DAE-Net: Dual Attention Mechanism and Edge Supervision Network for Image Manipulation Detection and Localization

Chunyin Shi, Chengyou Wang, *Member, IEEE*, Xiao Zhou, *Member, IEEE*, Zhiliang Qin, *Senior Member, IEEE*

***Abstract*—Image manipulation detection and localization aims to identify tampered regions from a suspicious image. Although existing works have achieved impressive results, they fail to capture sufficient tampering artifacts resulting in inaccurate localization of tampered regions. In addition, the problem of poor robustness to post-processing methods also occurs frequently. To address these problems, this paper presents a dual attention mechanism and edge supervision network (DAE-Net) to locate forged regions. Specifically, DAE-Net first extracts multi-scale semantic features from RGB images and captures global and local inconsistent noise features from noise maps. Subsequently, to mine the complementary relationship among different domains and obtain more discriminative feature representations, a dual attention mechanism module is designed to integrate semantic features and noise features. Then, we exploit a mask generation module consisting of fusion upsampling module (FUM) to generate tampered regions in a progressive decoding manner. FUM fully fuses multi-scale semantic features to enhance tampering traces. Finally, an edge supervision module is introduced to capture subtle edge information to optimize the boundary of tampered regions. The experimental results on five public datasets demonstrate that DAE-Net outperforms the state-of-the-art methods in terms of higher localization accuracy and stronger robustness under widely-used evaluation metrics. The F_1/AUC of the proposed DAE-Net reach 0.494/0.805, 0.871/0.984, 0.972/0.973, 0.330/0.741, and 0.338/0.826 on CASIA, NIST16, Columbia, Coverage, and IMD2020, respectively.**

***Index Terms*—Image forensics, image manipulation detection and localization (IMDL), dual attention mechanism, edge supervision, progressive decoding manner.**

I. Introduction

This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2021MF060; in part by the National Key Research and Development Program of China under Grant 2023YFC3321601; in part by the Joint Fund of Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003; in part by the National Natural Science Foundation of China under Grant 61702303; and in part by the Student Research Training Program (SRTP) at Shandong University, Weihai, under Grant A23264 and A24256. *(Corresponding author: Chengyou Wang)*

Chunyin Shi is with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China (e-mail: shicy@mail.sdu.edu.cn).

Chengyou Wang and Xiao Zhou are with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China, also with the Shandong Key Laboratory of Intelligent Communication and Sensing-Computing Integration, Shandong University, Jinan 250061, China, and also with the Shandong University–Weihai Research Institute of Industrial Technology, Weihai 264209, China (e-mail: wangchengyou@sdu.edu.cn; zhouxiao@sdu.edu.cn).

Zhiliang Qin is with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China, and also with the Weihai Beiyang Electrical Group Co. Ltd., Weihai 264209, China (e-mail: qinzhiliang@beiyang.com).

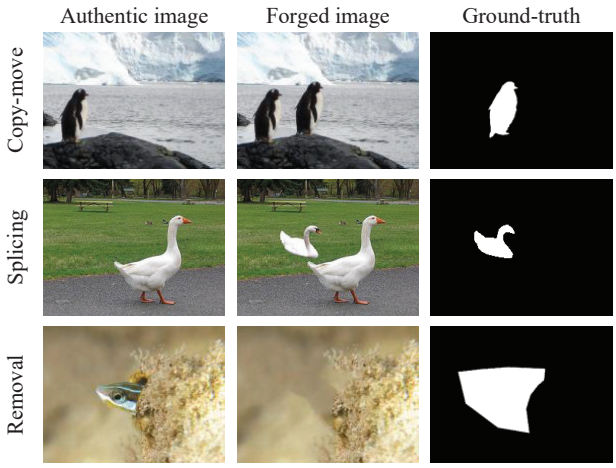


Fig. 1. Examples of content-changed based image manipulation techniques, i.e., copy-move, splicing, and removal. The first column is authentic image, the second column is forged image and the third column is ground-truth mask.

NOWADAYS, with the emergence of various user-friendly image editing tools, people can easily modify digital images [1]. These tampered images appear realistic and have no obvious difference from the common images [2]. When the tampered images appear in application scenarios that are extremely sensitive to image content, such as news media, academic research, and military communication [3], this will cause a trust crisis and threaten social stability. Therefore, it is necessary to develop trustworthy models for identifying forged images.

In general, image manipulation techniques can be divided into two categories: content-dependent process and content-independent process [4]. The former is more harmful than the latter since it alters the semantic contents of the image and is more misleading. The content-dependent process includes copy-move [5], splicing [6], and removal [7], as shown in Fig. 1. The content-independent process includes brightness change, blurring, nosing, and image compression [4]. They make global modifications and barely create any disinformation, but they may hide visual clues and discrepancies between authentic and manipulated regions.

The issue of image manipulation detection and localization (IMDL) has received widespread attention. Many researchers have proposed different methods for IMDL in the past. In the early stages, conventional IMDL methods relied on hand-

crafted features, such as feature point matching [8], color filter array (CFA) [9], noise inconsistency [10], discrete wavelet transform (DWT) [11], discrete cosine transform (DCT) [12], and JPEG compression artifact [13]. These methods are dependent on prior knowledge and can deal with specific types of tampering techniques. Nevertheless, when facing with complex and unknown tampering types, the accuracy and robustness of these methods are insufficient for practical applications. Recently, deep learning has demonstrated strong performance in computer vision tasks, such as image segmentation [14], image classification [15], and object detection [16], because of its adaptive feature extraction capability. Therefore, deep learning techniques have also been applied to address the issue of IMDL. A portion of methods are designed to handle single type of manipulation such as image copy-move forgery detection [17], image splicing detection [18], and image inpainting detection [19], which are not suitable for general IMDL. To detect multiple manipulation types, generic IMDL models have been proposed, such as ManTra-Net [20], RGB-N [21], and SPAN [22]. These methods can efficiently extract tampering artifacts and achieve manipulation localization at the pixel level.

These methods can be divided into two categories based on the network structure: single branch networks and multi-branch networks. Single branch networks [3], [23] directly process the forged images and exploit convolutional neural network (CNN) to extract the RGB features for localizing tampered regions. This could lose the tampering artifacts that cannot be captured in the RGB domain, resulting in false localization of tampered regions. Multi-branch networks utilize multiple CNNs in parallel or sequential order to capture diverse features to realize accurate localization, such as noise features [21], [24], frequency domain features [25], [26], and edge features [2]. These approaches based on noise features can enhance tampering traces and suppress image semantic content. Noise maps are extracted in three main categories: steganalysis rich model (SRM) filters [21], Bayar convolution (BayarConv) [27], and combined convolution [28]. To achieve more refined tampered boundaries, methods [2], [24] of the edge optimization are proposed.

Although the aforementioned approaches can perform general image manipulation detection and achieve tampered region localization at the pixel level. However, there are still three unsolved problems for IMDL: (1) Image forensic approaches based on CNN lose the image semantic information with the deepening of network, resulting in inaccurate manipulation localization in the process of mask generation. (2) Multi-branch methods perform simple concatenation only in the channel dimension when different features are fused, failing to explore the complementary relationship of the features in different domains and achieve effective feature fusion. (3) Existing methods suffer from rough and incomplete boundaries due to ignoring or destroying subtle edge artifacts. However, the existing methods cause degradation in detection performance when detecting some forged images that have been processed by high noise or high compression. This is because forged images after processing by high noise and high compression will seriously damage the tampering artifacts of the image and the inconsistency of the boundaries between the forged and real regions, which makes it difficult for the model to extract effective feature representations. In addition, since the size of the tampered area is generally random, which requires the model to have a strong sensitivity to the scale transformation of the tampered area. Since small tampered areas produce sample imbalance and contain fewer tampering artifacts, most of the existing image manipulation detection methods mainly focus on how to use the network to better extract the tampered features, ignoring the sample imbalance caused by small tampered areas, which affects the improvement of the model's detection performance.

To improve the accuracy of manipulation localization and enhance the robustness against image post-processing operations, we propose a dual attention mechanism and edge supervision network (DAE-Net) for IMDL, as shown in Fig. 2. DAE-Net focuses on detecting three common manipulated techniques, i.e., copy-move, splicing, and removal. Compared to the existing methods, the proposed DAE-Net exploits multiple tampering artifacts for IMDL. Specifically, we introduce a RGB branch to extract multi-scale semantic features for tampered regions prediction and edge information extraction. Meanwhile, noise branch employs the backbone of swin-transformer [29] to capture noise features that fuse global and local inconsistent information. To obtain more discriminative feature representations, a dual attention mechanism module is designed to mine the complementary relationship among different domain features by computing the self-attention in the spatial and channel dimensions. Subsequently, the mask generation module consisting of fusion upsampling module (FUM) localizes tampered regions in a progressive decoding manner. The FUM exploits multi-scale semantic information to improve the accuracy of localization. Finally, to optimize the boundary of tampered regions, the edge supervision module is introduced to enhance the subtle edge features and guide the network to predict more refined tampered regions. Extensive experiments are conducted on five public datasets. The experimental results show that the proposed DAE-Net outperforms state-of-the-art (SoTA) methods.

In summary, the main contributions of this work are as follows:

- We propose a dual attention mechanism and edge supervision network for IMDL, i.e., DAE-Net, which extracts multiple tampering clues to more accurately localize tampered region, including semantic features, fused global and local inconsistent noise features, and edge features.
- A dual attention mechanism module is designed to mine the complementary relationship between semantic features and noise features. This is achieved by computing self-attention in both spatial and channel dimensions, thus enabling sufficient fusion of features from different domains.
- In the mask generation stage, a fusion upsampling module is proposed to integrate multi-scale semantic features to accurately localize tampered regions in a progressive decoding manner. Meanwhile, an edge supervision module is utilized to optimize the boundary of tampered regions.
- Experimental results on five public datasets demonstrate

TABLE I
SUMMARY OF SOTA METHODS FOR IMAGE MANIPULATION DETECTION AND LOCALIZATION.

Category	Method	Forensic Clues		Backbone	Attention	Scales of Supervision		
		RGB	Noise			Pixel	Image	Edge
Unsupervised	CFA1 [9]	+	Local CFA	-	-	+	-	-
	NOI1 [10]	-	Noise	-	-	+	-	-
DNN-Based	ManTra-Net [20]	+	BayarConv, SRM Filter	Wider VGG	-	+	-	-
	SPAN [22]	+	BayarConv, SRM Filter	Wider VGG	Spatial Pyramid Attention	+	-	-
	DenseFCN [30]	+	-	Dense Block	-	+	-	-
	SATFL [31]	+	-	Wider VGG	Forgery Attention	+	-	-
	PSCC-Net [4]	+	-	HR-Net	Spatio-Channel Correlation	+	+	-
	TB-Net [25]	+	DCT	ResNet101	Adaptive Cross-Attention	+	-	-
	TA-Net [2]	+	-	Res2Net50	Reverse Attention Mining	+	-	+
	DMFF-Net [32]	+	BayarConv	ResNet50	Feature Fusion Module	+	-	-
	SF-Net [33]	+	DCT	HRNet	Spatial-Channel Fusion Excitation	+	-	+
	MVSS-Net [24]	+	BayarConv	ResNet50	Channel Attention, Position Attention	+	+	+
	DAE-Net (ours)	+	BayarConv	ResNet50, Swin-Transformer	Dual Attention, Attention Fusion Module	+	-	+

that the proposed DAE-Net outperforms the SoTA methods in terms of tampered localization accuracy and model robustness.

The rest of this paper is organized as follows: Section II presents the related work on IMDL. The proposed approach is described in details in Section III. Section IV presents the experimental results and corresponding discussions. Finally, Section V concludes this paper and discusses current limitations and future research directions.

II. RELATED WORK

In this section, we review related work on deep neural network (DNN) based image manipulation detection and localization. The methods for attention mechanism and edge supervision are introduced, and the differences between this work and the existing work are pointed out. The related work is briefly summarized as shown in Table I.

