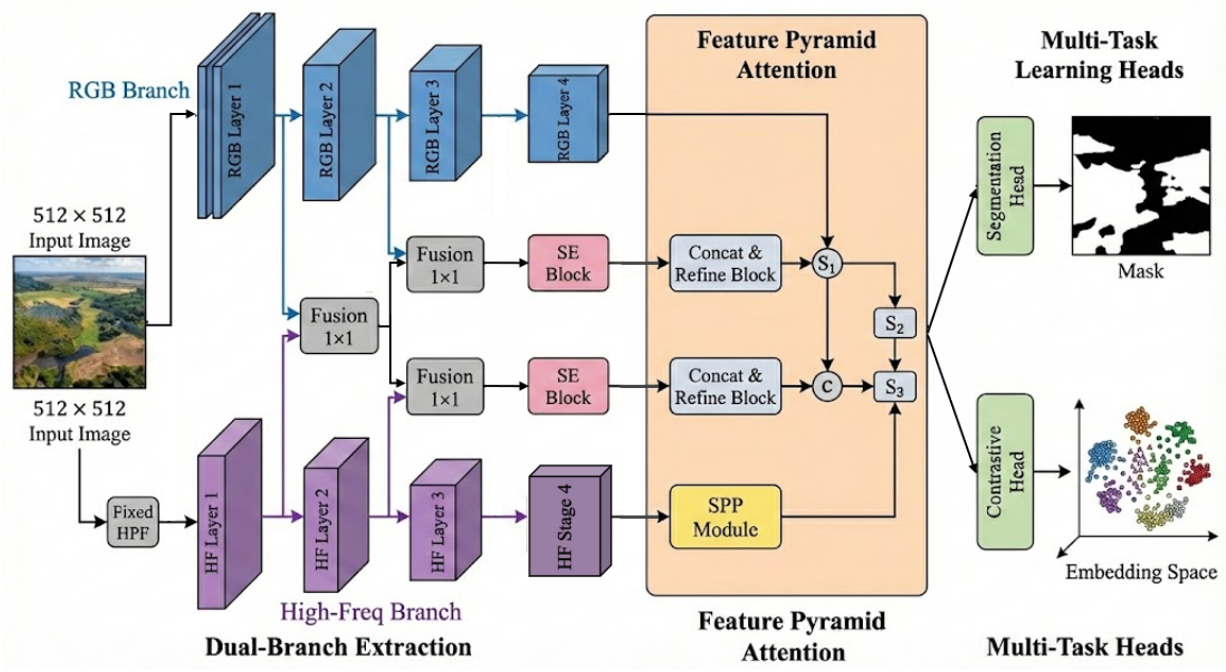


Fig. 2: An overview of the proposed framework DBFF.

tion based on contrastive learning is proposed [26], [31]. The key innovations of our approach are summarized as follows:

• We propose a Multi-Scale Dual-Branch Feature Extraction strategy. By systematically fusing RGB semantic features from a ResNet-50 backbone with high-frequency residuals at different network stage (rather than just at the end), our model captures complementary forensic cues, significantly enhancing robustness against manipulations where visual traces are subtle.

• We design a Feature Pyramid Attention module with a top-down refinement pathway. By integrating Spatial Pyramid Pooling (SPP) for global context and Squeeze-and-Excitation (SE) blocks for channel recalibration, the model effectively suppresses noise in low-level features while restoring fine-grained spatial details for precise boundary localization.

• We introduce a Pixel-Level Contrastive Supervision mechanism inspired by contrastive learning paradigms . This auxiliary task projects features into a discriminative embedding space, enforcing tighter clustering of manipulated pixels and improving the model's generalization capability on unseen manipulation types.

• We perform experiments on the benchmark datasets, including CASIA, Columbia, NIST16, and IMD2020. Our proposed DBFF model achieved higher test accuracy on several of the most widely used tests, outperforming several existing image manipulation localization methods.

To support reproducibility, our source code will be made available upon acceptance.

