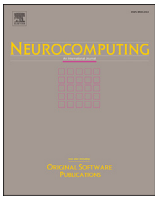
Neurocomputing 663 (2026) 131915

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Do not overestimate RGB: Improving image manipulation detection and localization via multi-noise-view fusion

Joonkyo Shim, Hyunsoo Yoon*

Department of Industrial Engineering, Yonsei University, 50 Yonsei-ro, Seoul, 03722, Seodaemun-gu, South Korea

HIGHLIGHTS

- MNVFusion balances RGB and noise-view modalities for accurate image manipulation localization.
- MB-CMM fuses multi-modal features efficiently with simple MLP-based channel mixing.
- Fixed GeM boosts detection via training-free, fixed operations on predicted maps.
- MNVFusion achieves state-of-the-art performance across six benchmark datasets in both localization and detection tasks.

ARTICLE INFO

Communicated by R. Cong

Keywords:
Image manipulation detection and localization
Image segmentation
Modality fusion
Noise-view generation

ABSTRACT

Image Manipulation Detection and Localization (IMDL) aims to identify tampered images and their altered regions. Existing RGB-centered approaches often overemphasize RGB information while overlooking complementary insights from noise-view modalities. This reliance on RGB limits their ability to detect subtle manipulation traces. To overcome these challenges, we propose Multi-Noise-View Fusion (MNVFusion), a framework that balances the contributions of RGB and noise-view modalities using a multi-branch encoder structure. MNVFusion incorporates the Multi-Branch Channel Mixing Module (MB-CMM), enabling efficient channel-wise fusion to integrate diverse modality features. Additionally, we introduce Fixed GeM, a training-free image-level detection module that enhances overall efficiency through fixed operations on localization maps. Experiments on six benchmark datasets show that MNVFusion delivers state-of-the-art performance in both detection and localization tasks.

1. Introduction

Image manipulation detection and localization (IMDL) is a task that detects whether images have been manipulated and localizes the tampered regions. While traditional image manipulation required technical expertise with editing tools, the recent emergence of deep generative models [1] has made it easier for users to realistically modify images. This surge in accessibility has increased the risk of manipulated images being used maliciously to mislead the public or create societal confusion. To counter this growing threat, the development of robust and effective IMDL frameworks has become crucial.

Unlike general-purpose semantic segmentation that specifies the semantic context of visual objects, IMDL aims to find evidence of manipulation distributed across the edited regions. This can be highlighted by transforming an RGB image into noise-views using high-pass filters or noise-sensitive fingerprints [2–4], which are designed to capture low-level feature inconsistencies. Since these methods target different aspects of an image and provide complementary information, an effective fusion strategy is essential for combining these visual evidences to accurately detect tampered regions. To address the challenges, previous works have introduced dual-branch frameworks for multi-modal feature fusion, incorporating additional noise-view modalities to supplement the RGB input. This approach, illustrated in Fig. 1(a), can be viewed as RGB-Centered Fusion, since RGB is considered the main modality and processed through its dedicated branch. The framework typically consists of an RGB branch and an X branch, where X represents either a single additional modality (dual fusion and late fusion) or a combination of features from multiple modalities (early fusion).

* Corresponding author.
Email addresses: shimjk@yonsei.ac.kr (J. Shim), hs.yoon@yonsei.ac.kr (H. Yoon).

https://doi.org/10.1016/j.neucom.2025.131915
Received 15 August 2025; Received in revised form 28 September 2025; Accepted 22 October 2025
Available online 7 November 2025

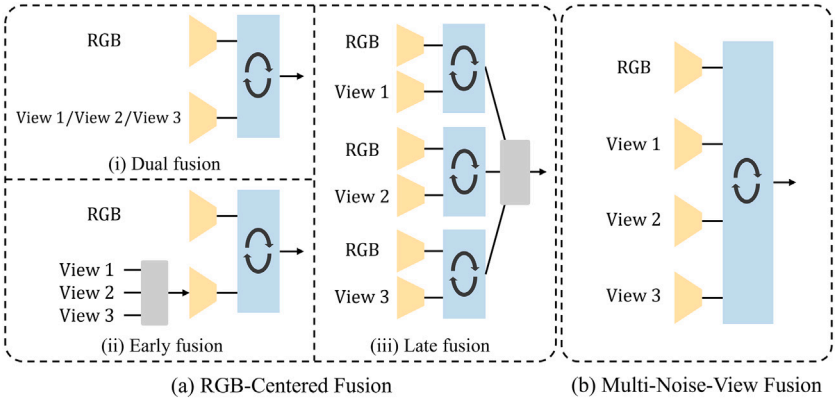


Fig. 1. Comparison between (a) a conventional RGB-centered fusion strategy, where RGB features dominate the integration process, and (b) the proposed Multi-Noise-View Fusion, which treats all modalities equally to exploit complementary information without bias toward a specific modality.

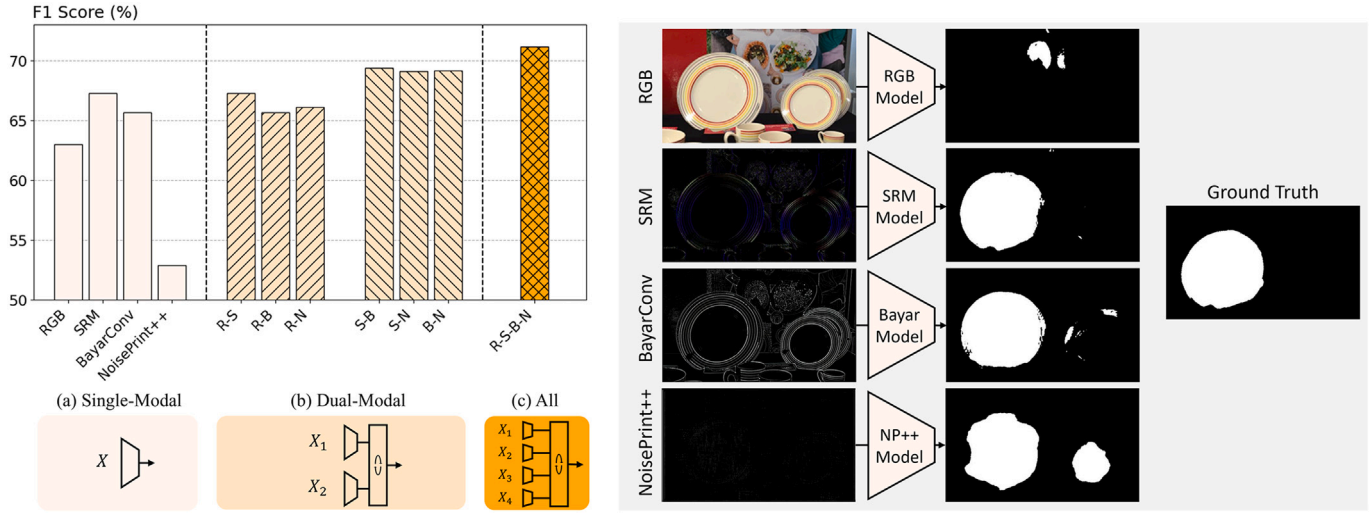


Fig. 2. Localization performance comparison on different combination of input modalities for training the model, including RGB and its three noise-views: Steganalysis Rich Model (SRM), Bayar Convolution, and NoisePrint++. (a) Single input modality without fusion. (b) Dual input modalities with dual fusion (e.g., R–S denotes fusion of RGB and SRM). (c) All input modalities with MNVFusion. On the right, we present a visual example of each modality and the corresponding predicted map in (a) Single-Modal scenario.