A. Image Manipulation Detection and Localization

Image manipulation detection and localization is one of the research topics in digital media forensics. Typical image manipulation techniques based on content changes can be divided into three categories: copy-move, splicing, and removal. Most of the early work focused on detecting specific types of manipulation [17]–[19]. Since manipulation type is unknown in the real world, it is important to develop generic models for multiple manipulation types.

Recently, some works that can detect multiple manipulation types have been proposed. ManTra-Net [20] formulated the forgery localization problem as a local anomaly detection problem and can detect multiple manipulation types. SPAN [21] utilized a pyramid of local self-attention blocks to model spatial relationship among image patches at multiple scales. ManTra-Net and SPAN have limitations in generalization because of failing to making full use of the spatial correlation. Yang et al. [34] constructed a constrained region-CNN, which utilized an attention region proposal network for coarse localization and fused low-level and high-level information to obtain refined tampered regions. Zhuang et al. [30] proposed a dense fully convolutional network, where dense connections and dilated convolutions were adopted for achieving better localization performance. Zeng et al. [35] proposed a proxy proposal contrastive learning task to exploit relationships of local features and obtain more generalizable features. Most of the above methods utilize the idea of segmentation to deal with the problem of image manipulation detection. However, Zhang et al. [36] first implemented image manipulation detection from a regression perspective and proposed a regression network based on CatmullRom splines. In addition, some works [37], [38] utilize uncertainty learning to improve the generalization of image manipulation detection. Ji et al. [37] proposed an uncertainty estimation network to measure the uncertainty of data and model and improve the results of manipulation localization. Xu et al. [38] designed an uncertainty supervised parallel network to avoid semantic noise while preserving more manipulation details. To fully employ the noise inconsistency between forged and real regions, Zhu et al. [39] proposed a two-step discriminative noise-guided method to enhance the noise inconsistency and improve the accuracy and robustness of forgery detection. Huang et al. [40] employed an RGB stream to extract both high-level and low-level tampering traces for coarse localization, and adopted a noise stream to expose local noise inconsistency for fine localization. In this work, the proposed DAE-Net adopts a two-branch coding structure to extract semantic features and noise features at multiple scales. Different from existing methods that directly fuse features from different domains by summation or concatenation, which makes it difficult to achieve effective feature fusion, the proposed dual attention mechanism module is employed to fuse semantic features and noise features by utilizing the self-attention mechanism in the spatial and channel dimensions, respectively. It can highlight the differences between tampered and real regions and obtain more discriminative feature representations.

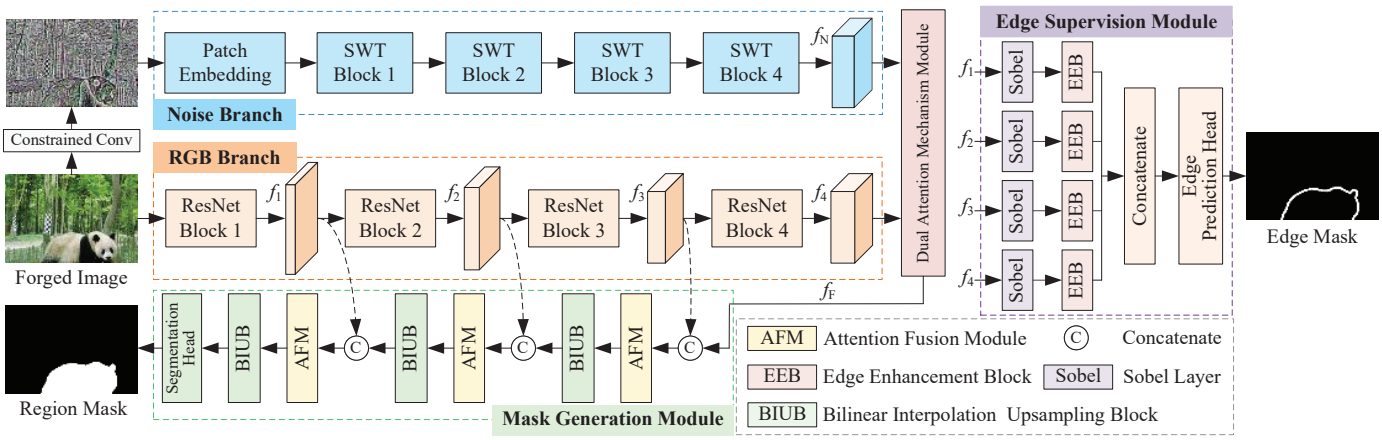


Fig. 2. Overview of the proposed DAE-Net for image manipulation localization. The input is a suspicious image ($H \times W \times 3$) and the output is a predicted mask ($H \times W \times 1$), which localizes the tampered regions. DAE-Net consists of five components: RGB branch, noise branch, dual attention mechanism module, mask generation module, and edge supervision module.

B. Attention Mechanism

Attention mechanism [18], [22] has been applied to IMDL to further improve the feature representation with relatively low cost, among which spatial attention [41], channel attention [42], and self-attention [31] are more widely implemented. Liu et al. [41] proposed an attention-aware hierarchical-feature fusion module, which adopted position attention to fuse hierarchical features from two different domains. Ganapathi et al. [42] utilized channel attention to extract attention-aware multiresolution features in the spatial and frequency domains to increase generalization on unseen manipulations. Li et al. [33] proposed a spatial-channel fusion excitation module to enhance the model's differentiation capability. A mutual attentive module was proposed to distill complementary signals from the foreground and the background features in [43]. Xu et al. [18] designed semantic-agnostic manipulation attention for mitigating the effect of rich image semantics and devised the subtractive operation to promote the network to learn general features. Zhuo et al. [31] utilized a self-attention mechanism to obtain contextual dependencies of intrinsic inconsistency in tampered regions. Wang et al. [44] employed the self-attention mechanism in the transformer architecture to produce refined global representations for IMDL. To realize the fusion of features from different domains, Xia et al. [32] adopted a hierarchical feature fusion strategy, which utilized the convolutional block attention module and self-attention to fuse RGB and noise features hierarchically. Inspired by the above attention mechanism, we design a dual attention mechanism module to fuse features from different domains. Meanwhile, different from the decoding manner of existing methods, an attention fusion module is designed in the mask generation process to fuse rich semantic information of different scales, enhancing the feature representation and improving the accuracy of manipulation localization.

C. Edge Supervision

To solve the problem of coarse and incomplete boundaries in current IMDL methods, recent works [3], [24] have introduced an edge supervision strategy to optimize the boundary of tampered regions. Lin et al. [28] enhanced fused tampering traces using the proposed edge artifact enhancement module and edge supervision strategy to find subtle edge artifacts hidden in images. Li et al. [3] applied morphological operations to extract multi-scale edge information, which can reduce feature redundancy by producing higher resolution edges. Gao et al. [45] designed a multi-scale edge-guided attention stream to obtain robust boundary information through multi-scale Sobel layer. Subsequently, the edge attention module accurately located the tampered regions based on the multi-scale edge features. Li et al. [46] proposed the threshold-adaptive differentiable binarization edge algorithm to provide the learnable edges. Zeng et al. [47] proposed a dual-pathway multi-scale network, which was mainly used for edge information extraction and adopted a dual-path structure to align visual features in RGB space and edge features in LAB space. MVSS-Net [24] introduced a Sobel layer to identify candidate edge pixels and re-weighted the edge features by attention maps for accurate edge localization. Therefore, inspired by the edge supervision strategy of MVSS-Net, we introduce an edge supervision module to optimize the boundary of tampered regions. In contrast to existing methodologies, we exploit the Sobel layer to identify the edge pixels at different scales. Then the edge enhancement block is designed to enhance the edge regions and retain more detailed information.

III. Proposed Method

Our goals are to localize tampered regions and improve localization accuracy by modeling visual artifact features at multiple scales and modalities. In this section, the network structure of DAE-Net is described in detail. The DAE-Net consists of a two-branch structure: the RGB branch and the noise branch, which extract artifact features of different modalities. Moreover, we introduce a dual attention mechanism module for feature fusion and a mask generation module for generating tampered mask. Finally, an edge supervision module is designed to optimize the boundary of tampered

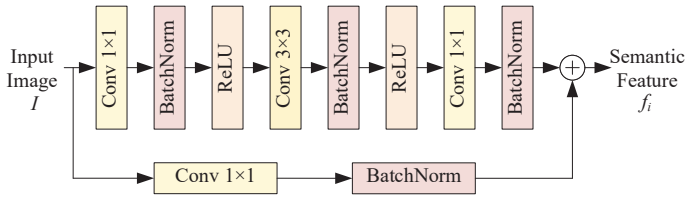


Fig. 3. Illustration of the basic structure of ResNet block.

regions. The details of the main network components are described below.

A. Overview

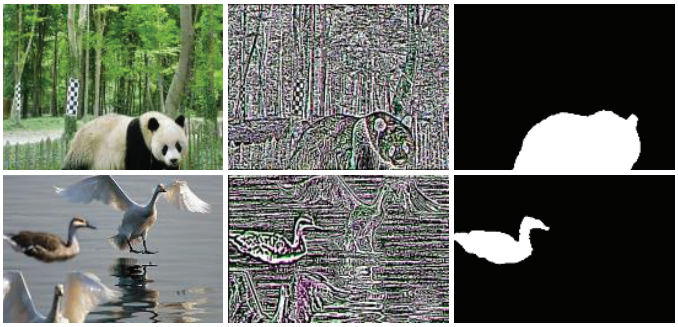
The overview of the proposed DAE-Net is shown in Fig. 2. An input image is denoted as $I \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the image, respectively. Our goal is to localize the tampered mask $M \in \mathbb{R}^{H\times W\times 1}$. As shown in Fig. 2, the whole network structure consists of five components: RGB branch, noise branch, dual attention mechanism module, mask generation module, and edge supervision module. We first employ the RGB branch and the noise branch to explore multimodal artifact features aiding the localization of tampered regions. The RGB branch is utilized to extract multi-scale semantic features. The noise branch is employed to capture the fused global and local inconsistent noise features. Moreover, the dual attention mechanism module is designed to fuse the semantic features and noise features to mine the complementary information between features and generate more generalized feature representations. Subsequently, the mask generation module is exploited to realize the localization of the tampered regions. During the mask generation process, the accuracy of the manipulation localization is improved by the fusion upsampling module to make full use of semantic features at different scales. Finally, to optimize the boundary of tampered regions, the edge supervision module is proposed to enhance the subtle edge features and guide the network to predict more refined tampered regions.

B. RGB Branch

The RGB branch aims to extract semantic features at multiple scales to provide better feature representations for image manipulation localization. The RGB branch is composed of four ResNet blocks in series, where the basic structure of each ResNet block is shown in Fig. 3. Each ResNet block contains a varying number of basic structures. Referring to the architecture of ResNet 50 [48], we set the number of basic structures in each block to be: 3, 4, 6, and 4, respectively. The CNN-based backbone is used to mine the local artifact features in the RGB domain of the input image. The input of the RGB branch is the forged image, denoted as $I \in \mathbb{R}^{H\times W\times 3}$. First, we perform the preprocessing operation to transform the input image via a 7×7 convolutional layer with a stride of 2 and 3×3 maximum pooling layer into the feature map $I' \in \mathbb{R}^{(H/4)\times(W/4)\times 64}$. To obtain multi-scale features containing rich semantic content and spatial information for assisting the final tampered mask prediction, the preprocessed image is passed through four RseNet blocks in series to extract semantic features at four scales: $f_1 \in \mathbb{R}^{(H/4)\times(W/4)\times 256}$, $f_2 \in \mathbb{R}^{(H/8)\times(W/8)\times 512}$, $f_3 \in \mathbb{R}^{(H/16)\times(W/16)\times 1024}$, and $f_4 \in \mathbb{R}^{(H/16)\times(W/16)\times 2048}$. The above process is expressed as Eq. (1):

$$\{f_i, i = 1, 2, 3, 4\} \leftarrow \mathrm{ResNet}\{\mathrm{Maxpool}[\mathrm{Conv}_{7\times 7}(I)]\}, \quad (1)$$

where $\mathrm{Conv}_{7\times 7}$ denotes the convolution operation with filter size 7×7 and $\mathrm{Maxpool}$ denotes the maximum pooling operation.