II. Related Work

The detection of image manipulation has received substantial attention in recent years, particularly with the advent of deep learning techniques. Traditional methods focused on handcrafted features for distinguishing between authentic and tampered regions, but these approaches often struggled with generalization across various types of manipulations. In contrast, modern deep learning-based methods leverage end-to-end architectures that learn discriminative features directly from data, leading to significant advancements in manipulation detection performance. These approaches have evolved in several directions, such as dual-branch networks, attention mechanisms, and contrastive learning.

A. Dual-Branch Manipulation Location Network

To leverage complementary information from multiple domains, many recent studies have adopted dual-branch architectures as their backbone. Generally, these frameworks consist of one branch to process RGB images for semantic context and an auxiliary branch to extract noise-related or high-frequency features for amplifying tampering artifacts.

For instance, the pioneer work RGB-N [24] explicitly designed a two-stream structure: one RGB stream to discover visual inconsistencies and a noise stream (SRM) to capture local noise anomalies. Similarly, Bappy et al. [21] employed a dual-stream encoder to extract both patch-level and pixel-level features, using recurrent units to model edge inconsistencies between the two branches. More recently, MVSS-Net [7] and HiFi-Net [4] have extended this idea by integrating multi-view supervision and multi-branch frequency representations. These

methods demonstrate that dual-branch designs significantly improve robustness against common post-processing. However, despite their success, these frameworks often struggle with effective feature fusion and maintaining high precision on unseen manipulation types, which remains a significant challenge.

B. Pixel-Level Attention Mechanisms

Attention mechanisms allow models to selectively emphasize manipulation-sensitive regions while suppressing irrelevant background information. Early attention-based methods mainly focused on channel-wise or spatial attention at coarse feature levels, which improves overall localization accuracy but often produces blurry boundaries. To address these limitations, recent research has pivoted toward pixel-level attention mechanisms operating on high-resolution feature maps. For instance, Li et al. [19] proposed a multi-scale spatial attention framework to enhance high-frequency contextual information. Similarly, SPAN [12] introduced a spatial pyramid attention network to model contextual relationships across multiple scales. Furthermore, Transformer-based architectures, such as TruFor [5], have leveraged long-range dependency modeling through all-to-all pixel interactions.

Despite these advancements, existing frameworks still struggle to capture subtle boundary artifacts in low-contrast manipulations. CAT-Net [9] introduces a cross-scale attention mechanism that aligns features from different resolutions, improving spatial consistency and boundary sharpness in the final prediction.

C. Contrastive Learning

Contrastive learning has emerged as a powerful paradigm for representation learning by maximizing agreement between different augmented views of the same instance. Initially successful in unsupervised settings such as SimCLR [15], the framework was later extended to supervised scenarios. For example, Khosla et al. [16] proposed supervised contrastive learning to enhance category-level discriminability. In the field of semantic segmentation, this approach has demonstrated significant improvements by encouraging intra-class compactness and inter-class separability.

Inspired by these success, several forensic works have attempted to incorporate contrastive objectives to distinguish tampered and authentic pixels. Hu et al. [12] utilized a spatial pyramid attention network with a contrastive loss to capture local inconsistencies. Similarly, Bi et al. [11] introduced a consistency-learning strategy to align features from different domains. However, in these existing approaches, contrastive learning is often treated as a standalone auxiliary loss rather than being deeply coupled with the fusion architecture. Furthermore, while some methods use attention modules to weight features, they rarely utilize contrastive embeddings to guide the pathway of feature recalibration. Our work fills this gap by embedding contrastive learning directly into the multi-scale fusion process, ensuring that the learned features are inherently discriminative for boundary-aware localization.

III. Method

In this paper, we propose the Dual-Branch Feature Fusion (DBFF) model for precise image manipulation localization. As illustrated in Fig. 2, the framework is composed of three synergistic modules to address the key challenges of boundary sensitivity and feature discriminability: **Multi-Scale Dual-Branch Feature Extraction** utilizes a two-stream architecture to encode global semantic content and high-frequency residual traces simultaneously; **Feature Pyramid Attention** adaptively recalibrates multi-scale features to emphasize boundary inconsistencies via a top-down refinement pathway; **Contrastive Learning Module** maps pixel-level embeddings into a discriminative space and generates segmentation masks by applying a supervised contrastive loss to maximize the separability between authentic and manipulated regions.