However, we observe that *RGB is not the most effective modality for the IMDL task*. As shown in Fig. 2, we vary the input modalities used for training and compare their pixel-level F1 scores averaged across six benchmark datasets, using the SegFormer [5] backbone. Notably, in the (a) Single-Modal scenario, RGB reports a lower performance compared to Steganalysis Rich Model (SRM) and Bayar Convolution, indicating that RGB is not the dominant modality for providing critical evidence of manipulation. This observation is further supported by a visual example in Fig. 2, where the manipulated area is clearly highlighted in the noise-view modalities but remains indistinct in the RGB image. Consequently, training with only RGB images fails to produce the desired localization map. Additionally, the performance gap between (a) Single-Modal and (b) Dual-Modal indicates that synergy between modalities is more significant in noise-noise fusion (S-B, S-N, and B-N) than in RGB-X fusion (R-S, R-B, and R-N). These findings underscore two major drawbacks of RGB-Centered Fusion: (1) the excessive influence of RGB on the overall framework, despite its limitations as a modality, and (2) its inability to fully leverage the complementary information from multiple modalities, as feature fusion between noise-view modalities is not achieved.

To address these issues, we propose Multi-Noise-View Fusion (MNVFusion), illustrated in Fig. 1(b), a novel multi-branch framework for the IMDL task. MNVFusion balances the influence of modalities by processing each modality through its own dedicated branch while enhancing synergy by combining features from all modalities. The performance improvement in Fig. 2(c) demonstrates that MNVFusion facilitates the full potential of multiple modalities. The core of MNVFusion is the Multi-Branch Channel Mixing Module (MB-CMM), which interactively fuses intermediate features via simple channel-wise feature mixing. Unlike previous fusion strategies limited to dual-modality fusion, MB-CMM supports feature fusion among more than two modalities, offering greater flexibility. Consequently, MNVFusion serves as a general IMDL framework capable of handling a wide range of modalities. Additionally, we propose Fixed GeM, a training-free module for image-level detection. Fixed GeM reduces overall training costs by relying on fixed operations while maintaining robust performance.

We verify the effectiveness of MNVFusion by leveraging various visual networks as backbones, including ConvNeXt [6], SegFormer [5], and VMamba [7]. When combined with VMamba, MNVFusion achieves state-of-the-art performance in both localization and detection tasks on an average of six benchmark datasets.

Our contributions are summarized as follows:

- We introduce MNVFusion, a new framework for detecting manipulated images and localizing tampered areas in such images.

MNVFusion processes multiple inputs separately and promotes full synergy between them.
- We propose MB-CMM for multi-modal feature fusion. MB-CMM performs simple MLP operations and interactively fuses features obtained from the multi-branch encoder.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance in both localization and detection tasks.

2. Related works

2.1. Image manipulation detection and localization

Previous works aiming to develop a general IMDL model [8–17] have incorporated additional modalities by generating noise-views of images through high-pass filters like Steganalysis Rich Model (SRM) [2] and BayarConv [3] or transforming RGB into frequency domain using Discrete Cosine Transform (DCT) [18]. TruFor [4] enhances NoisePrint [19], a CNN-based noise extractor, and proposes NoisePrint++, a learnable noise-sensitive fingerprint. Combined with a Transformer-based backbone [5], TruFor achieves superior performance compared to previous methods. Triaridis and Mezaris [20] extend TruFor to encompass three additional noise-view modalities and propose Early Fusion and Late Fusion, two different methods to fuse the intermediate features of multiple modalities. UnionFormer [21] proposes a parallel CNN-Transformer architecture to explore interactions between local and global features. While previous methods develop RGB-centered dual-branch frameworks, we introduce a multi-branch structure to fully utilize the complementary information of diverse modalities.

2.2. Multi-modal feature fusion

Various methods for fusing multi-modal features have been proposed for the IMDL task, especially for homogeneous modalities that are aligned with pixel-level. A naive concatenation of input modalities or intermediate features between different branches has been introduced in ManTraNet [9] and CAT-Net v2 [11]. MVSS-Net [12] explores dual-attention to enhance concatenated features of dual branches. Recently, interactive fusion that allows bidirectional cross-modal feature rectification has shown great performance in the multi-modal semantic segmentation field. Specifically, CMX [22] introduces the Cross-Modal Feature Rectification Module (CM-FRM), which adjusts one feature based on another, and the Feature Fusion Module (FFM), which combines the two features into a single map. The current state-of-the-art IMDL methods [4,20] adopt CMX fusion to facilitate effective fusion between RGB and noise-view modalities. However, these methods are limited to dual-modality fusion because of the explicit feature exchange mechanism between two encoder branches. In this paper, we propose a novel interactive fusion method that is capable of handling an arbitrary number of modalities.

3. Methods

3.1. Framework overview

Unlike RGB-Centered Fusion frameworks that over-rely on RGB inputs, MNVFusion assigns equal importance to all modalities through dedicated branches, as illustrated in Fig. 3. This design ensures a balanced contribution and maximized synergy between RGB and noise views. Furthermore, MNVFusion integrates multiple modalities for feature fusion using a multi-branch channel mixing module (MB-CMM). In the following sections, we provide a detailed explanation of our framework.

3.2. Modality extraction

To construct multi-modal inputs, we employ three noise-view extractors: SRM, BayarConv, and NoisePrint++, each capturing unique manipulation traces that complement RGB data. SRM [2] applies predefined high-pass filters to extract spatial inconsistencies, and we select three commonly used filters from prior studies [8,9] to ensure effectiveness. BayarConv [3] leverages a constrained convolutional layer to detect structural irregularities. NoisePrint++ [4] focuses on subtle editing traces by training a DnCNN [23] extractor in a self-supervised contrastive manner. We utilize pretrained weights of BayarConv and NoisePrint++ provided by Early Fusion and Late Fusion [20]. All three extracted modalities are pixel-aligned and spatially consistent, enabling integration without the need for further spatial registration in the subsequent fusion stage.

3.3. Multi-branch visual encoder

The MNVFusion framework employs a multi-branch encoder with four parallel branches, each processing a specific modality (RGB, SRM, BayarConv, NoisePrint++). To maintain parameter efficiency and reduce computational complexity, all branches share weights. This weight-sharing strategy minimizes redundancy and enables efficient multi-modal integration.

For the encoder structure, we explore various backbone networks, including ConvNeXt [6], SegFormer [5], and VMamba [7], leveraging their unique strengths. ConvNeXt improves upon traditional CNNs like ResNet [24] by optimizing design choices for enhanced efficiency. SegFormer employs a lightweight Transformer-based architecture, excelling in capturing global context for semantic segmentation tasks. VMamba, inspired by State Space Models (SSMs) [25], achieves a global receptive field with linear complexity using the 2D-Selective-Scan (SS2D) module, making it well-suited for dense prediction tasks. These backbones share a common hierarchical structure with four stages, seamlessly integrating with our multi-branch encoder.

