EAN: Edge-Aware Network for Image Manipulation Localization

Yun Chen, Hang Cheng, Haichou Wang, Ximeng Liu, *Senior Member, IEEE,* Fei Chen, *Member, IEEE,* Fengyong Li, Xinpeng Zhang, *Senior Member, IEEE,* Meiqing Wang

***Abstract*—Image manipulation has sparked widespread concern due to its potential security threats on the Internet. The boundary between the authentic and manipulated region exhibits artifacts in image manipulation localization (IML). These artifacts are more pronounced in heterogeneous image splicing and homogeneous image copy-move manipulation, while they are more subtle in removal and inpainting manipulated images. However, existing methods for image manipulation detection tend to capture boundary artifacts via explicit edge features and have limitations in effectively addressing subtle artifacts. Besides, feature redundancy caused by the powerful feature extraction capability of large models may prevent accurate identification of manipulated artifacts, exhibiting a high false-positive rate. To solve these problems, we propose a novel edge-aware network (EAN) to capture boundary artifacts effectively. This network treats the image manipulation localization problem as a segmentation problem inside and outside the boundary. In EAN, we develop an edge-aware mechanism to refine implicit and explicit edge features by the interaction of adjacent features. This approach directs the encoder to prioritize the desired edge information. Also, we design a multi-feature fusion strategy combined with an improved attention mechanism to enhance key feature representation significantly for mitigating the effects of feature redundancy. We perform thorough experiments on diverse datasets, and the outcomes confirm the efficacy of the suggested approach, surpassing leading manipulation localization techniques in the majority of scenarios.**

***Index Terms*—Image manipulation localization, Convolutional neural network, Feature fusion, Attention mechanism.**

I. INTRODUCTION

THE emergence of user-friendly advanced software has made it easy to edit even temper pictures, impacting their authenticity and giving rise to illegal forgery. Currently, image manipulation methods are divided into two categories: manipulation without altering semantics and manipulation involving changes in semantics. The former does less harm to society, whereas the latter is often exploited by criminals, leading to issues such as the dissemination of fake news [1], the misuse of manipulated images [2], and even using manipulated images as spurious evidence. Therefore, the detection of manipulated images is of significant importance. As shown in Fig. 1, in the context of image manipulation methods involving changes in semantics, techniques such as splicing transferring a section from one image to another, copy-move, which duplicates an object within the same image, and removal, where a specific region is erased, are commonly observed. Our goal is to leverage neural networks to discern the inherent distinctions between manipulated and authentic regions in images, with edge features playing a crucial role. In case of copy-move and splicing, the boundary artifacts are typically conspicuous, making them relatively easy for most neural networks to distinguish. However, extracting boundary artifacts can be challenging for manipulation methods like removal and repair, which lacks explicit boundary features. These techniques often involve subtle boundary alterations making it harder for neural networks to detect the implicit boundary artifacts.

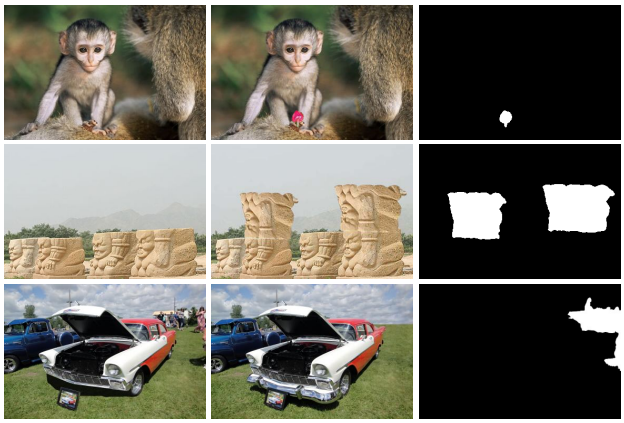

Fig. 1. Examples of image manipulation localization. The examples are the manipulation type of splicing, copy-move, and removal list from top to bottom.

Deep learning technology has seen significant growth in image manipulation localization in recent years. RGB-N [3] used Faster R-CNN as the main stem network, extracting features from RGB images and noise images to distinguish RGB feature differences and noise inconsistency between manipulated regions and authentic regions, but its output is limited to bounding boxes rather than segmenting masks. A network called Mantra-Net [4] transformed the localization problem of manipulated regions into local abnormality detec-

Yun Chen, Ximeng Liu, Fei Chen, are with the Colleage of Computer and Data Science, Fuzhou University, Fuzhou, Fujian 350108, China (e-mail: ychenfzu@163.com; snbnix@gmail.com; chenfei314@fzu.edu.cn).