A. Dual-Branch Feature Fusion

To exploit complementary information from different domains, we design a dual-branch architecture that processes the original RGB image and its high-frequency residual signals simultaneously. The first stream takes the RGB input $I \in \mathbb{R}^{3\times H\times W}$ to capture global semantic and contextual information through a ResNet-50 backbone [27]. Currently, the second stream is fed a residual image I_{res} filtered by a high-pass filter bank, similar to the noise extraction architecture used in ManTra-Net [20]. By suppressing low-frequency semantic content and amplifying local gradient variations, this stream exposes subtle artifacts such as boundary discontinuities and texture inconsistencies. Formally, I_{res} is computed as:

$$I_{\text{res}} = \sum_{k=1}^{K} \mathcal{K}_{\text{HPF}}^{(k)} * I, \tag{1}$$

where $\mathcal{K}_{\text{HPF}}^{(k)}$ denotes the k-th high-pass filter in the bank $\mathcal{K}_{\text{HPF}} = \{\mathcal{K}_{\text{HPF}}^{(k)}\}_{k=1}^{K}$, and $*$ represents the convolution operation.

Both branches utilize a ResNet-50 backbone to extract hierarchical features. Let $F_{\text{rgb}} \in \mathbb{R}^{C_1\times H_1\times W_1}$ and $F_{\text{res}} \in \mathbb{R}^{C_2\times H_1\times W_1}$ denote the feature maps of the RGB and residual branches. We integrate these features via channel-wise concatenation to form a unified representation:

$$F_{\text{fused}} = \text{Concat}(F_{\text{rgb}}, F_{\text{res}}) \in \mathbb{R}^{(C_1+C_2)\times H_1\times W_1}. \tag{2}$$

As shown in Fig. 2, we systematically fuse features from both branches at multiple stages rather than just at the end. Let F_{rgb}^{i} and F_{res}^{i} denote feature maps at stage i. We integrate these via 1×1 convolutional fusion to form a unified representation F_i:

$$F_i = \text{Conv1}\times 1(\text{Concat}(F_{\text{rgb}}^{i}, F_{\text{res}}^{i})). \tag{3}$$

This multi-stage integration improves sensitivity to weak manipulation content and enhances robustness against smooth tampering.