This architecture supports diverse backbones and adapts flexibly to various modalities in a model-agnostic manner, ensuring robust and

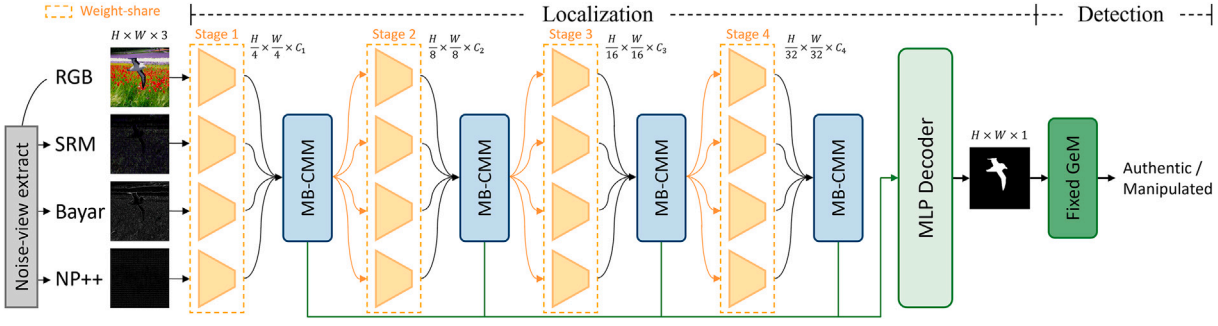


Fig. 3. An overview of the proposed MNVFusion. The model is trained for the localization task in a segmentation-based manner, where the input consists of four modalities and the output is a predicted map. Then, the model is evaluated for both localization and detection tasks utilizing a fixed operation for the image-level detection.

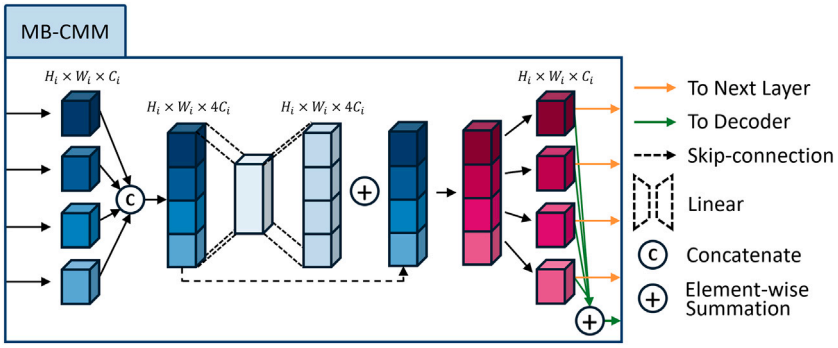


Fig. 4. A detailed view of the proposed MB-CMM.

scalable performance in image manipulation detection and localization tasks.

3.4. Multi-branch channel mixing module

To effectively refine the complementary features of multi-branch encoders, we propose the Multi-Branch Channel Mixing Module (MB-CMM) illustrated in Fig. 4, a simple feature fusion module designed for homogeneous modality fusion. Since each modality inherently encodes rich and spatially aligned information, we focus on facilitating effective information exchange across modalities rather than explicitly modeling their complex inter-modal relationships. In MB-CMM, each feature map from one branch interacts with feature maps from all other branches through channel mixing, leading to the full synergistic effect among input modalities. Furthermore, MB-CMM allows the fused features to influence subsequent layers in other branches, emphasizing features that contain critical evidence while encouraging other branches to focus on previously overlooked areas.

Given four feature maps extracted from each stage $\{F_{RGB}, F_{SRM}, F_{Bayar}, F_{NP++}\} \in \mathbb{R}^{H \times W \times C}$, MB-CMM can be formulated as follows:

$$F_{cat} = \text{Concat}(F_{RGB}, F_{SRM}, F_{Bayar}, F_{NP++}), \tag{1}$$

$$F_{mix} = \text{Linear}(\text{ReLU}(\text{Linear}(F_{cat}))), \tag{2}$$

$$F_{out} = \text{Split}(F_{cat} + F_{mix}). \tag{3}$$

First, the feature maps are concatenated along the channel dimension to form an aggregated feature map $F_{cat} \in \mathbb{R}^{H \times W \times 4C}$. Feature fusion across modalities is then performed through channel mixing by MLP layers. Specifically, the first linear layer reduces the channel dimension ($4C \rightarrow 2C$), followed by a ReLU function, and the second linear layer restores the original channel dimension ($2C \rightarrow 4C$), producing $F_{mix} \in \mathbb{R}^{H \times W \times 4C}$. After applying a skip connection, the mixed features are split back into the four feature maps, each with the same dimensions as the input. Finally, the output features $F_{out} = \{F'_{RGB}, F'_{SRM}, F'_{Bayar}, F'_{NP++}\} \in \mathbb{R}^{H \times W \times C}$ are processed in two ways: (1) passed to the next encoder layer and (2) passed to the decoder after the element-wise summation of the four feature maps. The colored arrows in Fig. 4 denote the flow of the feature map. Once the aggregated feature maps from the four stages are obtained, a localization map is predicted using the MLP decoder proposed in SegFormer [5]. It is a lightweight architecture consisting of MLP layers and bilinear upsampling.

Our simple *concat-mix-split* strategy enables the simultaneous fusion of multiple modalities and flexibly handles a varying number of feature maps. Therefore, it can be easily extended to accommodate scenarios with an even greater diversity of input modalities. In contrast, the CMX fusion [22], an attention-based approach adopted in previous state-of-the-art RGB-Centered Fusion methods [4,20], relies on feature exchange and cross-attention mechanisms designed for dual-branch fusion, making it unsuitable for scenarios involving more than two branches.

3.5. Fixed GeM for image-level detection

To enhance image-level detection performance, recent works have either incorporated image-level loss into the localization training phase [12,21] or introduced a separate training phase solely for detection while freezing the localization network [4,20], which increases the overall training burden. In contrast, we hypothesize that a well-trained localization network can accurately identify tampered regions without requiring image-level supervision, when trained on a sufficiently large dataset containing both manipulated and authentic images. Following on this hypothesis, we freeze our trained localization network and directly apply a fixed operation to the predicted map for image-level detection.

In this paper, we propose Fixed GeM for an effective fixed operation. Fixed GeM is a special case of generalized mean pooling (GeM), a learnable image-level detector employed in MVSS-Net [12] and defined as:

$$\text{GeM}(pred) = \frac{1}{H \times W}\left(\sum_{i=1}^{H}\sum_{j=1}^{W}\left(pred_{i,j}\right)^{p}\right)^{\frac{1}{p}} \tag{4}$$

where $pred$ is a predicted localization map with dimensions $H \times W \times 1$ and p is a trainable parameter that strikes a balance between global mean pooling ($p = 1$) and global max pooling ($p \approx \infty$). We develop Fixed GeM as a training-free module by fixing p at 7, which is an intermediate value between the two GeM branches (initial p of 3 and 10, respectively) of ConvGeM module proposed in MVSS-Net. Consequently, Fixed GeM does not require any additional parameters, loss terms, or training procedures. As a result, image-level detection can be obtained directly from the predicted localization map, improving the efficiency of the overall framework.

3.6. Loss function

During training for localization, weighted binary cross-entropy and dice loss are commonly used as follows:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i}\gamma_0(1-g_i)\log(1-p_i) + \gamma_1 g_i \log p_i, \tag{5}$$

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i p_i \cdot g_i}{\sum_i p_i^2 + \sum_i g_i^2}, \tag{6}$$

where g_i and p_i denote the i-th pixels of ground-truth and predicted map, respectively. Since the training data contain a greater number of pixels from authentic areas, this imbalance is addressed by setting the weights γ_0 to 1 and γ_1 to 5.