(a) (b) (c)

Fig. 4. Visualization results of extracted noise map by the constrained convolution layer: (a) Original forged image, (b) Noise map, and (c) Ground-truth binary mask.

C. Noise Branch

The RGB branch may ignore the invisible visual artifacts in the RGB domain when performing tampered features extraction, thereby reducing the generalization of manipulation localization. To improve the generalization ability and capture the invisible artifact features in the RGB domain, we introduce the noise branch to capture the noise features of the tampered images. Inspired by the swin-transformer [29] that can extract global and local features, the noise branch adopts the swin-transformer architecture as the backbone to extract noise features that fuse global and local inconsistencies. The noise branch consists of noise map extraction block, patch embedding, and swin-transformer (SWT) block.

The noise map extraction block, which consists of constrained convolutional layer [27], [34], adaptively converts the image into noise maps. The constrained convolutional layer is implemented as a set of trainable filter kernels with constraints that enable the filter kernels to resemble high-pass filters. The extracted noise maps are shown in Fig. 4. The constraints on the filter kernels are expressed as Eq. (2):

$$\begin{cases} w_k(c,c) = -1, \\ \sum_{m,n\neq c} w_k(m,n) = 1, \end{cases} \quad (2)$$

where w_k is the weight of k-th filter kernel, (c, c) is the center position coordinate of the filter kernel, and (m, n) is the non-center position coordinate of the filter kernel.

The noise maps are fed into the network with swin-transformer as the backbone to extract the noise features that fuse the global and local inconsistencies. First, the patch partition module is utilized to divide the noise maps into

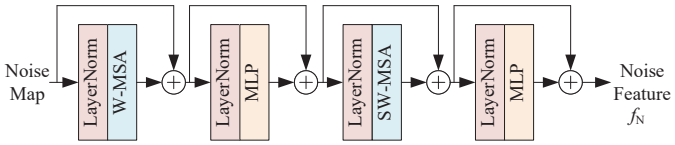


Fig. 5. Illustration of the basic structure of the SWT block.

non-overlapping patches of size $N_p \times N_p$. In this study, N_p is set to 4. Then a linear embedding layer is exploited to map the dimension of noise features to C, where the size of the noise features is $(H/4) \times (W/4) \times C$. Finally, four SWT blocks in series are used to extract the fused global and local inconsistent noise features from patches. The basic structure of the SWT block is shown in Fig. 5, which is made up of layernorm, window-based multi-head self-attention (W-MSA), multi-layer perceptron (MLP), and shifted window-based MSA (SW-MSA). Each SWT block consists of a varying number of SWT basic structures and patch merging. The SWT block employs the shifted window self-attention mechanism to extract global and local noise inconsistencies and reduce the computational complexity. The patch merging performs a downsampling operation on the features to reduce the feature resolution. The self-attention matrix operation is formulated as Eq. (3):

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^{\text{T}}}{\sqrt{d}} + B)V, \quad (3)$$

where Q, K, and V are the query, key, and value obtained after the liner projection of the patches, respectively; B is the relative position bias; and d is the dimension of Q and K.

The noise patches are fed into four successive SWT blocks to obtain features $f_N \in \mathbb{R}^{(H/32)\times(W/32)\times 8C}$, which fuse global and local inconsistent information. The above process is expressed as:

$$f_N \leftarrow \text{SWT}\{\text{PE}[\text{NEB}(I)]\}, \quad (4)$$

where NEB, PE, and SWT denote the noise extraction block, patch embedding, and successive swin-transformer blocks, respectively.

D. Dual Attention Mechanism Module

To fully mine the complementary relationship between semantic features and noise features, we introduce a dual attention mechanism (DAM) module [49], which leverages the self-attention mechanism in the spatial and channel dimensions to achieve effective fusion of different domain features. It solves the problem that previous method [50] simply concatenates only in the channel dimension and cannot discover the complementary relationship of different domains. For the two modality features, semantic feature f_4 and noise feature f_N, the noise feature is upsampled twice and concatenated with the semantic feature in the channel dimension. The input features $f_I \in \mathbb{R}^{(H/16)\times(W/16)\times 512}$ of the DAM module are obtained after a 1×1 convolutional layer.

The DAM module consists of two parallel branches: the spatial attention branch and the channel attention branch. The

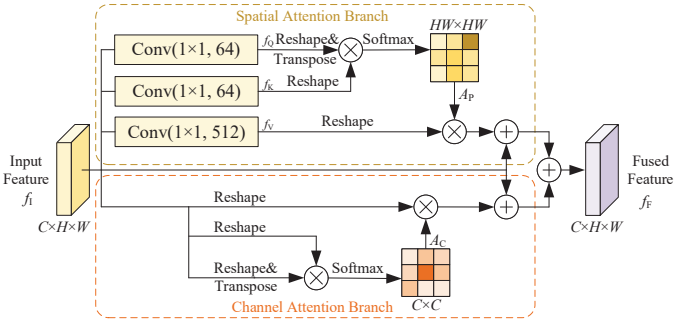


Fig. 6. The details of dual attention mechanism module, which contains spatial attention branch and channel attention branch.

DAM module structure is shown in Fig. 6. In the spatial attention branch, long-range spatial context information is captured by applying the self-attention mechanism in the spatial dimension. The input features are passed through three 1×1 convolutional layers to obtain query features $f_Q \in \mathbb{R}^{(H/16)\times(W/16)\times 64}$, key features $f_K \in \mathbb{R}^{(H/16)\times(W/16)\times 64}$, and value features $f_V \in \mathbb{R}^{(H/16)\times(W/16)\times 512}$, where the query features and key features are downscaled in channel dimension to reduce the model computation. The f_Q and f_V are reshaped to $f_Q \in \mathbb{R}^{N\times 64}$ and $f_V \in \mathbb{R}^{N\times 512}$, where $N = (H/16) \times (W/16)$. The f_K is reshaped and transposed to obtain $f_K \in \mathbb{R}^{64\times N}$. In the spatial attention branch, spatial attention matrix A_P of the feature maps f_I is computed as:

$$A_P^{ij} = \frac{\exp(f_Q^i \times f_K^j)}{\sum_{i=1}^{N} \exp(f_Q^i \times f_K^j)}, \quad (5)$$

where A_P^{ij} is the value of the i-th row and j-th column of the spatial attention matrix A_P, N is the total number of pixels. Spatial attention feature maps f_P are computed as:

$$f_P = f_V \times A_P + f_I, \quad (6)$$

where f_I denotes the input feature maps, f_V denotes the feature maps after 1×1 convolutional layer.

Similarly, in the channel attention branch, we reshape the input features to $f_R \in \mathbb{R}^{N\times 512}$ and transpose f_R to $f_T \in \mathbb{R}^{512\times N}$. Channel attention matrix A_C of the feature maps f_I is computed as:

$$A_C^{ij} = \frac{\exp(f_T^i \times f_R^j)}{\sum_{i=1}^{C} \exp(f_T^i \times f_R^j)}, \quad (7)$$

where A_C^{ij} is the value of the i-th row and j-th column of the channel attention matrix A_C, C is the number of channels. Channel attention feature maps f_C are computed as:

$$f_C = f_R \times A_C + f_I. \quad (8)$$

Finally, the outputs of the two branches are aggregated to enrich the representation of artifact features and to highlight the differences between the tampered and authentic regions. The spatial attention features f_P are summed with channel attention features f_C to obtain fusion features f_F, which is expressed as Eq. (9):

$$f_F = f_P + f_C. \quad (9)$$

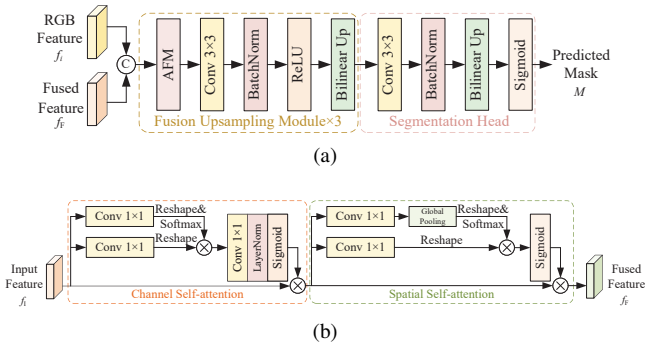


Fig. 7. Overall structure of mask generation module used in our proposed fusion upsampling module for tampered regions detection: (a) mask generation module and (b) attention fusion module.

E. Mask Generation Module

To solve the problem that previous methods [20], [24] cannot fully exploit the shallow features containing rich semantic information in the process of generating tampered mask, resulting in inaccurate prediction of tampered regions. As shown in Fig. 2, we adopt a progressive decoding manner to fully integrate low-level detailed features and high-level semantic features to improve the accuracy of manipulation localization. The mask generation module is composed of three FUMs and a segmentation head, as shown in Fig. 7 (a). The FUM introduces an attention fusion module (AFM) to fuse multi-scale semantic features, as shown in Fig. 7 (b).

For the given fused feature f_{F} and multi-scale semantic feature $\{f_i, i = 1, 2, 3\}$, they are first concatenated in the channel dimension to obtain feature f_{S}. Then it is fed into the AFM to achieve effective feature fusion and enhance representation of the tampered features. Inspired by the literature [51], we sequentially utilize self-attention in channel and spatial dimensions to fuse different features. The channel self-attention feature $f_{\mathrm{S}}^{\mathrm{C}}$ is computed as:

$$\begin{cases} S_{\mathrm{C}} = \mathrm{Softmax}[\mathrm{Conv}_{1\times1}(f_{\mathrm{S}})] \times \mathrm{Conv}_{1\times1}(f_{\mathrm{S}}), \\ f_{\mathrm{S}}^{\mathrm{C}} = \sigma\{\mathrm{LN}[\mathrm{Conv}_{1\times1}(S_{\mathrm{C}})]\} \otimes f_{\mathrm{S}}, \end{cases} \tag{10}$$

where $\mathrm{Conv}_{1\times1}$ denotes the convolution operation with filter size 1×1, S_{C} is channel self-attention, LN denotes LayerNorm, σ is the Sigmoid function, and $\otimes$ denotes element-wise multiplication.

The channel self-attention feature $f_{\mathrm{S}}^{\mathrm{C}}$ is used as input to compute the spatial self-attention feature $f_{\mathrm{S}}^{\mathrm{P}}$, which is formulated as:

$$\begin{cases} S_{\mathrm{P}} = \mathrm{Softmax}\{\mathrm{GP}[\mathrm{Conv}_{1\times1}(f_{\mathrm{S}}^{\mathrm{C}})]\} \times \mathrm{Conv}_{1\times1}(f_{\mathrm{S}}^{\mathrm{C}}), \\ f_{\mathrm{S}}^{\mathrm{P}} = \sigma(S_{\mathrm{P}}) \otimes f_{\mathrm{S}}^{\mathrm{C}}, \end{cases} \tag{11}$$

where GP is global pooling operation and S_{P} is spatial self-attention.

The fused features are then passed through a 3×3 convolution block and upsampled to twice the original size by bilinear interpolation, thus completing a fusion upsampling operation. The above process is expressed as Eq. (12):

$$f_{\mathrm{F}}^{1} = \mathrm{BU}\{\mathrm{ReLU}[\mathrm{BN}(\mathrm{Conv}_{3\times3}(f_{\mathrm{S}}^{\mathrm{P}}))]\}, \tag{12}$$

where $\mathrm{Conv}_{3\times3}$, BN, ReLU, and BU denote the convolution operation with filter size 3×3, a batch normalization layer, a ReLU activation function, and bilinear interpolation upsample, respectively.

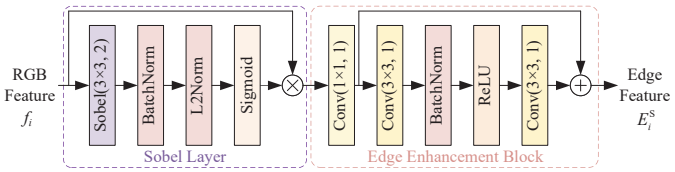


Fig. 8. Illustration of edge supervision module, which composed of the Sobel layer and edge enhancement block for tampered edge detection.