Hang Cheng, Haichou Wang, Meiqing Wang, are with the School of Mathematics and Statistics, Fuzhou University, Fuzhou, Fujian 350108, China (e-mail: hcheng@fzu.edu.cn; haichouwang@163.com; mqwang@fzu.edu.cn).

Fengyong Li is with the College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China (e-mail: fyli@shiep.edu.cn).

Xinpeng Zhang is with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: zhangxinpeng@fudan.edu.cn).

Corresponding author: Hang Cheng.

tion problems, employing the Z-Score feature to detect these abnormalities, and evaluated using the Long Short Term Network (LSTM). SPAN [5] effectively modeled the correlation among multi-scale image blocks by creating feature pyramids composed of local self-attention blocks. Additionally, location projections were devised to encode the spatial positions of the image patches. MVSS-Net [6] employed multi-scale feature learning and multi-scale supervision to perform manipulation localization, strengthening the learning of boundary features using noise distribution and boundary artifacts around the manipulated region. Nevertheless, the aforementioned methods that don't incorporate implicit edge features or rely solely on explicit edge features may struggle with accurately positioning predicted masks. In addition, attributed to the powerful feature extraction capability of current neural network models, feature redundancy rather hinders performance improvement [7]. Balancing the utilization of features and considering potential redundancies are crucial for optimizing the performance of image manipulation localization.

In this work, the proposed edge-aware network to capture manipulated artifacts by finely fusing noise features, RGB features, and edge features. The localization of the manipulated region is essentially a problem of binary semantic segmentation of an image, so the attributes and semantic features of the image itself are significant. We use ResNet34 [8] as the backbone of the feature extractor. Inspired by Unet [9] and DenseNet [10], we design an RGB feature extractor that incorporates both low-level image features with high-level semantic features through skip connections. Similarly, we improve the noise feature extractor to capture noise inconsistencies between the authentic and manipulated regions. Then, the edge-aware module is responsible for merging the edge features extracted from the RGB and noise streams. Notably, the encoder is directed to focus on the boundary region by edge-aware module. The above features are combined and enhanced by the feature fusion module according to our proposed attention mechanism for mitigating the effects of feature redundancy. Finally, the decoder predict whether every pixel of the image has been modified or not. In summary, our primary contributions are as follows:

- We present an edge-aware mechanism to achieve feature refinement from the explicit and implicit edge features by the interaction of adjacent layer driven features. Different from the existing methods that we make edge features a powerful cue for discriminating manipulated regions and the localization performance is improved.
- We propose a multi-feature fusion module with an improved attention mechanism to optimize the contribution of different features for better discrimination of boundary artifacts, which mitigates the effects of feature redundancy.
- We analyze the effectiveness of EAN in locating manipulated regions, which verifies the feasibility of an encoder for boundary feature-guided neural networks. We perform thorough experiments on diverse datasets, surpassing leading other techniques in accuracy in manipulation localization and the robustness of the network model.

The remainder of this paper is organized as follows. Section II lists the abbreviations used in the paper. In Section III, we review related works on image manipulation localization. Section IV describes the proposed network architecture. Section V presents the experimental results and analyzes the performance of the proposed method. Finally, Section VI concludes the paper and discusses the future directions.

II. List of Abbreviations

Abbreviation	Definition
IML	Image Manipulation Localization
EAN	Edge-Aware Network
DCT	Discrete Cosine Transform
PCA	Principal Component Analysis
CNN	Convolutional Neural Network
DWT	Discrete Wavelet Transform
FT	Fourier Transform
PCL	Proposal Contrastive Learning
LSTM	Long Short Term Network
SIFT	Scale Invariant Feature Transform
ORB	Oriented FAST and Rotated BRIEF
ViT	Vision Transformer Model
RFE	RGB Feature Extractor
NFE	Noise Feature Extractor
EAM	Edge-Aware Module
FFM	Feature Fusion Module
ESAB	Enhanced Self-Attention Block
SAB	Spatial Attention Block
DB	Decoder Block
NIST16	NIST Nimble 2016
AUC	Area Under the Curve
GradCAM	Gradient-Weighted Class Activation Map