In addition, we employ edge loss, which is known to be effective in learning semantic-agnostic features by focusing on the boundary of the tampered region [12,16,26]. Following IML-ViT [26], we obtain the edge mask M' from the ground-truth mask M by using the dilation ($\oplus$) and erosion ($\ominus$) operations as follows:

$$M' = |(M \ominus B(k)) - (M \oplus B(k))|, \tag{7}$$

where $B(k)$ produces a $(2k+1) \times (2k+1)$ cross matrix with only the k-th column and row containing a value of 1, and the rest containing 0. Using the edge mask, the edge loss can also be defined as binary cross-entropy:

$$\mathcal{L}_{edge} = -\frac{1}{N}\sum_i (1 - g'_i)\log(1 - p'_i) + g'_i \log p'_i, \tag{8}$$

where $g' = g * M'$ and $p' = p * M'$; $*$ denotes the point-wise product. The total loss is formulated as follows:

$$\mathcal{L}_{total} = \lambda_{bce}\mathcal{L}_{bce} + (1 - \lambda_{bce})\mathcal{L}_{dice} + \lambda_{edge}\mathcal{L}_{edge}, \tag{9}$$

where λ_{bce} and λ_{edge} are set to 0.3 and 6, respectively.

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

To train our model, we use the datasets following CAT-Net v2 [11], which contain authentic and manipulated images with the ground-truth masks. The training datasets include Casia v2 [27], FantasticReality [28], IMD2020 [29], and a dataset proposed in [11] that applies splicing and copy-move using the COCO [30] training set or RAISE [31] as a source. We evaluate our model on six benchmark datasets commonly used in previous works: Casia v1+[1] [27], DSO-1 [33], Columbia [34], Coverage [35], CocoGlide [4], and NIST16 [36]. Especially, CocoGlide contains locally AI-created images using a diffusion-based generative model [1]. In Table 1, we provide the number of images and types of manipulation in the datasets we used.

4.1.2. Evaluation metrics

We follow the common criteria used in previous works to evaluate our model. For reporting localization performance, we measure the pixel-level F1 score with a fixed threshold of 0.5, which is a practical metric for real-world scenarios. For detection performance, we use image-level Area Under Curve (AUC) and balanced accuracy, which is the arithmetic mean of sensitivity and specificity.

4.1.3. Implementation details

We leverage different types of visual backbones and develop four versions of the MNVFusion framework: MNVFusion-ConvNeXt with ConvNeXt-Small [6], MNVFusion-SegFormer with SegFormer-B2 [5] that introduces Mix Transformer-B2 encoder, MNVFusion-VMamba (T) with Vanilla-VMamba-T [7], and MNVFusion-VMamba (S) with Vanilla-VMamba-S. All backbone networks are initialized with ImageNet-pretrained weights. We follow the data setup procedure proposed by [4]. For each training epoch, an equal number of images from the training datasets are randomly sampled to avoid data imbalance. The input images are resized in the range of [0.5–1.5], randomly cropped to 512 × 512, and performed JPEG compression with a quality factor in the range of [30–100] for data augmentation. We train our model for 70 epochs with a batch size of 4 and perform gradient accumulation to achieve an effective batch size of 24. We use an SGD optimizer with a momentum of 0.9 and weight decay of 0.0005 and set the learning rate to start at 0.005 and decay to zero by using a polynomial scheduler.

[1] Casia v1+ [12] is a fixed version of Casia v1 that replaces duplicated images existing in Casia v2 with images from Corel [32] dataset.

Table 1
Details on training and testing datasets (Sp: Splicing, CM: Copy-Move, Inp: Removal/Inpainting).

Dataset	Number of images		Manipulation		
	Authentic	Manipulated	Sp	CM	Inp
Casia v2	7491	5105	✓		
FantasticReality	16,592	19,423	✓		
IMD2020	414	2010	✓	✓	✓
tampered COCO	–	400 K	✓	✓	
tampered RAISE	24,462	400 K		✓	
Casia v1+	800	921	✓	✓	
DSO-1	100	100	✓		
Columbia	183	180	✓		
Coverage	100	100		✓	
CocoGlide	512	512			✓
NIST16	160	160	✓	✓	✓

4.2. Comparison with state-of-the-art methods

We compare our MNVFusion with existing methods for the general IMDL task in terms of localization and detection performance. The methods include MVSS-Net [12], CAT-Net v2 [11], ManTraNet [9], TruFor [4], Early Fusion and Late Fusion [20], UnionFormer [21], ForMA [37], and KLMN [38]. Most of them utilize the RGB-Centered Fusion approach for multi-modal fusion. Specifically, MVSS-Net, CAT-Net v2, TruFor, and UnionFormer adopt the dual fusion approach. Early Fusion and KLMN adopt the early fusion by applying convolution blocks or learnable combining parameters to mix the noise-view modalities in the early stage. In addition, ForMA can also be regarded as an early fusion approach, which incorporates noise features into the decoder stage. Late Fusion adopts the late fusion that first performs separate fusion for each RGB–X pair, and then merges the results at a later stage. ManTraNet adopts input fusion, which utilizes fusion in the input space by concatenation in the channel dimension.

4.2.1. Localization results

We report the localization performance in Table 2. Overall, MNVFusion significantly outperforms previous methods across all datasets. Compared to MVSS-Net, CAT-Net v2, and ManTraNet, which leverage CNN-based backbones, MNVFusion-ConvNeXt demonstrates superior performance. Similarly, MNVFusion-SegFormer surpasses TruFor, Early Fusion, and Late Fusion in the average performance of the six datasets, while utilizing the same MiT-B2 backbone. This highlights the superiority of our MNVFusion over previous RGB-Centered Fusion methods. Additionally, MNVFusion-SegFormer achieves better results than UnionFormer on the Casia v1+, Columbia, Coverage, and CocoGlide datasets.

When combined with VMamba backbone, MNVFusion achieves the higher performance. MNVFusion-VMamba (T) delivers a better average F1 score than MNVFusion-SegFormer, while maintaining the same number of total parameters. In addition, when compared with ForMA, which also employs the VMamba-T backbone, MNVFusion-VMamba (T) attains superior performance, further validating the effectiveness of our fusion strategy. When scaled up to a larger model, MNVFusion-VMamba (S) achieves state-of-the-art results, outperforming Early Fusion and Late Fusion by 6.7 % on average.

In addition, we compare the number of total parameters. The weight-sharing strategy and Fixed GeM allow MNVFusion to require fewer parameters than previous methods. For a fair comparison of computational efficiency, we further measure the FLOPs of different fusion strategies under the MiT-B2 backbone. In particular, we compare the FLOPs of CM-FRM & FFM module adopted in Early Fusion and Late Fusion with the proposed MB-CMM in MNVFusion-SegFormer. All FLOPs are computed with an input resolution of 512 × 512. The results, summarized in Table 4, show that MB-CMM requires substantially fewer

Table 2
Comparison of pixel-level F1 performance of image manipulation localization. The best results are represented in **bold**. The asterisk (*) marks indicate the performance evaluation conducted using the model checkpoint released in the official code. For UnionFormer, we do not report Params (M) and DSO-1 results since they not reported in their paper and the code is not released.