To fuse the three scales of semantic features $\{f_i, i = 1, 2, 3\}$, we adopt a progressive decoding manner and perform three fusion upsampling operations to obtain fused features $\{f_{\mathrm{F}}^{i}, i = 1, 2, 3\}$ at different scales, as shown in Fig. 2. The obtained fused feature f_{F}^{3} are input into the segmentation head to predict the final tampered mask $M \in \mathbb{R}^{H\times W\times1}$. This process can be represented as Eq. (13):

$$M = \sigma\{\mathrm{BU}[\mathrm{BN}(\mathrm{Conv}_{3\times3}(f_{\mathrm{F}}^{3}))]\}, \tag{13}$$

where bilinear interpolation upsampling multiple is set to 2.

F. Edge Supervision Module

To address the problem of rough and incomplete boundary for tampered region localization, existing methods [3], [24], [28] extract edge features to optimize the boundary of tampered regions. Inspired by the literature [24], the edge supervision module is designed to assist in the localization of tampered region. The edge supervision module consists of a Sobel layer and an edge enhancement block, as shown in Fig. 8. The edge supervision module takes the multi-scale semantic features extracted by the RGB branch as inputs, which contains four different scales of inputs in total. Therefore, the edge supervision module is designed as four parallel Sobel branches to extract the edge information of the four scales of semantic features. The edge strategy is implemented to optimize the boundary of tampered regions.

For the given multi-scale semantic feature $\{f_i, i = 1, 2, 3, 4\}$, we first extract the edge features with the Sobel convolution layer, which is initialized using the Sobel operator. The multi-scale semantic features are fed into the Sobel convolution layer to identify the edge pixels. Then the attention maps of the edge pixels are obtained after batch normalization layer, L2Norm, and sigmoid function. After element-wise multiplication of the attention maps with the input feature maps, multi-scale edge features are obtained. This process is expressed as Eq. (14):

$$E_i = \sigma\{\mathrm{L}_2[\mathrm{BN}(\mathrm{Sobel}(f_i))]\} \otimes f_i, \tag{14}$$

where $\{E_i, i = 1, 2, 3, 4\}$ are edge features; L_2 and Sobel denote the L2Norm layer and Sobel convolution layer, respectively.

Subsequently, the edge enhancement module is exploited to enhance the edge features E_i so that the extracted edge features retain more edge detail information. Specifically, a 1×1 convolution layer is first used to change the number of

TABLE II
TRAINING-TESTING SPLIT AND DESCRIPTION OF STANDARD DATASETS. ✔ AND ✗ INDICATE WHETHER OR NOT THE MANIPULATION TYPE IS INVOLVED.

Dataset	Training-Testing Split		Involved Manipulation Type			Is Processed	Total
	Training	Testing	Copy-move	Splicing	Removal		
CASIA [52]	5123	921	✔	✔	✗	Yes	6044
NIST16 [53]	404	160	✔	✔	✔	Yes	564
Columbia [54]	130	50	✗	✔	✗	No	180
Coverage [55]	75	25	✔	✗	✗	Yes	100
IMD2020 [56]	1610	400	✔	✔	✔	Yes	2010

channels. Then the edge features are enhanced by two 3×3 convolution layers. The enhanced edge features are represented as Eq. (15):

$$E_i^{\mathrm{S}} = \mathrm{Conv}_{3\times3}\{\mathrm{ReLU}[\mathrm{BN}(\mathrm{Conv}_{3\times3}(E_i))]\} + E_i, \quad (15)$$

where $\{E_i^{\mathrm{S}}, i = 1, 2, 3, 4\}$ denote enhanced edge features.

Finally, the four enhanced edge features are upsampled to the same size as the tampered image and are concatenated in the channel dimension to fuse the different scales of edge information. The edge mask is predicted with the edge prediction head. This process is represented as:

$$M_{\mathrm{E}} = \mathrm{Conv}_{1\times1}[\mathrm{Cat}(E_i^{\mathrm{S}})|_{i=1}^{4}], \quad (16)$$

where $\mathrm{Cat}(\cdot)$ denotes feature concatenation along the channel dimension. The network performs supervised training by calculating the loss between edge prediction mask and authentic edge mask to optimize the boundary of tampered regions.

G. Loss Function

During the end-to-end training of the network, we consider a multi-task supervision strategy, i.e., tampered mask supervision and edge supervision. The objective of the tampered mask supervision is to enhance the model's sensitivity to the boundary localization. Consequently, two loss functions were established to optimize the model's parameters: the tampered region loss L_{M} and the edge loss L_{E}.

Tampered Region Loss. The prediction of the tampered region can be viewed as a binary classification of pixels. Thus, the tampered region loss is generally computed using binary cross entropy L_{B}, which is denoted as:

$$L_{\mathrm{B}} = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log_2(\hat{y}_i) + (1-y_i)\log_2(1-\hat{y}_i)], \quad (17)$$

where m is the number of pixels, y_i and $\hat{y}_i$ are ground-truth and prediction of the i-th pixel, respectively. However, in the image tampering, the size of the tampered region is uncertain. Sometimes the tampered region is only a small part of the whole image, which suffers from the problem of sample unbalance. Inspired by the literature [57], the dice loss L_{D} is introduced into the tampered region loss for optimization. The dice loss ignores a large number of background pixels in the calculation, so the problem of sample unbalance is well solved. The dice loss L_{D} can be expressed as:

$$L_{\mathrm{D}} = 1 - \frac{2\sum_{i=1}^{m} y_i \hat{y}_i}{\sum_{i=1}^{m} {y_i}^2 + \sum_{i=1}^{m} \hat{y}_i^2}. \quad (18)$$

The tampered region loss L_{M} is the loss between the ground-truth mask and the predicted mask and is expressed as:

$$L_{\mathrm{M}} = \lambda L_{\mathrm{B}} + \mu L_{\mathrm{D}}, \quad (19)$$

where L_{B} and L_{D} are the binary cross entropy loss and dice loss between the ground-truth mask and the predicted mask, respectively; λ and μ are two hyper-parameters in the range of [0,1].

Edge Loss. The edge loss is the loss between the ground-truth boundary and the predicted boundary. Since the edge pixels of the tampered region are typically in the minority, we apply the dice loss which effectively copes with unbalanced positive and negative samples. The edge loss L_{E} is expressed as:

$$L_{\mathrm{E}} = 1 - \frac{2\sum_{i=1}^{m} e_i \hat{e}_i}{\sum_{i=1}^{m} {e_i}^2 + \sum_{i=1}^{m} \hat{e}_i^2}, \quad (20)$$

where e_i and $\hat{e}_i$ refer to the value of the i-th pixel in ground truth boundary and prediction boundary, respectively.

Overall, the total loss L_{T} for our model can be defined by combining the tampered region loss and the edge loss and is expressed as:

$$L_{\mathrm{T}} = \gamma_{\mathrm{M}} L_{\mathrm{M}} + \gamma_{\mathrm{E}} L_{\mathrm{E}}, \quad (21)$$

where γ_{M} and γ_{E} are the weights of the tampered region loss and edge loss, respectively.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we evaluate our model quantitatively and qualitatively. The experimental setup is introduced at Section IV-A and the comparison results with the SoTA methods are presented in Section IV-B. Subsequently, we perform ablation study to verify the effectiveness of different components and losses in Section IV-C. Furthermore, robustness study is performed to evaluate the robustness of the model against different post-processing attacks and specific conditions in Section IV-D. Finally, we show the visualization results in Section IV-E and compare the computational complexity of different methods in Section IV-F.

A. Experimental Setup

1) Datasets: The experimental images are mainly from five standard datasets, CASIA [52], NIST16 [53], Columbia [54], Coverage [55], and IMD2020 [56]. For a fair comparison, we adopt the most popular training-testing splitting configurations

TABLE III
QUANTITATIVE COMPARISON OF DAE-NET WITH THE NINE SOTA METHODS ON FIVE PUBLIC DATASETS: CASIA, NIST16, COLUMBIA, COVERAGE, AND IMD2020. "-" INDICATES THAT THE RESULTS ARE NOT PROVIDED IN THE LITERATURE.

Method	Category	CASIA		NIST16		Columbia		Coverage		IMD2020		Average	
		F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
CFA1 [9]	Unsupervised	0.207	0.522	0.174	0.501	0.467	0.720	0.190	0.485	-	-	-	-
NOI1 [10]	Unsupervised	0.263	0.612	0.285	0.487	0.574	0.546	0.269	0.587	-	-	-	-
ManTra-Net [20]	DNN-based	0.223	0.648	0.462	0.795	-	0.824	0.283	0.777	0.265	0.785	-	0.799
SPAN [22]	DNN-based	0.213	0.709	0.582	0.961	0.815	0.936	0.325	0.791	-	-	-	-
DenseFCN [30]	DNN-based	0.209	0.631	0.704	0.954	0.710	0.881	0.185	0.754	0.286	0.723	0.419	0.789
SATFL [31]	DNN-based	0.246	0.697	0.613	0.937	0.804	0.892	**0.347**	0.767	0.300	0.796	0.462	0.818
MVSS-Net [24]	DNN-based	0.390	0.748	0.827	0.981	0.703	0.719	0.284	**0.808**	**0.411**	0.817	0.523	0.815
DMFF-Net [32]	DNN-based	0.386	0.791	0.843	0.976	0.837	0.945	0.282	0.727	0.328	0.784	0.535	0.845
SF-Net [33]	DNN-based	0.464	**0.921**	**0.911**	0.928	0.805	0.948	0.311	0.737	0.344	0.768	0.567	0.860
DAE-Net (ours)	DNN-based	**0.494**	0.805	0.871	**0.984**	**0.872**	**0.973**	0.330	0.741	0.338	**0.826**	**0.581**	**0.866**

on these five datasets, as in [3], [22], [44]. The details of these datasets are presented in Table II. We will briefly introduce these datasets as follows.

- **CASIA** [52]: CASIA contains two versions, i.e. CASIA V1.0 and CASIA V2.0, where CASIA V1.0 has 920 tampered images and CASIA V2.0 has 5123 tampered images. CASIA provides spliced and copy-moved images of various objects. We follow the methods in [3], [44] and used CASIA V2.0 as training set and CASIA V1.0 as testing set.
- **NIST16** [53]: NIST16 is a challenging dataset in which we perform extensive experiments. NIST16 includes 564 tampered images covering three manipulation types: copy-move, splicing, and removal. The ground-truth masks are provided.
- **Columbia** [54]: Columbia contains only 180 spliced images. We follow the method in [3] and randomly use 130 images as training set and the remaining 50 images as testing set.
- **Coverage** [55]: Coverage provides 100 images generated by copy-move techniques and also includes ground-truth masks.
- **IMD2020** [56]: IMD2020 includes 2010 real-life tampered images collated from Internet, and includes all three manipulation types.

2) Compared Methods: We compare DAE-Net with state-of-the-art methods, which can be divided into two categories: unsupervised methods e.g., CAF1 [9], NOI1 [10], and DNN-based methods, e.g., ManTra-Net [20], SPAN [22], DenseFCN [30], SATFL [31], MVSS-Net [24], DMFF-Net [32], and SF-Net [33]. These methods are briefly described below:

- **CFA1** [9]: CFA1 performs image forgery localization by analyzing CFA artifacts.
- **NOI1** [10]: NOI1 leverages noise inconsistency for manipulation detection.
- **ManTra-Net** [20]: ManTra-Net detects and localizes tampered regions by a local anomaly detection network.
- **SPAN** [22]: SPAN utilizes a pyramid structure and models the relationship among image blocks by local self-attention blocks.
- **DenseFCN** [30]: DenseFCN leverages a fully convolutional encoder-decoder structure with dense connections and dilated convolution for achieving better localization performance.
- **SATFL** [31]: SATFL employs self-attention mechanism to localize tampered regions, where self-attention module is based on a channel-wise high pass filter block.
- **MVSS-Net** [24]: MVSS-Net jointly utilizes multi-view input, boundary artifacts and the holistic information in an end-to-end manner to extract semantic-agnostic and generalizable features.
- **DMFF-Net** [32]: DMFF-Net uses a graded feature fusion strategy and a multi-supervision strategy to achieve accurate tampered region localization.
- **SF-Net** [33]: SF-Net utilizes a feature enhancement module to enhance the fine-grained features in RGB stream and adpots a fusion excitation strategy to effectively fuse spatial and channel features.