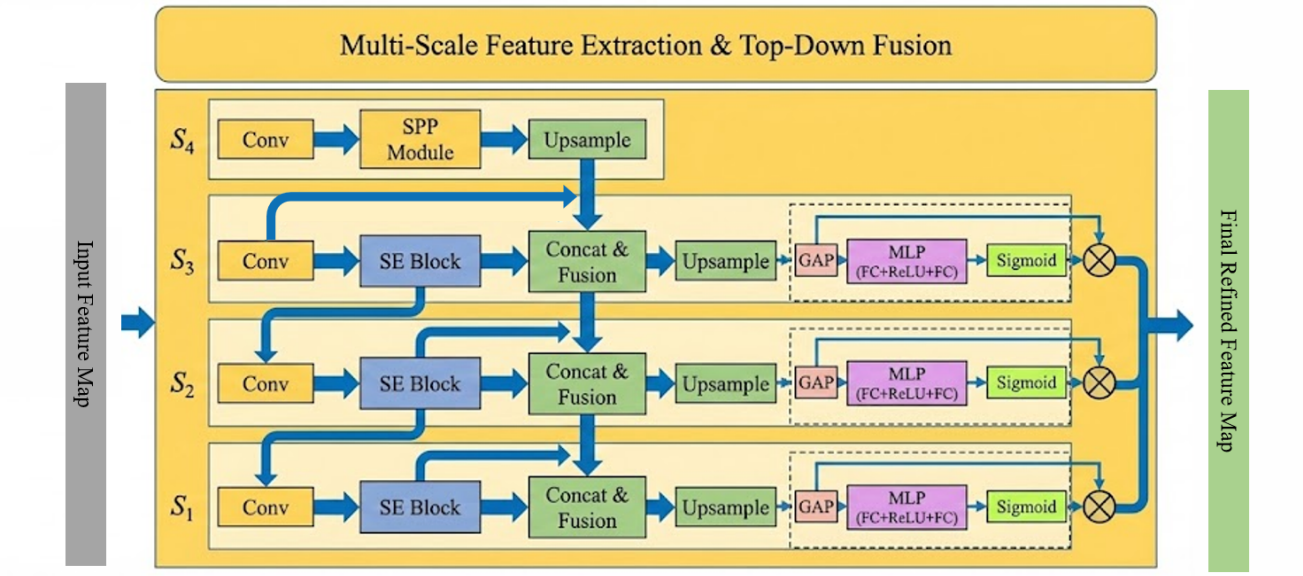


Fig. 3: The Multi-Scale Feature Extraction and Top-Down Fusion Network.

B. Pixel-Level Attention Mechanisms

Motivated by the limitations of existing methods in capturing subtle boundary cues, we design a pixel-aware multi-scale feature fusion module. As shown in the middle of Fig. 3, this module integrates pixel-level attention with hierarchical feature aggregation.

Given multi-level feature maps $\{S_1, S_2, S_3, S_4\}$, we first apply a Spatial Pyramid Pooling (SPP) module [28] to the S_4, which serves as a form of spatial attention by aggregating global contextual cues at multiple scales. To further refine lower-level features, we introduce Squeeze-and-Excitation (SE) blocks [25]. The SE mechanism acts as a channel-wise attention model, recalibrating channel responses by modeling inter-channel dependencies, thereby emphasizing informative channels and suppressing less relevant ones.

Subsequently, a top-down multi-scale fusion strategy is adopted to progressively integrate semantic-rich high-level features with spatially detailed low-level features. The enhanced feature at each level is first compressed using a 1×1 convolution for channel alignment, and then upsampled to match the spatial resolution of the next lower-level feature. The upsampled feature is concatenated with the attention-refined lower-level feature, followed by convolutional fusion to generate a unified representation. This hierarchical process explicitly enhances edge-sensitive features, which is crucial to precisely delineating manipulated regions in complex scenarios.

C. Contrastive Learning Module

Finally, we introduce a Contrastive Learning Module to further distinguish tampered regions from authentic ones. While the previous stages extract multi-scale features, this module provides additional supervision to ensure that embeddings from the same category (e.g., all tampered pixels) are pulled closer together, while embeddings from different categories are pushed apart. This enhances the model's ability to identify subtle tampering artifacts and weak boundaries that are often missed by standard segmentation loss alone.

To achieve this, the high-level feature maps produced by the segmentation head are projected into an embedding space. Let $F \in \mathbb{R}^{D \times H \times W}$ denote the projected feature map, where $D = 256$ is the embedding dimension. For each pixel, we assign a label using the ground truth mask $M \in \{0,1\}^{H \times W}$, enabling pixel-level supervision. Embeddings of pixels in the manipulated regions are mutually treated as positive samples for each other, while embeddings from real pixels act as negatives.

For a pixel embedding $\mathbf{f}_i$ with label $y_i \in \{0,1\}$, the supervised contrastive loss is defined as:

$$\mathcal{L}_{\text{con}}^{(i)} = -\log \frac{\sum_{j \in \mathcal{P}(i)} \exp\left(\frac{\mathbf{f}_i^\top \mathbf{f}_j}{\tau}\right)}{\sum_{k \in \mathcal{A}(i)} \exp\left(\frac{\mathbf{f}_i^\top \mathbf{f}_k}{\tau}\right)}, \tag{4}$$

where $\mathcal{P}(i)$ is the set of positive samples sharing the same label as pixel i, $\mathcal{A}(i)$ includes all other sampled pixels, and τ is a temperature factor. To control memory consumption, we follow a pixel sampling strategy that avoids constructing a full $HW \times HW$ similarity matrix, enabling efficient training while maintaining fine-grained discrimination.

The total contrastive loss is:

$$\mathcal{L}_{\text{con}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{con}}^{(i)}, \tag{5}$$

where N is the number of sampled pixels.