Method	Extra modality	Backbone	Params (M)	Ca.	DSO.	Col.	Cov.	Coco.	NIST.	AVG
MVSS-Net	Bayar	ResNet-50	146.9	.528	.358	.729	.514	.486	.320	.489
CAT-Net v2	DCT	HRNet	114.3	.752	.584	.859	.381	.434	.308	.553
ManTraNet	SRM, Bayar	VGG-16	4.7	.180	.412	.508	.317	.516	.172	.351
TruFor	NP++	MiT-B2	68.7	.737	.930	.859	.600	.523	.399	.675
Early Fusion	SRM, Bayar, NP++	MiT-B2	68.9	.784	.863	.888	.663	.553	.417*	.695
Late Fusion	SRM, Bayar, NP++	MiT-B2	151.8	.775	.899	.864	.641	.574	.419*	.695
UnionFormer	Bayar, DCT	ResNet, ViT	–	.760	–	.861	.592	.536	.413	–
ForMA	SRM, Bayer, NP++	VMamba-T	37.0	.729	.387	.949	.587	.453	.454	.593
KLMN	SRM, Bayar, DCT	MiT-B0	10.9	.476	.918	.892	.387	.464	.357	.582
MNVFusion-ConvNeXt	SRM, Bayar, NP++	ConvNeXt-S	63.8	.729	.713	.893	.576	.607	.466	.664
MNVFusion-SegFormer	SRM, Bayar, NP++	MiT-B2	33.5	.788	.883	.948	.699	.582	.369	.712
MNVFusion-VMamba (T)	SRM, Bayar, NP++	VMamba-T	36.5	.806	.831	.952	**.746**	**.658**	.482	.746
MNVFusion-VMamba (S)	SRM, Bayar, NP++	VMamba-S	58.0	**.809**	**.970**	**.966**	.695	.613	**.518**	**.762**

Table 3
Comparison of image-level AUC and balanced accuracy in detecting image manipulation.

Method	Casia v1+		DSO-1		Columbia		Coverage		CocoGlide		NIST16		AVG	
	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc
MVSS-Net	.932	.808	.552	.485	.984	.667	.733	.545	.654	.536	.579	.538	.739	.597
CAT-Net v2	.942	.838	.747	.525	.977	.803	.680	.635	.667	.580	.750	.597	.794	.663
ManTraNet	.644	.500	.874	.500	.810	.500	.760	.500	.778	.500	.624	.500	.748	.500
TruFor	.916	.813	.984	.930	.996	**.984**	.770	.680	.752	.639	.760	.662	.863	.785
Early Fusion	.929	.845	.966	**.935**	.996	.962	**.839**	**.770**	.755	.660	.693*	.641*	.863	.802
Late Fusion	.930	**.860**	.958	.830	.977	.822	.792	.720	.760	.677	.695*	.647*	.852	.759
UnionFormer	**.951**	.843	–	–	.998	.979	.783	.694	**.797**	.682	.793	.680	–	–
MNVFusion-ConvNeXt	.841	.773	.969	.880	.993	.956	.755	.655	.714	.651	.769	.669	.840	.764
MNVFusion-SegFormer	.899	.828	.980	**.935**	.995	.962	.810	.710	.754	.699	.719	.644	.860	.796
MNVFusion-VMamba (T)	.899	.833	.960	.905	.991	.981	.835	.755	.771	.683	.768	.681	.871	.806
MNVFusion-VMamba (S)	.908	.847	**.987**	.910	**1.000**	.975	.817	.720	.773	**.704**	**.802**	**.712**	**.881**	**.811**

Table 4
FLOPs comparison between CMX Fusion (CM-FRM & FFM) and MB-CMM under the MiT-B2 backbone with 512 × 512 input resolution.

Methods	Fusion method	FLOPs (G)
Early Fusion	CM-FRM & FFM	6.40
Late Fusion	CM-FRM & FFM	19.21
MNVFusion-SegFormer	MB-CMM	4.90

FLOPs while maintaining superior accuracy, thereby confirming its effectiveness as a lightweight and efficient fusion design.

4.2.2. Detection results

In Table 3, we compare the detection performance. Since the proposed Fixed GeM computes the final image-level decision score as a byproduct of the predicted localization map, the detection performance is greatly affected by the quality of the localization map. Consequently, the average AUC and balanced accuracy of the MNVFusion series follow the same trend as their localization performance. While not achieving the top performance on every individual dataset, MNVFusion-VMamba (S) demonstrates the highest average performance across six benchmark datasets, surpassing Early Fusion by 1.8 % in AUC and 0.9 % in balanced accuracy. Compared to UnionFormer, however, the improvement in detection performance is relatively limited. Considering that MNVFusion does not utilize any trainable module for image-level detection, these results still highlight the competitiveness of our approach in achieving comparable detection accuracy with a simpler design.

4.2.3. Qualitative results

We compare the visual results of MNVFusion-VMamba (S) with state-of-the-art methods for both manipulated and authentic images in Fig. 5. Compared to existing methods, MNVFusion clearly identifies the tampered regions in manipulated images. Furthermore, the proposed model does not respond to authentic images, proving that MNVFusion effectively prevents false alarms and demonstrates robust performance in the image-level detection task.

4.3. Comparison on AI-generated manipulation artifacts

With the rapid advancements in image generation and transformation, recent generative models have been increasingly employed to create highly realistic manipulations. To further evaluate the robustness of our framework under such challenging conditions, we conduct additional experiments on two recently released AI-generated manipulation datasets: AutoSplice [39] and CSI-IMD [40]. The AutoSplice dataset leverages the DALL·E2 [41] language-image model to automatically generate and splice masked regions guided by text prompts. Human verification was incorporated to ensure the realism of the manipulations, resulting in a dataset of 5894 authentic and manipulated images. The CSI-IMD dataset focuses on the semantic significance of manipulations by providing detailed annotations that capture the semantic impact of edits. It consists of a gold-standard set of 1000 manually annotated manipulations and an extended set of 500,000 automatically generated examples. The manipulations are created using advanced generation techniques, primarily based on the Stable Diffusion model [42].

For evaluation, we utilize our pre-trained MNVFusion-VMamba (T) and MNVFusion-VMamba (S) models to perform the image manipulation localization task on these datasets. The quantitative results are reported in Tables 5 and 6. MNVFusion-VMamba (T) achieves

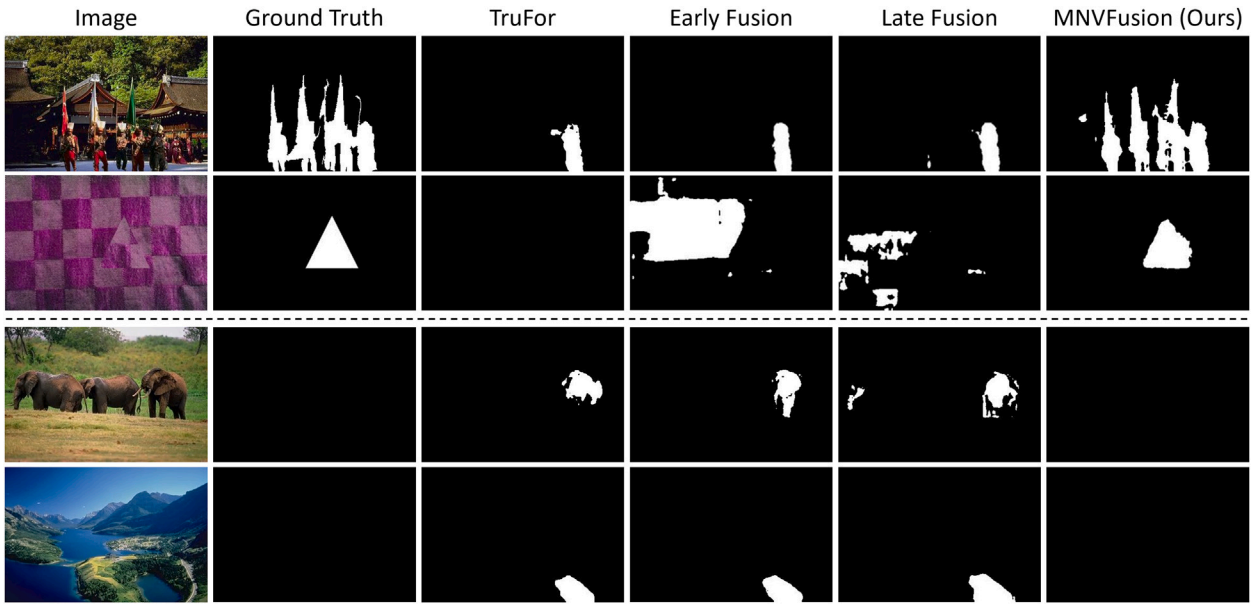


Fig. 5. Visualization examples of predicted localization maps for manipulated (top) and authentic (bottom) images.