3) Implementation Details: The proposed model is implemented based on Pytorch and is trained over the platform with two NVIDIA GeForce RTX 2080Ti GPUs. Due to the memory limitations of the computing devices, all images are resized to 512×512. We use Adam for optimization with a learning rate of 1e-5 and a weight decay of 5e-5. We train the whole model for 60 epochs with a batch size of 8. In our tampered region loss function, the hyperparameters λ and μ are set to 0.5. In the total loss function, the tampered region loss weight γ_{M} and the edge loss weight γ_{E} are set to 0.8 and 0.2, respectively.

4) Evaluation Metrics: We evaluate the performance of the proposed method for image manipulation localization at the pixel level. Following previous works [2], [3], [24], pixel-level F_1 score and area under the receiver operating curve (AUC) are adopted to evaluate the performance of binary classification for every pixel. Both F_1 score and AUC are in the range of [0,1]. Larger values of them indicate better localization performance of the model. F_1 score is calculated as:

$$F_1 = \frac{2pr}{p+r}, \tag{22}$$

where p and r represent precision and recall, respectively. The F_1 is harmonic average of the precision and recall. The F_1 and

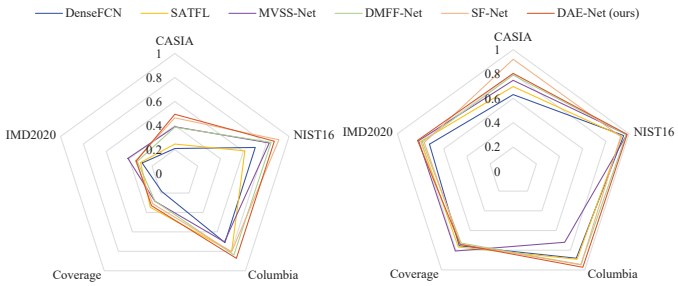


Fig. 9. Comparison results visualization with radar charts of the proposed method with five SoTA methods on five public datasets: (a) The results of F_1 and (b) The results of AUC.

AUC are calculated on each image, respectively. The mean value of each metric for all images in a dataset is used to make comparisons in the following experiments.

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

We compare and analyze DAE-Net with the state-of-the-art image manipulation localization methods on five benchmark datasets. The compared IMDL methods are divided into two categories: unsupervised methods, e.g., CAF1 [9], NOI1 [10], and DNN-based methods, e.g., ManTra-Net [20], SPAN [22], DenseFCN [30], SATFL [31], MVSS-Net [24], DMFF-Net [32], and SF-Net [33]. For a fair and reproducible comparison, the networks are trained using the same training strategy for methods with accessible source code. Their performance is then evaluated on five benchmark datasets. If the source code is not available, the reported results are from the relevant references. In particular, the reported results of ManTra-Net [20], DenseFCN [30], MVSS-Net [24], DMFF-Net [32], and SF-Net [33] are obtained by retaining them with their public codes on our training samples. The reported results of SPAN [22] and SATFL [31] are obtained by considering the corresponding and related literatures [3], [28]. We analyze the comparison results from both quantitative and qualitative perspectives. The generalization of the proposed model is discussed.

1) Quantitative Comparison: For a fair comparison between proposed method and nine SoTA methods, we utilize F_1 and AUC to measure the difference between the ground-truth mask and the predicted mask. Table III and Fig. 9 show quantitative comparison of the F_1 and AUC on five benchmark datasets, with the best performance bolded. It can be seen that DNN-based methods show better performance compared to unsupervised methods. It is observed that the F_1 of the DAE-Net achieve optimal results on CASIA and Columbia, and suboptimal results on the remaining three datasets. The values of AUC achieve optimal results on NIST16, Columbia, and IMD2020. Moreover, the DAE-Net achieves the best results for the average of both F_1 and AUC on all five datasets, which are 0.581 and 0.866, respectively. On the CASIA dataset, the F_1 of the proposed method is 0.271, 0.281, 0.285, 0.248, 0.104, 0.108, and 0.030 higher than that of ManTra-Net [20], SPAN [22], DenseFCN [30], SATFL [31], MVSS-Net [24], DMFF-Net [32], and SF-Net [33], respectively. This indicates that our method effectively improves the performance of manipulation localization by extracting noise and edge features, which ensures the generalization ability of the model. Our method obtains performance improvements of the F_1 with gains of 4.4% on NIST16, 16.9% on Columbia, and 4.6% on Coverage when compared with MVSS-Net [24]. For the AUC measurement, DAE-Net achieves 82.6% on the real-world dataset IMD2020, and outperforms MVSS-Net and DMFF-Net by 0.9% and 4.2%, respectively. This suggests that our method adopts a progressive decoding manner to make full use of semantic features containing rich semantic information, which effectively improves the capability of manipulation localization and can generalize well to real-world dataset.

2) Qualitative Comparison: In this section, we present qualitative results by comparing DAE-Net and five most competitive method: ManTra-Net [20], DenseFCN [30], MVSS-Net [24], DMFF-Net [32], and SF-Net [33]. Since the NIST16 dataset contains three manipulation types, we selected the NIST16 dataset for manipulation localization visualization. Fig. 10 shows qualitative visualization of manipulation localization for DAE-Net with five methods on the NIST16 dataset. It is observed that our method can localize tampered regions regardless of the manipulation types. The other five methods produce a large number of false alarms and incomplete predictions, i.e., tampered pixels cannot be localized accurately and completely. Among them, ManTra-Net [20] cannot achieve effective localization of removal manipulation type as it fails to extract the more generalized tampered features. Moreover, ManTra-Net [20] and DenseFCN [30] do not extract the edge features, resulting in inferior boundary localization compared to the proposed method. MVSS-Net [24] suffers from the problem of missing detection of tampered regions because it directly upsamples the extracted high-level tampered features during the mask generation process, making it difficult to effectively recovery the tampered regions. DMFF-Net [32] and SF-Net [33] still suffer from the loss of tampered regions and poor boundary localization. These localization visualization results can further verified that the proposed method could localize more complete tampered regions. The utilization of fully multi-scale features, which contain rich semantic information, in a progressive decoding manner, yields more accurate manipulation localization results.

3) Generalization Ability Evaluation: We further evaluate the generalization ability to unseen datasets. We train the model on the CASIA V2.0 dataset, and perform the evaluation on the other unseen datasets, including CASIA V1.0, NIST16, and Columbia. The experimental results are shown in Table IV. It is observed that there is a degradation in the detection performance on the unseen dataset due to the larger gaps between different datasets. However, compared to the remaining three methods, our model achieves better performance on three datasets, which proves that our model has better generalization ability. This is attributed to the fact that DAE-Net captures and fuses different forgery clues to obtain more generalized feature representation. Our DAE-Net outperforms Mantra-Net [20], DenseFCN [30], and MVSS-Net [24] by 0.245, 0.140, and 0.029 on unseen Columbia dataset in terms of AUC, respectively, which demonstrates the superior performance of proposed model. However, the model's generalization for cross-dataset detection still need to

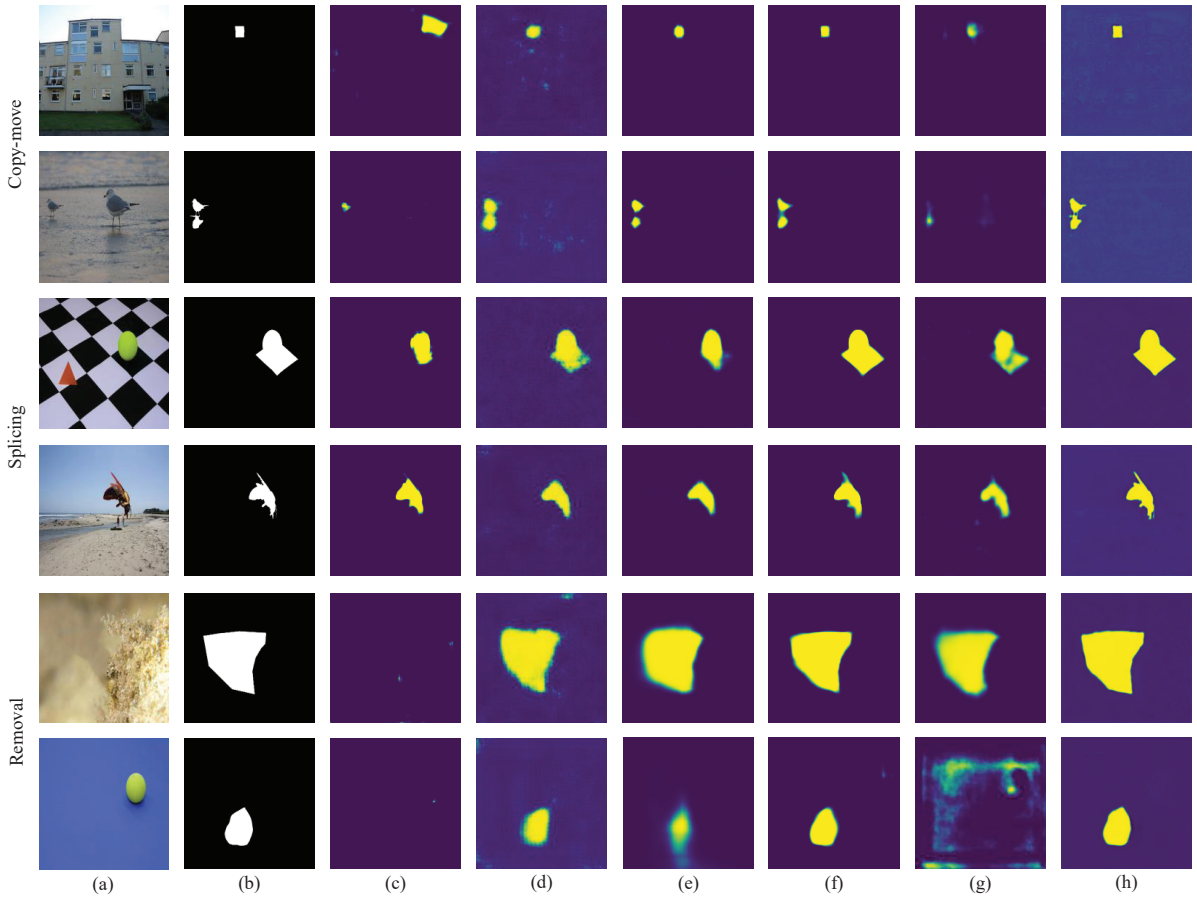


Fig. 10. Qualitative visualization results of the proposed DAE-Net compared with five SoTA methods against copy-move, splicing, and removal manipulations on NIST16 dataset: (a) Original forged image, (b) Ground-truth binary mask, (c) ManTra-Net, (d) DenseFCN, (e) MVSS-Net, (f) DMFF-Net, (g) SF-Net, and (h) DAE-Net (ours).

TABLE IV
GENERALIZATION PERFORMANCE ON UNSEEN DATASETS. F_1 AND AUC ARE REPORTED IN THIS EXPERIMENT.

Method	CASIA		NIST16		Columbia	
	F_1	AUC	F_1	AUC	F_1	AUC
ManTra-Net [20]	0.223	0.648	0.095	0.475	0.386	0.481
DenseFCN [30]	0.209	0.631	0.051	0.604	0.317	0.586
MVSS-Net [24]	0.390	0.748	0.246	0.643	0.471	0.697
DAE-Net (ours)	**0.494**	**0.805**	**0.299**	**0.663**	**0.486**	**0.726**

be further improved.