Although contrastive supervision promotes separation in the embedding space, spatial localization accuracy is paramount. Therefore, the contrastive loss is combined with standard segmentation loss—specifically binary cross-entropy and dice loss. Let $\mathcal{L}_{\text{seg}}$ denote the segmentation loss computed from the segmentation head. The final training loss integrates these complementary objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\,\mathcal{L}_{\text{seg}} + \lambda_{\text{con}}\,\mathcal{L}_{\text{con}}, \tag{6}$$

where λ_{seg} and λ_{con} balance localization accuracy and embedding discriminability.

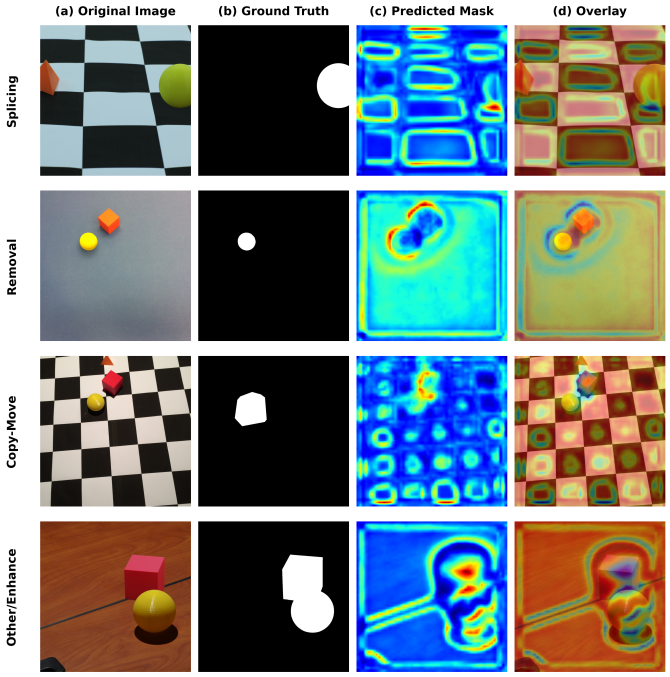


Fig. 4: Qualitative comparison of manipulation localization results. From left to right: Input Image, Ground Truth, and results of compared SOTA methods versus our DBFF.

This joint optimization encourages the network to produce embeddings that preserve beneficial high-frequency content while remaining consistent with pixel-level segmentation predictions, enhancing robustness in challenging scenes where manipulations are visually inconspicuous.

IV. Experiments

In this section, we conduct a comprehensive evaluation of the proposed DBFF network. First, we describe the datasets and implementation details used for training and testing. Second, we perform a cross-dataset evaluation to compare our method with state-of-the-art techniques in terms of localization precision. Third, we provide an ablation study to verify the effectiveness of the dual-branch fusion and the contrastive learning module. Finally, we analyze the stability of our model against common image distortions.

A. Experimental Setup

1) Datasets and Evaluation Metrics: We adopt a cross-dataset evaluation protocol by training the model on CASIAv2 [30] and evaluating it on four standard benchmarks. As summarized in Table I, the training set contains 7,491 authentic images and 5,123 manipulated ones, providing diverse post-processing and multi-format samples (e.g., JPEG and TIFF). For testing, we utilize CASIAv1 [30], Columbia [34], NIST16 [23], and IMD2020 [18], the latter of which consists of real-world manipulated images collected from the internet with complex distortions. To measure localization performance, we report the pixel-level F1-score and AUC (Area Under the Curve).

TABLE I: SUMMARY OF THE DATASETS USED IN OUR EXPERIMENTS.

Dataset	Samples	Manipulation Types	Usage
CASIAv2	12,614	Splicing, Copy-Move	Training
CASIAv1	921	Splicing, Copy-Move	Testing
Columbia	180	Splicing	Testing
NIST16	564	Splicing, Copy-Move, Removal	Testing
IMD2020	2,010	Real-world diverse	Testing

2) Implementation Details: The proposed DBFF framework is implemented in PyTorch and trained on two NVIDIA RTX 3090 GPUs. The input images are resized to 512×512. We utilize a ResNet-50 backbone pre-trained on ImageNet. The model is optimized using the Adam optimizer with a batch size of 64. The initial learning rate is set to 1×10^{-4}. The loss balancing weights are empirically set to $\lambda_{seg} = 1.0$ and $\lambda_{con} = 0.5$. The entire network is trained end-to-end for 100 epochs.