Table 5
Pixel-level localization performance comparison on the AutoSplice dataset. The evaluation is conducted on 3621 manipulated images.

Methods	F1	IoU	Prec.	Acc.
NoisePrint	.333	.217	.390	.480
ManTraNet	.179	.120	.639	.586
ForensicsGraph	.362	.289	.393	.530
CAT-Net	.751	.648	.884	.827
MVSS-Net	.330	.238	.734	.677
PSCC-Net	.558	.447	.899	.725
ViT-VAE	.156	.115	.245	.560
Early Fusion	.790	.685	.836	.791
Late Fusion	.789	.687	.821	.786
MNVFusion-VMamba (T)	**.864**	**.785**	**.938**	**.879**
MNVFusion-VMamba (S)	.828	.734	.900	.837

Table 6
Pixel-level localization performance comparison on the CSI-IMD dataset, evaluated on the 1000 gold-standard set images.

Methods	F1	IoU	Prec.	Acc.
ManTraNet	.020	.009	.176	.010
CAT-Net v2	.718	.614	.671	.772
TruFor	.737	.652	.713	.786
CRCNN	.007	.002	.060	.004
HiFi-Net	.040	.018	.117	.024
IF-OSN	.110	.041	.204	.075
ObjectFormer	.012	.006	.035	.008
PSCC-Net	.232	.139	.239	.226
RRU-Net	.218	.102	.278	.180
SPAN	.000	.000	.007	.000
Early Fusion	**.814**	**.763**	**.796**	**.912**
Late Fusion	.768	.717	.755	.850
MNVFusion-VMamba (T)	.736	.658	.737	.855
MNVFusion-VMamba (S)	.740	.660	.747	.872

the highest score on the AutoSplice dataset, surpassing representative baselines such as NoisePrint [19], ManTraNet, ForensicsGraph [43], CAT-Net [44], MVSS-Net, PSCC-Net [45], ViT-VAE [46], Early Fusion, and Late Fusion. On the CSI-IMD dataset, MNVFusion-VMamba models demonstrate superior performance over the majority of existing approaches such as ManTraNet, CAT-Net v2, TruFor, CRCNN [47], HiFi-Net [48], IF-OSN [49], ObjectFormer [50], PSCC-Net, RRU-Net [51], and SPAN [10], although their results remain marginally lower than those achieved by the Early Fusion. These findings highlight that our proposed framework maintains robust performance even under unseen and semantically challenging manipulations, supporting the generalizability of MNVFusion.

4.4. Ablation studies

4.4.1. Effect of core designs

In this section, we remove or replace each component of our method one at a time and compare the localization results to show the impact of our design choices. To consider all components, we select a smaller baseline model and training scheme by training MNVFusion-VMamba (T) solely on the Casia v2 dataset. All other training details remain the same. For performance evaluation, we report the pixel-level F1 scores of the manipulated images in the test datasets and the relative decrease in average F1 scores compared to the baseline. The results are listed in Table 7.

Model architecture. Through *#1*–*#3*, we analyze the effectiveness of our core designs. The effect of MB-CMM is revealed in *#1*, where its removal leads to a substantial decrease (-9.8 %). In *#2*, we explore the integration of additional channel embedding module from FFM [22] during the fusion of feature maps at the final stage of MB-CMM, which introduces extra parameters and computations. However, the results reveal that this added complexity does not lead to improved performance. Instead, our simple summation strategy remains more effective and computationally efficient for final feature fusion. In *#3*, we replace the MLP decoder with the Mamba decoder proposed in Sigma [52], which includes the Visual State Space (VSS) block for enhancing inter-channel information. Again, the results indicate that the simple MLP decoder performs better on the IMDL task.

Fusion method in dual-branch. We verify the advantages of the proposed multi-branch encoder framework and demonstrate the effectiveness of our MB-CMM fusion strategy compared to other fusion

Table 7
Ablation studies of MNVFusion for localization performance. (↓) indicates a relative decrease in average F1 score percentage compared to the baseline. In all experiments, the Casia v2 dataset is used for training.

Method	Ca.	DSO.	Col.	Cov.	Coco.	NIST.	AVG (↓)
MNVFusion(Proposed)	.714	.649	.701	.479	.489	.340	.562 (0.0 %)
Model Architecture							
#1. w/o MB-CMM	.692	.319	.647	.430	.437	.264	.465 (−9.8 %)
#2. w/ Channel Embedding to Decoder	.718	.435	.698	.523	.518	.326	.536 (−2.6 %)
#3. Mamba Decoder as Decoder	.698	.436	.646	.501	.502	.328	.519 (−4.3 %)
Fusion Method in Dual-Branch							
#4. MB-CMM	.704	.396	.742	.507	.496	.301	.524 (−3.8 %)
#5. CM-FRM & FFM	.561	.590	.477	.466	.598	.295	.498 (−6.4 %)
#6. CroMB & ConMB	.696	.304	.672	.404	.433	.343	.475 (−8.7 %)
Modality							
#7. RGB	.700	.275	.730	.423	.442	.306	.479 (−8.3 %)
#8. RGB + SRM	.704	.318	.720	.490	.495	.291	.503 (−5.9 %)
#9. RGB + BayarConv	.715	.300	.731	.429	.444	.295	.486 (−7.6 %)
#10. RGB + NoisePrint++	.685	.623	.659	.405	.452	.361	.531 (−3.1 %)
Loss							
#11. w/o Edge loss	.703	.388	.657	.484	.491	.316	.507 (−5.5 %)

mechanisms through *#4* − *#6*. We modify the proposed model in a dual-branch manner by using Early Convolution employed in Early Fusion [20], which first integrates the three extra input modalities into a single mixed feature. For feature fusion of the RGB and X branches, we evaluate three approaches: MB-CMM, CM-FRM & FFM [22], which is based on attention mechanisms, and CroMB & ConMB [52], which is built upon State Space Models (SSMs). The performance drop in *#4* compared to our baseline indicates that the proposed multi-branch framework outperforms the dual-branch framework in processing multi-modal inputs. Notably, the superior performance of *#4* over its counterparts (*#5*, *#6*) suggests that our simple MLP-based fusion strategy in MB-CMM is more effective than both attention-based and SSM-based fusion approaches. Additionally, MB-CMM can be directly extended to multi-branch fusion, while CM-FRM & FFM and CroMB & ConMB can only be applied to dual-branch fusion.

Modality. Through *#7* − *#10*, we compare the effects of different input modalities. The use of RGB alone as the input for a single-branch encoder without fusion module leads to the worst result. Performance improves when additional modalities are applied, and the best average F1 metric is achieved when using the baseline, in which all modalities are employed.

Loss. In *#11*, we remove the edge loss in Eq. (8) from the total loss. The resulting performance degradation indicates that the edge loss enhances the learning of the semantic-agnostic features substantially.