C. Ablation Study

The proposed method designs the noise branch to capture fused global and local inconsistent noise features and an edge supervision module to refine the boundary for localization of the tampered regions. Meanwhile, a progressive decoding manner is adopted to achieve the localization of the tampered region by fusion upsampling modules. To evaluate the effectiveness of each module, we perform ablation study on noise branch (NB), DAM, edge supervision module (ESM), and mask generation module (MGM). Furthermore, we discuss the effect of different loss functions. All the ablation study in this section is conducted on the NIST16 or Columbia datasets.

1) Effectiveness of Different Components: Firstly, we test the performance with/without NB and the results are shown in Table V. It is observed that the performance of DAE-Net is significantly improved by adding semantic-agnostic noise features. Subsequently, we conduct ablation study on the effectiveness of the MGM module and the results are shown in Table VI. The localization accuracy of model is improved after adding MGM. This is attributed to the fact that MGM utilizes the attention fusion module to fuse multi-scale semantic features in a progressive manner, effectively reducing the false alarm rate of tampered regions. In addition, Table VII demonstrates the effectiveness of DAM. We compare three attention fusion methods such as spatial attention (SA), channel attention (CA), and convolutional block attention module (CBAM) [58]. The DAM outperforms SA, CA, and CBAM in the NIST16 and Columbia datasets. This is because DAM utilizes the global information in spatial and channel dimensions to mine the complementary relationship between semantic features and noise features to obtain discriminative feature representation.

Finally, we construct five model variants, where the baseline model contains only the RGB branch. The remaining four model variants are then constructed by sequentially adding

NB, DAM, ESM, and MGM to the baseline model. All model variants employ the same training settings. The F_1 and AUC for five model variants on the NIST16 dataset are presented in Table VIII. It is observed that the proposed components are effective and significantly improve the F_1 and AUC. Compared with baseline, the noise branch is designed to extract the fused global and local inconsistent noise features in baseline+NB. The F_1 of the baseline+NB is improved by 4.5%. Baseline+NB+DAM+ESM further improves both F_1 and AUC by introducing edge supervision module. It verifies that ESM can extract the subtle edge artifacts to optimize the boundary of tampered region. Finally, we utilize a specially designed mask generation module to localize the tampered region by fusion upsampling module in a progressive decoding fashion, obtaining the optimal performance with 0.871 in F_1 and 0.984 in AUC. In addition, we visualize the localization results for different model variants. The corresponding results are shown in Fig. 11. The noise branch and DAM can basically complete the overall localization of the tampered region, which significantly improves the accuracy of manipulation localization. The ESM effectively assists the localization of the tampered boundary as well as the MGM adopts progressive decoding to solve the problem of insufficient tampered region segmentation.

TABLE V
ABLATION STUDY OF NOISE BRANCH ON NIST16 AND COLUMBIA DATASETS. F_1 AND AUC ARE REPORTED IN THIS EXPERIMENT.

Model Variant	NIST16		Columbia	
	F_1	AUC	F_1	AUC
w/o NB	0.621	0.947	0.792	0.938
w/ NB + w/o Noise Map	0.734	0.963	0.822	0.945
w/ NB + Noise Map	**0.871**	**0.984**	**0.872**	**0.973**

TABLE VI
ABLATION STUDY OF MASK GENERATION MODULE ON NIST16 AND COLUMBIA DATASETS. F_1 AND AUC ARE REPORTED IN THIS EXPERIMENT.

Model Variant	NIST16		Columbia	
	F_1	AUC	F_1	AUC
w/o MGM	0.601	0.969	0.738	0.915
w/ MGM + w/o AFM	0.773	0.979	0.808	0.949
w/ MGM + AFM	**0.871**	**0.984**	**0.872**	**0.973**

TABLE VII
ABLATION STUDY OF DIFFERENT ATTENTION FUSION STRATEGIES ON NIST16 AND COLUMBIA DATASETS. F_1 AND AUC ARE REPORTED IN THIS EXPERIMENT.

Model Variant	NIST16		Columbia	
	F_1	AUC	F_1	AUC
w/o DAM	0.666	0.944	0.773	0.927
w/ SA	0.784	0.955	0.832	0.956
w/ CA	0.741	0.953	0.816	0.942
w/ CBAM [58]	0.802	0.967	0.853	0.964
w/ DAM	**0.871**	**0.984**	**0.872**	**0.973**

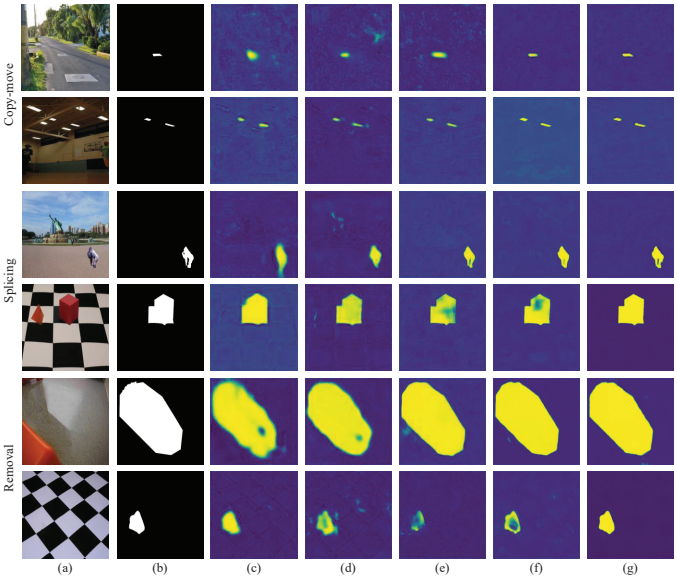


Fig. 11. Qualitative visualization results of different model variants against copy-move, splicing, and removal manipulations on NIST16 dataset: (a) Original forged image, (b) Ground-truth binary mask, (c) Baseline, (d) Baseline+NB, (e) Baseline+NB+DAM, (f) Baseline+NB+DAM+ESM, and (g) DAE-Net.

2) Effectiveness of Different Losses: We discuss the effect of different loss functions and hyperparameters in Eq. (21). Six combinations of loss function hyperparameters, ($\lambda_{\mathrm{M}} = 1, \lambda_{\mathrm{E}} = 0$), ($\lambda_{\mathrm{M}} = 0.8, \lambda_{\mathrm{E}} = 0.2$), ($\lambda_{\mathrm{M}} = 0.6, \lambda_{\mathrm{E}} = 0.4$), ($\lambda_{\mathrm{M}} = 0.5, \lambda_{\mathrm{E}} = 0.5$), ($\lambda_{\mathrm{M}} = 0.4, \lambda_{\mathrm{E}} = 0.6$), ($\lambda_{\mathrm{M}} = 0.2, \lambda_{\mathrm{E}} = 0.8$), are set to observe the experimental results, as shown in Table IX. It is observed that when ($\lambda_{\mathrm{M}} = 0.8, \lambda_{\mathrm{E}} = 0.2$), the model achieves better performance. Therefore, our model choose ($\lambda_{\mathrm{M}} = 0.8, \lambda_{\mathrm{E}} = 0.2$) as the default parameter setting for the experiment. Meanwhile, compared with only binary cross entropy loss L_{B}, adding Dice loss L_{D} to tampered region can further improve the performance with gains of 13% in F_1 and 2.9% in AUC. This result shows that Dice loss can effectively solve the problem of sample imbalance. On this basis, an edge loss L_{E} is introduced to optimize the boundary of tampered regions. By integrating all loss components, the F_1 and AUC achieve the best results.

D. Robustness Study

When the tampered images are delivered on the Internet, they may undergo some image post-processing operations that inevitably weaken or hide the tampering artifacts, which in turn affects the detection performance of the model. In this section, we further investigate the impact of robustness under different image post-processing operations and different sizes of tampered areas. In addition, the performance of the model in specific conditions is discussed.

We apply three image post-processing operations on the NIST16 dataset: Gaussian blur with a kernel size k, JPEG compression with a quality factor q, and Gaussian noise with a variance σ to evaluate the robustness of DAE-Net. We compared the robustness of DAE-Net with ManTra-Net [20], SPAN [22], DenseFCN [30], and MVSS-Net [24] on

TABLE VIII
QUANTITATIVE COMPARISON OF DIFFERENT MODEL VARIANTS ON NIST16 DATASET. ✔INDICATES THAT THE MODULE IS INVOLVED.

Model Variant	Component					NIST16	
	RGB	NB	DAM	ESM	MGM	F_1	AUC
Baseline	✔					0.621	0.947
Baseline+NB	✔	✔				0.666	0.944
Baseline+NB+DAM	✔	✔	✔			0.722	0.965
Baseline+NB+DAM+ESM	✔	✔	✔	✔		0 773	0.979
Baseline+NB+DAM+ESM+MGM	✔	✔	✔	✔	✔	**0.87**1	**0.984**

TABLE IX
ABLATION STUDY WITH DIFFERENT LOSSES AND WEIGHTS OF LOSS FUNCTION ON NIST16 DATASET. F_1 AND AUC ARE REPORTED IN THIS EXPERIMENT.

Hyperparameter	Loss Function		NIST16	
	L_{M}	L_{E}	F_1	AUC
$\lambda_{\mathrm{M}}=1,\lambda_{\mathrm{E}}=0$	L_{B}	-	0.622	0.887
	$L_{\mathrm{B}}+L_{\mathrm{D}}$	-	0.677	0.903
$\lambda_{\mathrm{M}}=0.8,\lambda_{\mathrm{E}}=0.2$	L_{B}	L_{E}	0.741	0.955
	$L_{\mathrm{B}}+L_{\mathrm{D}}$	L_{E}	**0.871**	**0.984**
$\lambda_{\mathrm{M}}=0.6,\lambda_{\mathrm{E}}=0.4$	L_{B}	L_{E}	0.657	0.947
	$L_{\mathrm{B}}+L_{\mathrm{D}}$	L_{E}	0.762	0.959
$\lambda_{\mathrm{M}}=0.5,\lambda_{\mathrm{E}}=0.5$	L_{B}	L_{E}	0.652	0.954
	$L_{\mathrm{B}}+L_{\mathrm{D}}$	L_{E}	0.733	0.961
$\lambda_{\mathrm{M}}=0.4,\lambda_{\mathrm{E}}=0.6$	L_{B}	L_{E}	0.620	0.940
	$L_{\mathrm{B}}+L_{\mathrm{D}}$	L_{E}	0.703	0.956
$\lambda_{\mathrm{M}}=0.2,\lambda_{\mathrm{E}}=0.8$	L_{B}	L_{E}	0.624	0.923
	$L_{\mathrm{B}}+L_{\mathrm{D}}$	L_{E}	0.710	0.943

these corrupted data. The comparison results of F_1 and AUC between the proposed method and the SoTA methods on corrupted NIST16 dataset are shown in Fig. 12, where the values of kernel size k, quality factor q, and variance σ are $[3, 9, 15]$, $[50, 75, 100]$, and $[3, 9, 15]$, respectively. It is observed that whichever image post-processing operations are used, the F_1 and AUC of the proposed method are higher than those of the other four SoTA methods. It indicates that our method has better robustness against multiple distortion attacks. The performance of the SoTA methods is decreased when dealing with the post-processed images, because of failing to effectively extract the tampering traces hidden by the post-processing operations. However, the proposed method extracts complementary information of semantic features and noise features to enhance the tampering feature representation. Meanwhile the edge supervision strategy is utilized to extract weakened edge artifacts to assist the localization of tampered regions. Since most of the post-processing operations have a severe impact on the noise branch, we fuse multi-scale semantic features in a progressive decoding manner to reduce the false and missed alarm rates for localization of tampered regions. In summary, the strategy of fusing different tampered features (i.e., semantic features, noise features, and edge features) and progressive mask generation can effectively handle post-processing attacks.