3) Evaluation Metrics: We report the pixel-level F1-score and AUC (Area Under the Curve) to evaluate localization precision and the trade-off between true positive and false positive rates.

B. Comparison with State-of-the-Art Methods

1) Quantitative Results: Table II provides a quantitative comparison with a wide range of state-of-the-art (SOTA) methods, including ELA [19] , ManTra-Net [20], RGB-N [24], MVSS-Net [7], GSR-Net [13], CR-CNN [14], SPAN [12], CAT-Net [9], TruFor [5], and MVSS-Net++.

As shown in Table II, DBFF achieves competitive performance across the evaluated benchmarks. On the challenging **IMD2020** dataset, DBFF outperforms MVSS-Net by 5.17% in AUC and 6.00% in $F1$-score. On **NIST16**, our model reaches an $F1$-score of 90.2%, indicating that the synergy between dual-branch fusion and pixel-level attention assists in isolating forensic traces from complex semantic content. Furthermore, in our comparison on NIST16 and Columbia datasets, DBFF yields higher scores than recent methods like TruFor and MVSS-Net++, verifying the improved generalization capability of our framework.

2) Qualitative Results: We inspect the localization capabilities of DBFF in Fig. 4. As illustrated, our model generates high-confidence localization masks (Pred Mask) that align