4.4.2. Effect of detection module

In Table 8, we compare our proposed Fixed GeM with two Fixed GeM variants and the ConvGeM proposed in MVSS-Net [12]. For a fair comparison, we freeze our trained localization network, MNVFusion-VMamba (S), and evaluate performance by varying only the detection module. The Fixed GeM variants include $p = 1$, which is equivalent to global mean pooling of all pixels, and $p = 100$, which approximates global max pooling. The three Fixed GeM methods are training-free modules with fixed p values. In contrast, ConvGeM is a trainable module that incorporates convolutional layers and a GeM operation with a learnable p. Using the same datasets for localization training, we train only the ConvGeM module for 30 epochs with a batch size of 16 and an initial learning rate of 0.001. We initialize p value to 7. The results show that our Fixed GeM ($p = 7$) achieves a comparable average AUC and balanced accuracy performance to ConvGeM (initial $p = 7$), indicating that the addition of learnable components to Fixed GeM does not significantly improve overall performance. In contrast, Fixed GeM shows poor balanced accuracy when extreme p values are used. Further, we investigate the sensitivity of Fixed GeM to the pooling parameter p by conducting experiments over a wide range of p values, as illustrated in Fig. 6. The results demonstrate that Fixed GeM maintains strong AUC and balanced accuracy across different p values, based on the average performance over six datasets, confirming its robustness.

Table 8
Ablation results on benchmark datasets using different image-level detection methods.

Dataset	Metric	Method			
		Fixed GeM ($p = 7$)	Fixed GeM ($p = 1$)	Fixed GeM ($p = 100$)	ConvGeM ($p = 7$)
Casia v1 +	AUC	.908	.848	**.918**	.902
	bAcc	.847	.502	.616	**.863**
DSO-1	AUC	.987	.917	.981	**.988**
	bAcc	.910	.500	.500	**.940**
Columbia	AUC	**1.000**	.988	.995	.995
	bAcc	.975	.536	.577	**.983**
Coverage	AUC	**.817**	.764	.805	.805
	bAcc	.720	.500	.560	**.745**
CocoGlide	AUC	**.773**	.753	.763	.756
	bAcc	**.704**	.528	.621	.687
NIST16	AUC	.802	.656	.814	**.817**
	bAcc	**.712**	.503	.509	.672
AVG	AUC	**.881**	.821	.879	.877
	bAcc	.811	.512	.564	**.815**

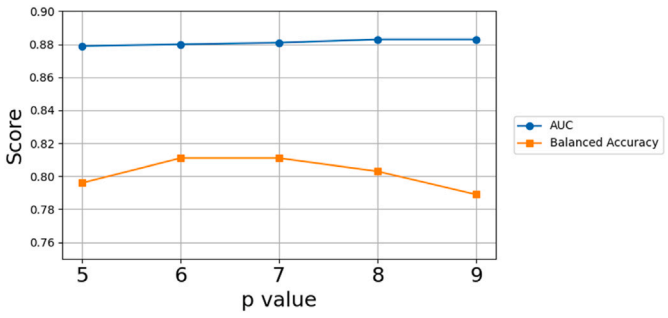


Fig. 6. Average AUC and balanced accuracy using different p values for Fixed GeM.

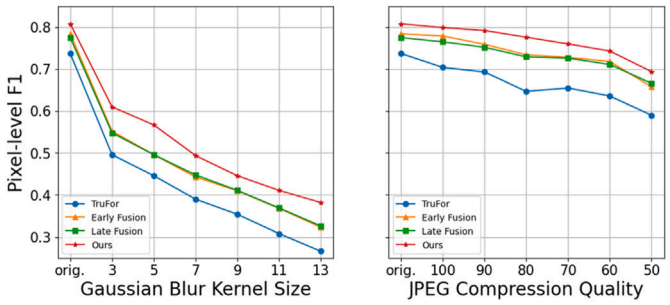


Fig. 7. Robustness analysis under Gaussian blur (left) and JPEG compression (right).

4.5. Robustness analysis

We perform experiments using corrupted images to verify the robustness of our model. For evaluation, we use the Casia v1 + dataset and apply two degrading transformations—Gaussian blur with different kernel sizes and JPEG compression with different quality factors—both of which are commonly used in prior studies to simulate post-processing artifacts. We compare MNVFusion-VMamba (S) with TruFor, Early Fusion, and Late Fusion, the previous state-of-the-art methods. The results in Fig. 7 show that MNVFusion-VMamba (S) consistently outperforms previous methods across various degradation levels, demonstrating its robustness to image distortions.

5. Conclusion

In this paper, we introduce the MNVFusion framework for image manipulation detection and localization. MNVFusion integrates the complementary information of diverse noise-view modalities through separate branches, thereby reducing reliance on RGB input. To fully exploit the synergy between modalities, we propose MB-CMM, a multi-branch fusion module that interactively refines the features extracted by the encoder. MB-CMM flexibly handles features from more than two modalities using a simple channel-mixing fusion strategy. Additionally, MNVFusion enhances efficiency by training exclusively for the localization task, while detection results are directly derived from the predicted localization map. The proposed method achieves state-of-the-art results on benchmark datasets and demonstrates robust performance when applied to images degraded by various types of distortion.

6. Limitations

In this work, we mainly focus on utilizing noise-view modalities to supplement the RGB image. While these noise-aware representations yield promising results, we do not explore other representation domains. Frequency-domain transforms, such as the Discrete Cosine Transform (DCT) or wavelet-based methods, may offer complementary information to further enhance performance. In addition, our robustness evaluation is limited to common post-processing artifacts. Evaluating the model under more diverse and challenging conditions, such as adversarial perturbations or unseen corruptions, would offer a more comprehensive understanding of its robustness. We leave these aspects as potential directions for future work to further improve the applicability and generalization of the proposed framework.

CRediT authorship contribution statement

Joonkyo Shim: Writing – original draft, Validation, Methodology, Investigation. **Hyunsoo Yoon:** Writing – review & editing, Supervision, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work the author(s) used none of the generative AI and AI-assisted technologies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2025-02305884) and by the National Research Foundation of Korea (NRF) grant funded by the Korean government (RS-2023-00213798).

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neucom.2025.131915.

Data availability

Data will be made available on request.

References

[1] A.Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, M. Chen, Glide: towards photorealistic image generation and editing with text-guided diffusion models, in: International Conference on Machine Learning, PMLR, 2022, pp. 16784–16804.
[2] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 7 (3) (2012) 868–882.
[3] B. Bayar, M.C. Stamm, A DEEP learning approach to universal image manipulation detection using a new convolutional layer, in: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016, pp. 5–10.
[4] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, Trufor: leveraging all-round clues for trustworthy image forgery detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20606–20615.
[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[7] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: visual state space model, arXiv preprint arXiv:2401.10166, 2024.
[8] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Learning rich features for image manipulation detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1053–1061.
[9] Y. Wu, W. AbdAlmageed, P. Natarajan, Mantra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9543–9552.
[10] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, R. Nevatia, Span: spatial pyramid attention network for image manipulation localization, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, 2020, pp. 312–328.
[11] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, C. Kim, Learning JPEG compression artifacts for image manipulation detection and localization, Int. J. Comput. Vis. 130 (8) (2022) 1875–1895.
[12] C. Dong, X. Chen, R. Hu, J. Cao, X. Li, MVSS-Net: multi-view multi-scale supervised networks for image manipulation detection, IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2022) 3539–3553.
[13] I.I. Ganapathi, S. Javed, S.S. Ali, A. Mahmood, N.-S. Vu, N. Werghi, Learning to localize image forgery using end-to-end attention network, Neurocomputing 512 (2022) 25–39.
[14] D. Xu, X. Shen, Z. Shi, N. Ta, Semantic-agnostic progressive subtractive network for image manipulation detection and localization, Neurocomputing 543 (2023) 126263.
[15] X. Jin, W. Yu, W. Shi, Image manipulation localization VIA dynamic cross-modality fusion and progressive integration, Neurocomputing 610 (2024) 128607.
[16] H. Wang, H. Cheng, Y. Chen, Y. Xu, M. Wang, Image manipulation localization VIA semantic-guided feature enhancement and DEEP multi-scale EDGE supervision, Neurocomputing 639 (2025) 130255.