To test the model's performance in detecting different sizes of tampered area, two types of tampering datasets are constructed, i.e., the large tampered area dataset and the small

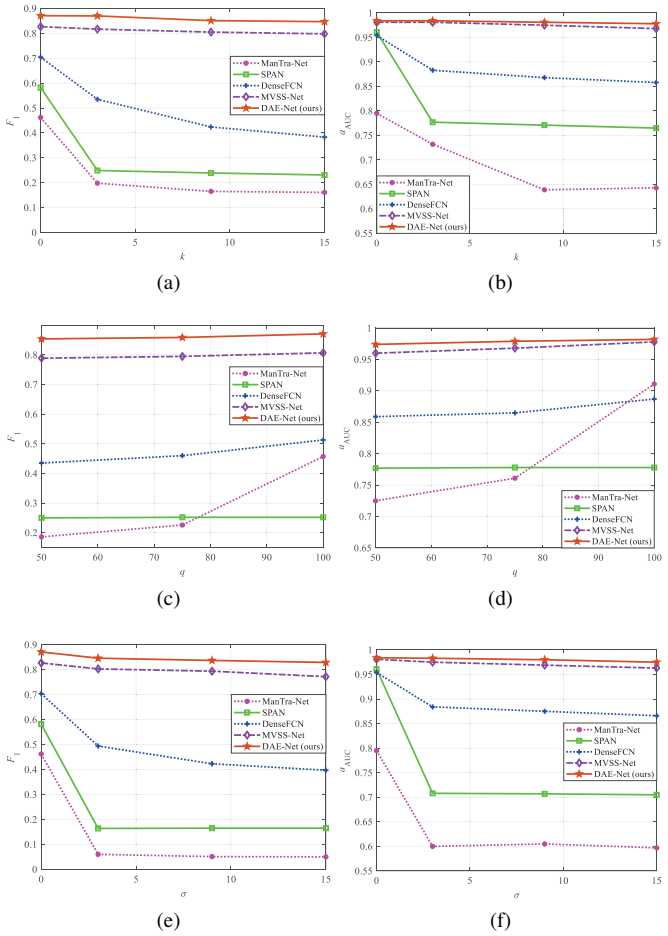


Fig. 12. Robustness comparison of the proposed DAE-Net with four SoTA methods under three image processing techniques, i.e., (a) and (b) Gaussian blur with kernel size k, (c) and (d) JPEG compression with quality factor q, (e) and (f) Gaussian noise with variance σ. The results of F_1 and AUC are reported, respectively.

tampered area dataset. These two datasets each contain 100 images from five datasets mentioned above. The datasets in which the pixels of tampered areas account for less than 12% of the pixels of the whole image are defined as small tampered area, which is mainly based on the definition of the relative size of small targets. The test results on the two datasets are shown in Table X. It is observed that the proposed method achieves the optimal detection results compared with four SoTA methods in terms of F_1 and AUC, which are 0.902 and 0.956 for the large tampered area dataset, and 0.708 and 0.953

TABLE X
DETECTION PERFORMANCE OF DAE-NET IS COMPARED WITH FOUR SOTA METHODS ON LARGE TAMPERED AREA DATASET AND SMALL TAMPERED AREA DATASET. THE RESULTS OF F_1 AND AUC ARE REPORTED.

Method	Large Tampered Area		Small Tampered Area	
	F_1	AUC	F_1	AUC
ManTra-Net [20]	0.572	0.725	0.108	0.447
DenseFCN [30]	0.506	0.709	0.245	0.711
MVSS-Net [24]	0.807	0.947	0.376	0.821
DMFF-Net [32]	0.873	0.943	0.561	0.871
DAE-Net (ours)	**0.902**	**0.956**	**0.708**	**0.953**

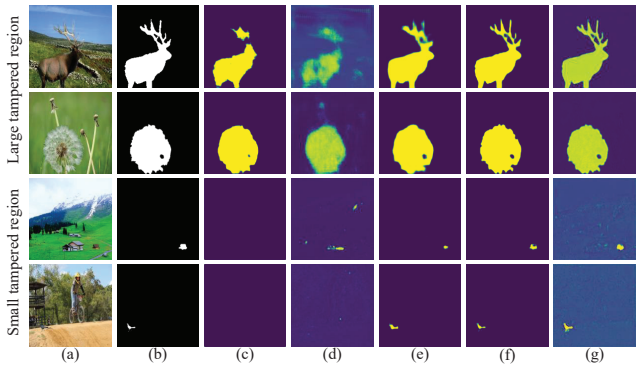


Fig. 13. Qualitative visualization results of the proposed DAE-Net on large tampered area dataset and small tampered area dataset: (a) Original forged image, (b) Ground-truth binary mask, (c) ManTra-Net, (d) DenseFCN, (e) MVSS-Net, (f) DMFF-Net, and (g) DAE-Net (ours).

for the small tampered area dataset, respectively. In addition, we visualize the results of manipulation localization on both datasets, as shown in Fig.13. In summary, the proposed method can effectively cope with different sizes of tampered area, which is mainly attributed to the fact that the proposed model fully utilizes semantic features at different scales as well as edge information.

Moreover, the performance of the model under specific conditions (low resolution and high noise environment) is discussed. The experiments are performed on the NIST16 and Columbia datasets. For the low resolution setting, the original image input 512×512 are reduced to 0.75, 0.5, and 0.25 times of the original, i.e., the input image sizes become 384×384, 256×256, and 128×128, respectively. For the high-noise environment setting, different intensities of Gaussian noise are added to the input images, where the noise variance is set to 3, 9, and 15, respectively. The detection results of the proposed DAE-Net on NIST16 and Columbia in low resolution and high noise environments are shown in Table XI. It is observed that the proposed model is robust to high noise environments. For the low-resolution case, the performance of model shows the decrease, indicating that the proposed method has some limitations for low resolution detection, which is mainly due to the fact that as the image resolution decreases, the forged boundaries and forged traces of the tampered image are blurred, making it difficult for the model to identify the tampered and authentic region efficiently.

TABLE XI
DETECTION PERFORMANCE OF THE PROPOSED DAE-NET ON NIST16 AND COLUMBIA IN LOW RESOLUTION AND HIGH NOISE ENVIRONMENTS. THE RESULTS OF F_1 AND AUC ARE REPORTED.

Specific Condition	Parameter	NIST16		Columbia	
		F_1	AUC	F_1	AUC
Original	-	0.871	0.984	0.872	0.973
Low Resolution	384×384	0.772	0.973	0.780	0.932
	256×256	0.754	0.965	0.660	0.855
	128×128	0.677	0.952	0.602	0.827
High Noise Environments	$\sigma = 3$	0.846	0.983	0.869	0.968
	$\sigma = 9$	0.837	0.980	0.834	0.960
	$\sigma = 15$	0.829	0.975	0.828	0.952

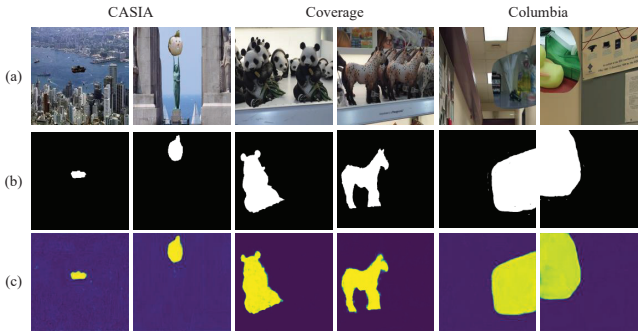


Fig. 14. Qualitative visualization results of the proposed DAE-Net on CASIA, Coverage, and Columbia: (a) Original forged image, (b) Ground-truth binary mask, and (c) DAE-Net (ours).

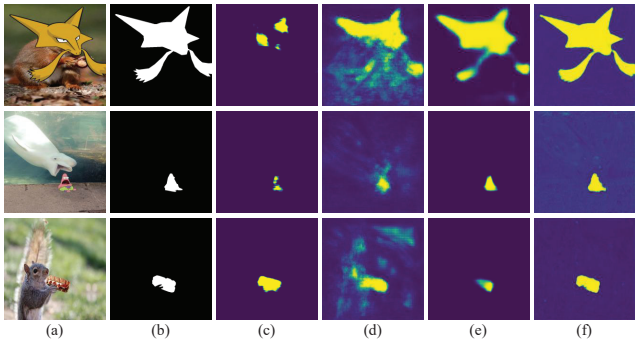


Fig. 15. Qualitative visualization results of the proposed DAE-Net on IMD2020 real-life dataset: (a) Original forged image, (b) Ground-truth binary mask, (c) ManTra-Net, (d) DenseFCN, (e) MVSS-Net, and (f) DAE-Net (ours).

E. Qualitative Visualizations

1) Qualitative Results on CASIA, Coverage, and Columbia: To verify the generalization of the proposed DAE-Net on different datasets, the predicted masks on CASIA [52], Coverage [55], and Columbia [54] are shown in Fig. 14. It is observed that our method achieves accurate localization results for forged images on different datasets. No matter whether they are small manipulated targets or similar targets, our method can localize them accurately. It also illustrates that the proposed DAE-Net has the superior generalization ability. The better localization results of DAE-Net are attributed to extensive exploration of tampered features (i.e., semantic

features, noise features, and subtle edge features) and the enhancement of subtle artifacts.

2) Qualitative Results on the IMD2020 Real-life Dataset: We compare the proposed DAE-Net with ManTra-Net [20], DenseFCN [30], and MVSS-Net [24] on the IMD2020 [56] real-life dataset. The qualitative visualization results are shown in Fig. 15. Compared with the other datasets, the IMD2020 [56] dataset contains more real forged scenarios, which can make the evaluation more convincing. As can be seen in Fig. 15, the proposed DAE-Net can segment the tampered regions completely and accurately. Nevertheless, ManTra-Net [20] cannot locate the tampered regions accurately and has poor generalization to the real-life dataset. Both DenseFCN [30] and MVSS-Net [24] can roughly localize the main body of the tampered regions, but the integrity of the tampered regions and the accuracy of boundary localization are unsatisfactory. In addition, it is seen from the second row from Fig. 15, DAE-Net also identifies small tampered regions. These localization results indicate that the edge supervision module and the progressive decoding manner can effectively capture edge artifacts and locate the tampered regions completely.

F. Computational Complexity Analysis

In this section, we compute and compare the complexity of different IMDL methods, i.e., ManTra-Net [20], DenseFCN [30], and MVSS-Net [24]. The computational complexity was calculated and compared to the above models in NVDIA GeForce RTX 2080Ti GPU device. We compared model parameters, floating point operations (FLOPs), inference time per image, and performance. Table XII presents the experimental results, where the inference time is the average of 150 randomly selected samples and performance is the average AUC on the five datasets. From Table XII, it is observed that proposed method consumes less inference time compared with ManTra-Net [20], with only 105.56 ms per image. Moreover, our model parameters are smaller compared with MVSS-Net. In summary, the proposed method achieves relatively small computational complexity, which means that our method can obtain a better experience in real scenarios compared with the SoTA methods.

TABLE XII
COMPUTATIONAL COMPLEXITY COMPARED WITH THREE SOTA METHODS, WHERE "M" REPRESENTS MILLION, "G" REPRESENTS BILLION, AND "MS" REPRESENTS MILLISECONDS.

Method	Params (M)	FLOPs (G)	Inference (ms)	Performance
ManTra-Net [20]	3.80	249.54	145.61	0.799
DenseFCN [30]	0.59	7.71	97.78	0.789
MVSS-Net [24]	142.78	163.38	108.57	0.815
DAE-Net (ours)	85.32	86.51	105.56	0.866

V. CONCLUSION

In this work, we propose a novel dual attention mechanism and edge supervision network to tackle the challenge of advanced image manipulation techniques. To obtain the invisible artifact information in the RGB domain, we extract the fused global and local inconsistent noise features in the noise domain as complementary information. Subsequently, we design a dual attention mechanism module to enhance tampering feature representations by mining the complementary relationships of features from different domains. Furthermore, the progressive mask generation module is utilized to achieve manipulation localization at the pixel level, in which the designed fusion upsampling module makes full use of the multi-scale semantic features to improve the accuracy of manipulation localization. Finally, the edge supervision module is designed to extract subtle edge features and refine the boundary of tampered regions. Extensive experiments on different datasets demonstrate the effectiveness and viability of the proposed method, which has better or comparable performance compared to the SoTA methods.