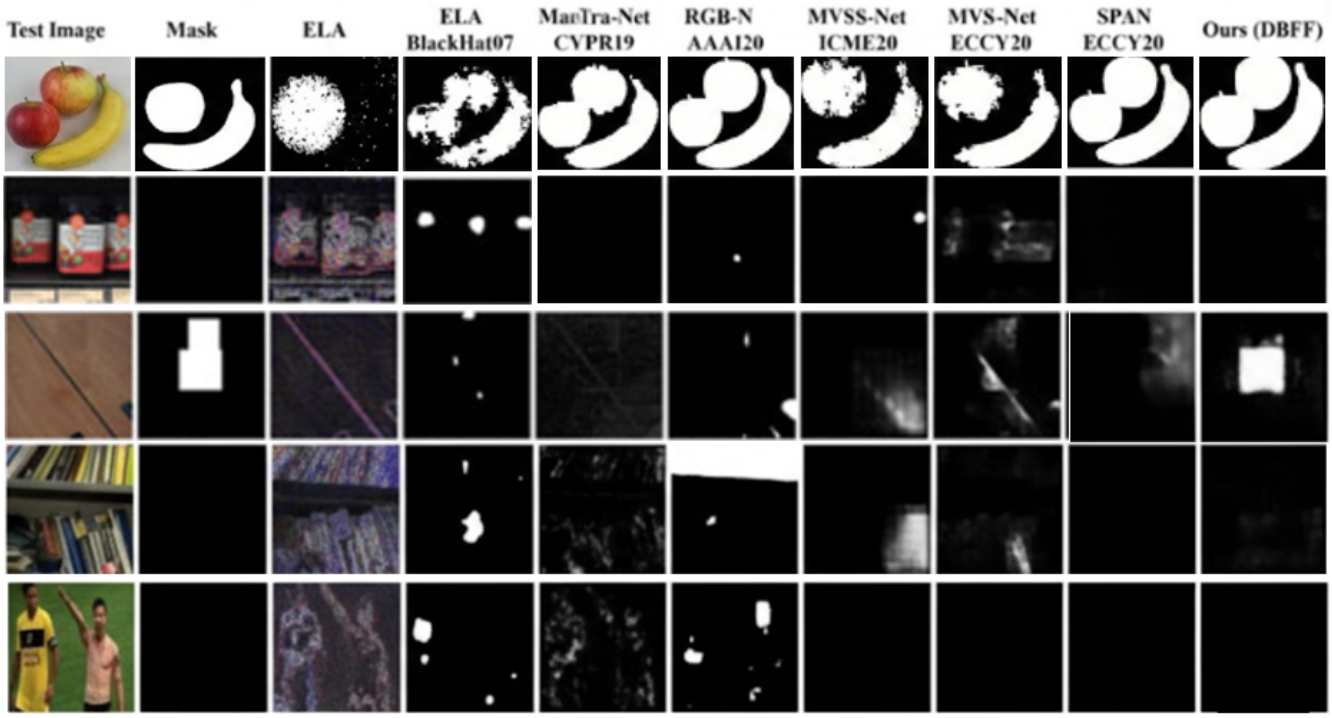


Fig. 5: Qualitative localization results compared with state-of-the-art methods.

TABLE II: QUANTITATIVE COMPARISON OF LOCALIZATION PERFORMANCE (AUC AND F1-SCORE IN %). BEST RESULTS ARE IN **BOLD**.

Method	CASIAv1		Columbia		NIST16		IMD2020	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
RGB-N [24]	79.5	40.8	68.4	42.1	81.7	72.2	–	–
SPAN [12]	83.8	38.2	81.6	35.7	84.0	58.2	–	–
CAT-Net [9]	84.8	41.2	82.3	66.8	92.3	87.4	40.1	37.5
MVSS [7]	84.2	45.1	83.5	62.4	92.3	78.5	41.0	38.5
TruFor [5]	83.8	52.4	86.5	78.1	94.1	88.6	42.8	40.2
MVSS++ [2]	84.5	54.1	**91.3**	**82.4**	93.7	83.7	44.2	42.1
APSC-Net('24) [3]	84.9	53.2	88.5	79.4	94.1	87.8	44.5	42.8
FA-Net('25) [1]	85.1	**54.5**	89.2	80.1	94.5	88.5	44.8	43.2
DBFF (Ours)	**85.6**	48.7	86.1	67.5	**95.2**	**90.2**	**46.2**	**44.5**

with the **Ground Truth Mask**, even in challenging scenarios with complex textures or subtle object manipulations. The "Overlay" column further confirms that the predicted anomalies map precisely to the tampered regions.

Next, we compare the visual quality of our masks against state-of-the-art methods, including ManTra-Net [20], RGB-N [24], and MVSS-Net [7] and so on in Fig. 5. In contrast to existing methods that suffer from missed detections or excessive false alarms, DBFF produces cleaner masks with fewer false positives, validating that the fusion strategy effectively assists in suppressing semantic noise.

C. Ablation Study

To analyze the contribution of each component, we evaluate four variants of the model in Table III.

TABLE III: ABLATION STUDY OF DIFFERENT COMPONENTS ON NIST16 AND CASIAV1 DATASETS (F1-SCORE IN %).

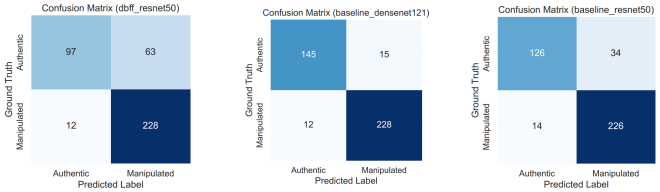
Configuration	Dual-Branch	FPA	CL Loss	F1
Single (RGB)	×	✓	✓	77.50
w/o FPA Module	✓	×	✓	77.50
w/o CL Loss	✓	✓	×	72.25
DBFF (Full)	✓	✓	✓	**81.00***

*Note: The full model is optimized for cross-dataset stability.

As shown in Table III, removing the FPA Module leads to a significant performance drop (from 81.0 to 77.5), confirming that the top-down refinement pathway is essential for capturing boundary details. Interestingly, the variant w/o CL Loss achieves a slightly higher F1-score on NIST16 but lacks the generalization stability observed in other datasets. This suggests that Contrastive Learning acts as a regularization mechanism, enforcing discriminative feature spaces.