[17] X. Yang, X. Chai, Z. Gan, L. Cao, Y. Zhang, MSHRT-net: multi-scale hierarchical residual transfer network for image manipulation detection and localization, Neurocomputing 648 (2025) 130788.
[18] N. Ahmed, T. Natarajan, K.R. Rao, Discrete cosine transform, IEEE Trans. Comput. 100 (1) (1974) 90–93.
[19] D. Cozzolino, L. Verdoliva, Noiseprint: a CNN-based camera model fingerprint, IEEE Trans. Inf. Forensics Secur. 15 (2019) 144–159.
[20] K. Triaridis, V. Mezaris, Exploring multi-modal fusion for image manipulation detection and localization, in: International Conference on Multimedia Modeling, Springer, 2024, pp. 198–211.
[21] S. Li, W. Ma, J. Guo, S. Xu, B. Li, X. Zhang, Unionformer: unified-learning transformer with multi-view representation for image manipulation detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12523–12533.
[22] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, R. Stiefelhagen, CMX: cross-modal fusion for RGB-X semantic segmentation with transformers, IEEE Trans. Intell. Transp. Syst. 24 (12) (2023) 14679–14694.
[23] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a Gaussian denoiser: residual learning of DEEP CNN for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.
[24] K. He, X. Zhang, S. Ren, J. Sun, DEEP residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[25] A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752, 2023.
[26] X. Ma, B. Du, X. Liu, A.Y.A. Hammadi, J. Zhou, IML-ViT: image manipulation localization by vision transformer, arXiv preprint arXiv:2307.14863, 2023.
[27] J. Dong, W. Wang, T. Tan, Casia image tampering detection evaluation database, in: 2013 IEEE China Summit and International Conference on Signal and Information Processing, IEEE, 2013, pp. 422–426.
[28] V.V. Kniaz, V. Knyaz, F. Remondino, The point where reality meets fantasy: mixed adversarial generators for image splice detection, Adv. Neural Inf. Process. Syst. 32 (2019).
[29] A. Novozamsky, B. Mahdian, S. Saic, IMD2020: a large-scale annotated dataset tailored for detecting manipulated images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, 2020, pp. 71–80.
[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part v 13, Springer, 2014, pp. 740–755.
[31] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, Raise: a RAW images dataset for digital image forensics, in: Proceedings of the 6th ACM Multimedia Systems Conference, 2015, pp. 219–224.
[32] J.Z. Wang, J. Li, G. Wiederhold, Simplicity: semantics-sensitive integrated matching for picture libraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (9) (2001) 947–963.
[33] T.J. De Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, A. de Rezende Rocha, Exposing digital image forgeries by illumination color classification, IEEE Trans. Inf. Forensics Secur. 8 (7) (2013) 1182–1194.
[34] Y.-F. Hsu, S.-F. Chang, Detecting image splicing using geometry invariants and camera characteristics consistency, in: 2006 IEEE International Conference on Multimedia and Expo, IEEE, 2006, pp. 549–552.
[35] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, S. Winkler, Coverage–A novel database for copy-move forgery detection, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 161–165.
[36] H. Guan, M. Kozak, E. Robertson, Y. Lee, A.N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, J. Fiscus, MFC datasets: large-scale benchmark datasets for media forensic challenge evaluation, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, 2019, pp. 63–72.
[37] K. Guo, G. Cao, Z. Lou, X. Huang, J. Liu, A lightweight and effective image tampering localization network with vision Mamba, IEEE Signal Process. Lett. 32 (2025) 2179–2183.
[38] H. Huang, Y. Liu, X. Jin, S. Xiao, B. Liu, KLMN: knowledge distillation based lightweight multi-clue image forgery detection and localization, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.
[39] S. Jia, M. Huang, Z. Zhou, Y. Ju, J. Cai, S. Lyu, Autosplice: a text-prompt manipulated image dataset for media forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 893–903.
[40] Y. Chen, M.-C. Chang, M. Kirchner, Z. Zhang, X. Li, A. Basharat, A. Hoogs, A semantically impactful image manipulation dataset: characterizing image manipulations using semantic significance, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 7659–7668.
[41] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125, 2022 1 (2) 3.
[42] H. Face, Runwayml stable diffusion v1. 5 (2024).
[43] O. Mayer, M.C. Stamm, Exposing fake images with forensic similarity graphs, IEEE J. Sel. Top. Signal Process. 14 (5) (2020) 1049–1064.
[44] M.-J. Kwon, I.-J. Yu, S.-H. Nam, H.-K. Lee, CAT-Net: compression artifact tracing network for detection and localization of image splicing, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 375–384.
[45] X. Liu, Y. Liu, J. Chen, X. Liu, PSCC-Net: progressive spatio-channel correlation network for image manipulation detection and localization, IEEE Trans. Circuits Syst. Video Technol. 32 (11) (2022) 7505–7517.
[46] T. Chen, B. Li, J. Zeng, Learning traces by yourself: blind image forgery localization VIA anomaly detection with ViT-VAE, IEEE Signal Process. Lett. 30 (2023) 150–154.
[47] C. Yang, H. Li, F. Lin, B. Jiang, H. Zhao, Constrained R-CNN: a general image manipulation detection model, arXiv preprint arXiv:1911.08217, 2019.
[48] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, X. Liu, Hierarchical fine-grained image forgery detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3155–3165.
[49] H. Wu, J. Zhou, J. Tian, J. Liu, Robust image forgery detection over online social network shared images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13440–13449.
[50] J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, Y.-G. Jiang, Objectformer for image manipulation detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2364–2373.
[51] X. Bi, Y. Wei, B. Xiao, W. Li, RRU-Net: the ringed residual U-Net for image splicing forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[52] Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, Y. Xie, Sigma: siamese mamba network for multi-modal semantic segmentation, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 1734–1744.

Author biography

Joonkyo Shim earned a BS degree in Department of Industrial Engineering from Yonsei University, Seoul, Republic of Korea, in 2023. Currently, he is pursuing a Ph.D's degree in industrial engineering at Yonsei University with research interests encompassing representation learning, deepfake disruption, and image manipulation detection.

Hyunsoo Yoon is an associate professor in the Department of Industrial Engineering at Yonsei University. He received his BS and MS in industrial engineering from Korea University in 2010 and 2012, respectively. He completed his second MS in statistics from Georgia Institute of Technology in 2013, followed by a Ph.D. in industrial engineering from Arizona State University in 2018. Before joining Yonsei University, he held positions as an assistant professor at the State University of New York at Binghamton and as a post-doctoral fellow at the ASU-Mayo Clinic Center for Innovative Imaging. His research focuses on transfer learning of heterogeneous data sources for predictive analytics, domain adaptation using, and anomaly detection. He is a member of IISE, INFORMS, and IEEE.