DAE-Net improves the accuracy of manipulation localization by extracting multiple tampered traces. However, the generalization of the proposed method in cross-dataset detection still needs to be further improved. In further work, we will reconstruct the network architecture for strong generalization. On the one hand, the backbone of the noise branch is improved to extract global and local features by using a more lightweight architecture. On the other hand, more components can be designed to explore more tampered artifacts (e.g., compression artifacts, frequency domain features) for IMDL. Moreover, as AI-generated models evolve and a large number of AI images are produced, the compatibility of model to detect AI-generated images can be further considered.

REFERENCES

[1] M. Barni, Q.-T. Phan, and B. Tondi, "Copy move source-target disambiguation through multi-branch CNNs," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1825–1840, Jan. 2021.

[2] Z. Shi, H. Chen, and D. Zhang, "Transformer-auxiliary neural networks for image manipulation localization by operator inductions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4907–4920, Sep. 2023.

[3] F. Li, Z. Pei, X. Zhang, and C. Qin, "Image manipulation localization using multi-scale feature fusion and adaptive edge supervision," *IEEE Trans. Multimedia*, vol. 25, pp. 7851–7866, Dec. 2023.

[4] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7505–7517, Nov. 2022.

[5] Y. Zhu, C. Chen, G. Yan, Y. Guo, and Y. Dong, "AR-Net: Adaptive attention and residual refinement network for copy-move forgery detection," *IEEE Trans. Ind. Inform.*, vol. 16, no. 10, pp. 6714–6723, Oct. 2020.

[6] Y. Liu, X. Zhu, X. Zhao, and Y. Cao, "Adversarial learning for constrained image splicing detection and localization based on atrous convolution," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 10, pp. 2551–2566, Oct. 2019.

[7] H. Wu and J. Zhou, "IID-Net: Image inpainting detection network via neural architecture search and attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1172–1185, Mar. 2022.

[8] C.-M. Pun, X.-C. Yuan, and X.-L. Bi, "Image forgery detection using adaptive oversegmentation and feature point matching," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 8, pp. 1705–1716, Aug. 2015.

[9] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.

[10] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1497–1503, Sep. 2009.

[11] S. P. Jaiprakash, M. B. Desai, C. S. Prakash, V. H. Mistry, and K. L. Radadiya, "Low dimensional DCT and DWT feature based model for detection of image splicing and copy-move forgery," *Multimed. Tools Appl.*, vol. 79, no. 39–40, pp. 29 977–30 005, Oct. 2020.

[12] G. Gani and F. Qadir, "Copy move forgery detection using DCT, patchmatch and cellular automata," *Multimed. Tools Appl.*, vol. 80, no. 21–23, pp. 32 219–32 243, Sep. 2021.

[13] I. Amerini, T. Uricchio, L. Ballan, and R. Caldelli, "Localization of JPEG double compression through multi-domain convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops*, Honolulu, HI, USA, 2017, pp. 1865–1871.

[14] C. Zhai, L. Yang, Y. Liu, and H. Yu, "DBMA-Net: A dual-branch multiattention network for polyp segmentation," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–16, Mar. 2024.

[15] H. Shi, Y. Zhang, G. Cao, and D. Yang, "MHCFormer: Multiscale hierarchical conv-aided Fourierformer for hyperspectral image classification," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, Dec. 2024.

[16] Y. Luo, F. Shao, Z. Xie, H. Wang, H. Chen, B. Mu, and Q. Jiang, "HFMDNet: Hierarchical fusion and multilevel decoder network for RGB-D salient object detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, Feb. 2024.

[17] B. Chen, W. Tan, G. Coatrieux, Y. Zheng, and Y.-Q. Shi, "A serial image copy-move forgery localization scheme with source/target distinguishment," *IEEE Trans. Multimedia*, vol. 23, pp. 3506–3517, Oct. 2021.

[18] Y. Xu, M. Irfan, A. Fang, and J. Zheng, "Multiscale attention network for detection and localization of image splicing forgery," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, Sep. 2023.

[19] Y. Zhang, Z. Fu, S. Qi, M. Xue, Z. Hua, and Y. Xiang, "Localization of inpainting forgery with feature enhancement network," *IEEE Trans. Big Data*, vol. 9, no. 3, pp. 936–948, Jun. 2023.

[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.*, Long Beach, CA, USA, 2019, pp. 9535–9544.

[21] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 1053–1061.

[22] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, UK, 2020, pp. 312–328.

[23] Z. Shi, C. Chang, H. Chen, X. Du, and H. Zhang, "PR-Net: Progressively-refined neural network for image manipulation localization," *Int. J. Intell. Syst.*, vol. 37, no. 5, pp. 3166–3188, May 2022.

[24] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "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.

[25] Z. Gao, C. Sun, Z. Cheng, W. Guan, A. Liu, and M. Wang, "TBNet: A two-stream boundary-aware network for generic image manipulation localization," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 7541–7556, Jul. 2023.

[26] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 3155–3165.

[27] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2691–2706, Nov. 2018.

[28] X. Lin, S. Wang, J. Deng, Y. Fu, X. Bai, X. Chen, X. Qu, and W. Tang, "Image manipulation detection by multiple tampering traces and edge artifact enhancement," *Pattern Recognit.*, vol. 133, pp. 1–11, Jan. 2023.

[29] 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 Conf. Comput. Vis. Pattern Recognit.*, Montreal, QC, Canada, 2021, pp. 9992–10 002.

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

[31] L. Zhuo, S. Tan, B. Li, and J. Huang, "Self-adversarial training incorporating forgery attention for image forgery localization," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 819–834, Mar. 2022.

[32] X. Xia, L. C. Su, S. P. Wang, and X. Y. Li, "DMFF-Net: Double-stream multilevel feature fusion network for image forgery localization," *Eng. Appl. Artif. Intell.*, vol. 127, pp. 1–13, Jan. 2024.

[33] F. Li, H. Zhai, X. Zhang, and C. Qin, "Image manipulation localization using spatialchannel fusion excitation and fine-grained feature enhancement," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, Feb. 2024.

[34] C. Yang, H. Li, F. Lin, B. Jiang, and H. Zhao, "Constrained R-CNN: A general image manipulation detection model," in *Proc. IEEE Int. Conf. Multimedia Expo*, London, UK, 2020, pp. 1–6.

[35] Y. Zeng, B. Zhao, S. Qiu, T. Dai, and S.-T. Xia, "Toward effective image manipulation detection with proposal contrastive learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4703–4714, Sep. 2023.

[36] L. Zhang, M. Xu, D. Li, J. Du, and R. Wang, "CatmullRom splines-based regression for image forgery localization," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, Canada, 2024, pp. 7196–7204.

[37] K. Ji, F. Chen, X. Guo, Y. Xu, J. Wang, and J. Chen, "Uncertainty-guided learning for improving image manipulation detection," in *Proc. IEEE Int. Conf. Comput. Vision*, Paris, France, 2023, pp. 22 399–22 408.

[38] D. Xu, X. Shen, and Y. Lyu, "UP-Net: Uncertainty-supervised parallel network for image manipulation localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6390–6403, Nov. 2023.

[39] J. Zhu, D. Li, X. Fu, G. Yang, J. Huang, A. Liu, and Z.-J. Zha, "Learning discriminative noise guidance for image forgery detection and localization," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, Canada, 2024, pp. 7739–7747.

[40] Y. Huang, S. Bian, H. Li, C. Wang, and K. Li, "DS-UNet: A dual streams UNet for refined image forgery localization," *Inf. Sci.*, vol. 610, pp. 73–89, Sep. 2022.

[41] Y. Liu, B. Lv, X. Jin, X. Chen, and X. Zhang, "TBFormer: Two-branch transformer for image forgery localization," *IEEE Signal Process. Lett.*, vol. 30, pp. 623–627, Jun. 2023.

[42] I. I. Ganapathi, S. Javed, S. S. Ali, A. Mahmood, N.-S. Vu, and N. Werghi, "Learning to localize image forgery using end-to-end attention network," *Neurocomputing*, vol. 512, pp. 25–39, Nov. 2022.

[43] D. Xu, X. Shen, Y. Lyu, X. Du, and F. Feng, "MC-Net: Learning mutually-complementary features for image manipulation localization," *Int. J. Intell. Syst.*, vol. 37, no. 5, pp. 3072–3089, May 2022.

[44] J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, and Y.-G. Jiang, "ObjectFormer for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 2354–2363.

[45] Z. Gao, S. Chen, Y. Guo, W. Guan, J. Nie, and A. Liu, "Generic image manipulation localization through the lens of multi-scale spatial inconsistence," in *Proc. ACM Int. Conf. Multimed.*, Lisbon, Portugal, 2022, pp. 6146–6154.

[46] D. Li, J. Zhu, M. Wang, J. Liu, X. Fu, and Z.-J. Zha, "Edge-aware regional message passing controller for image forgery localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 8222–8232.

[47] N. Zeng, P. Wu, Y. Zhang, H. Li, J. Mao, and Z. Wang, "DPMSN: A dual-pathway multiscale network for image forgery detection," *IEEE Trans. Ind. Inform.*, vol. 20, no. 5, pp. 7665–7674, May 2024.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[49] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 3141–3149.

[50] S. Li, S. Xu, W. Ma, and Q. Zong, "Image manipulation localization using attentional cross-domain CNN features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5614–5628, Sep. 2023.

[51] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise mapping," *Neurocomputing*, vol. 506, pp. 158–167, Sep. 2022.

[52] 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, 2013, pp. 422–426.

[53] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus, "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops*, Waikoloa, HI, USA, 2019, pp. 63–72.

[54] 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, Canada, 2006, pp. 549–552.

[55] 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.*, Phoenix, AZ, USA, 2016, pp. 161–165.

[56] A. Novozamsky, B. Mahdian, and S. Saic, "IMD2020: A large-scale annotated dataset tailored for detecting manipulated images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops*, Snowmass, CO, USA, 2020, pp. 71–80.

[57] H. Ding, L. Chen, Q. Tao, Z. Fu, L. Dong, and X. Cui, "DCU-Net: A dual-channel U-shaped network for image splicing forgery detection," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5015–5031, Mar. 2023.

[58] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 3–19.

Chunyin Shi received the B.E. degree in electronic information engineering from Hohai University, China, in 2022. He is currently pursuing the M.E. degree in information and communication engineering with Shandong University, China. His current research interests include image forgery detection, deepfake detection, and computer vision.

Chengyou Wang (Member, IEEE) received the B.E. degree in electronic information science and technology from Yantai University, China, in 2004, and the M.E. and Ph.D. degrees in signal and information processing from Tianjin University, China, in 2007 and 2010, respectively. He is currently an Associate Professor and a Supervisor of Ph.D. students with Shandong University, Weihai, China. His current research interests include signal and information processing, digital image/video processing, computer vision, artificial intelligence, and wireless communication technology.

Xiao Zhou (Member, IEEE) received the B.E. degree in automation from Nanjing University of Posts and Telecommunications, China, in 2003, the M.E. degree in information and communication engineering from Inha University, South Korea, in 2005, and the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2013. She is currently an Associate Professor and a Supervisor of master's students with Shandong University, Weihai, China. Her current research interests include wireless communication technology, intelligent information processing, image communication, digital image/video processing and analysis, computer vision, and artificial intelligence.

Zhiliang Qin (Senior Member, IEEE) received the B.E. degree from the Beijing Institute of Technology in 1995, the M.E. degree from the Graduate School of China Academy of Engineering Physics in 1998, and the Ph.D. degree from the Nanyang Technological University, Singapore in 2003. From 2002 to 2019, he worked at the Agency for Science, Technology, and Research in Singapore as the Scientist in the area of algorithm developments for machine learning, signal processing, data analytics, optimization theories, and data storage systems. From 2019 to present, he is the Deputy Chief Engineer at the Weihai Beiyang Electrical Group Co. Ltd., Weihai, China, and also with the Adjunct Distinguished Professor affiliated with the School of Mechanical, Electrical and Information Engineering, Shandong University, China.