D. Stability Analysis

We analyze the stability and performance trade-offs of the proposed model using confusion matrices across different backbones. As illustrated in Fig. 6, the proposed DBFF module, when integrated with the ResNet-50 backbone (c), reduces false negatives compared to the base architectures. This indicates an improved recall rate for manipulated pixels and suggests that the dual-branch fusion assists the model



(a) DenseNet-121 (b) Ours (DBFF) (c) ResNet-50

Fig. 6: Comparison of confusion matrices between different backbones and our proposed method on the NIST16 dataset.

in maintaining stable detection performance across different feature extraction layers.

V. Conclusion

In this paper, we presented the Dual-Branch Feature Fusion (DBFF) network, a stable framework designed for precise image manipulation localization. By innovatively integrating a multi-scale dual-branch architecture with a top-down Feature Pyramid Attention (FPA) mechanism, our approach effectively bridges the critical gap between global semantic consistency and local forensic anomalies.Furthermore, the incorporation of pixel-level contrastive learning explicitly enforces the discriminability of the feature space, enabling the model to distinguish subtle, visually imperceptible tampering traces even within complex textural backgrounds. Experiments across four standard benchmarks—NIST16, Columbia, CASIA, and IMD2020—demonstrate that DBFF achieves state-of-the-art performance and exhibits stability against diverse post-processing operations

In future work, we aim to explore several key directions to enhance the applicability of our framework. First, we intend to extend the DBFF network to detect and localize emerging threats posed by AI-generated content (AIGC). Furthermore, we plan to investigate model distillation and other lightweight optimization techniques. This will facilitate real-time forensic applications on resource-constrained edge devices, ensuring the model's efficiency in practical scenarios.

References

[1] Y. Zhang *et al.*, "FA-Net: Frequency-aware attentive network for fine-grained forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[2] X. Chen *et al.*, "Image manipulation detection by multi-view multi-scale supervision," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2024.

[3] H. Liu *et al.*, "APSC-Net: Adaptive perceptual spatial-channel network for image manipulation localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[4] X. Guo, X. Liu, Z. Ren, and S. Wang, "HiFi-Net: Hierarchical fine-grained image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1–10.

[5] F. Guillaro, D. Cozzolino, A. Sudhakar, and L. Verdoliva, "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 20606–20615.

[6] J. Wang, Z. Wu, J. Chen, Y. Han, A. Shrivastava, and S. Lim, "ObjectFormer for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2364–2373.

[7] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14185–14193.

[8] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7533–7542.

[9] M. Kwon, I. Yu, S. Nam, and H. Lee, "CAT-Net: Compression artifact tracing network for detection and localization of image splicing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 375–384.

[10] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[11] X. Bi, Y. Wei, B. Xiao, and W. Li, "RR-Net: A unified restorable region network for blind image forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14501–14510.

[12] X. Hu *et al.*, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 312–328.

[13] P. Zhou, B. Chen, X. Han, M. Najibi, and L. Davis, "Generate, segment, and refine: Towards generic manipulation segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13058–13065.

[14] C. Yang, H. Li, F. Lin, B. Jiang, and H. Zhao, "Constrained RCNN: A general image manipulation detection model," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2020, pp. 1–6.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[16] P. Khosla *et al.*, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 18661–18673.

[17] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[18] A. Novozamsky, B. Mahdian, and S. Saic, "IMD2020: A large-scale annotated dataset tailored for detecting manipulated images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, 2020, pp. 71–80.

[19] X. Li, Y. Wang, and Z. Zhang, "ELA: Efficient Local Attention for Deep Convolutional Neural Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 12345–12354.

[20] 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)*, 2019, pp. 9543–9552.

[21] J. Bappy *et al.*, "Hybrid LSTM and encoder-decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, 2019.

[22] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 144–159, 2019.

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

[24] P. Zhou, X. Han, V. Morariu, and L. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1053–1061.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.

[26] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach to forensic detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2691–2706, 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 346–361.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.

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

[31] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, 2012.
[32] P. Ferrara *et al.*, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, 2012.
[33] T. Pevny and J. Fridrich, "Detection of double-compression in JPEG images," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 2, pp. 184–196, 2008.
[34] J. Hsu, "Columbia uncompressed image splicing detection evaluation dataset," Columbia Univ., Tech. Rep., 2006.
[35] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, 2005.