III. Related Work

A. Image Manipulation Localization

Early works primarily relied on manual feature, which involved modeling images to unveil statistical dependencies among pixels. Anomalous pixels were identified based on the captured statistical features. For example, in [11], novel probability models were introduced for the Discrete Cosine Transform (DCT) coefficients of singly and doubly compressed areas. Additionally, a dependable approach for estimating the main quantization factor in instances of double compression was developed. In [12], the Markov model is suggested for detecting passive digital image splicing. The patches-based approach first divides the image into two categories patches: overlap and non-overlap and then uses Principle Component Analysis (PCA) [13], DCT [14], Local Binary Pattern (LBP) [15], Discrete Wavelet Transform (DWT) [16], Fourier Transform (FT) [17] to extract relevant features from these image patches. There were also methods based on keypoint detection, such as Scale Invariant Feature Transform (SIFT) [18], Oriented FAST and Rotated BRIEF (ORB) [19], etc. In [20], the Proposal Contrastive Learning (PCL) was proposed, which

exploited the relationships between local features through a proxy suggestion contrastive learning task by attracting and rejecting proposal-based positive/negative sample pairs.

Conventional methods rely on handcrafted features to identify manipulated areas, which are easily obsolete by new manipulation techniques. In contrast, deep learning-based approaches can learn deep semantic features of manipulated pixels and effectively capture manipulation artifacts. Many deep learning-based approaches have proposed to locate the manipulation region of a specific type, *e.g.*, splicing [1], [12], [21], [22], copy-move [23]–[27], and removal [28]–[30]. Recently, some works have focused on manipulation localization for post-processed images [31], [32]. In this work, we design a network architecture to localize the manipulated regions by mining the powerful clues of the forgery artifacts from the redundant features that is applicable to all types of manipulation methods.

B. Attention Mechanism

Neural networks are commonly perceived as black-box models making it impractical to manually adjust weights for specific areas of interest. This characteristic make it challenging to directly interpret how the network makes its decisions or to intervene in the learning process to prioritize certain features over others. Despite this limitation, researchers have been exploring methods to enhance the interpretability of neural networks, such as using attention mechanisms. The attention mechanism was first proposed by [33] and has significantly benefited various fields, such as Machine Translation [34], where it enables models to concentrate on pertinent sections of the input sequence during translation. In Object Detection [35], the attention mechanism enhanced the ability to detect objects in cluttered scenes by selectively attending to important regions in an image. The ability to prioritize and weigh the significance of different features leads to more efficient learning and better performance in tasks like image manipulation localization. SPAN [5] used the self-attention mechanism to model global features at multiple scales and calculated a new representation based on the relationship between each pixel and adjacent pixels in the self-attention layer. Based on the self-attention mechanism, a novel deep learning model called Transformer has been proposed and achieves better performance than traditional CNN in many aspects. Recently, the Vision Transformer model (ViT) [36] has made the first attempt to directly apply the transformer to image patch sequences, with state-of-the-art performance on several image recognition benchmarks. ObjectFormer [37] combined CNN with Transformer to propose an Object-level intermediate representation and introduce image patches position coding in frequency and RGB domain. In this work, we enhance the awareness of edge features through the edge-aware mechanism and use an improved attention mechanism to optimize feature fusion. Consequently, the adverse effects stemming from feature redundancy are mitigated by consistently emphasizing features that are pertinent for precise manipulation localization.

IV. PROPOSED METHOD

A. Overview

In this section, we demonstrate the edge-aware network. We first provide an overview of our EAN and the details of the individual components are then presented. Fig. 2 gives an overview of our EAN, which consists of an RGB feature extractor (RFE) (Section IV-B), a noise feature extractor (NFE) (Section IV-C), an edge-aware module (EAM) (Section IV-D) for aggregating edge features to guide EAN to focus on boundary region, and a feature fusion module (FFM) (Section IV-E) for feature enhancement by improved attention mechanism.

Specifically, we denote an input image as $X \in \mathbb{R}^{(H\times W\times 3)}$. As for the choice of noise extraction, we apply Constrain Conv [38] to transform X to X_n, representing the noise characteristic distribution. Then, we use the RGB feature extractor and noise feature extractor to extract the RGB feature map $F_r \in \mathbb{R}^{(H\times W\times C)}$ and the noise feature map $F_n \in \mathbb{R}^{(H\times W\times C)}$. Edge-aware module is responsible for combining explicit edge feature from RFE and implicit edge feature from NFE to form edge feature $F_e \in \mathbb{R}^{(H\times W\times C)}$.

$$\tilde{M} = Decoder(FFM(F_r, F_n, F_e)). \tag{1}$$

The data flow of EAN is conceptually expressed by Eq.1, where H, W, and C are the height, width, and channel of the feature map, respectively. The fusion result of these features is then fed to FFM. At last, $X \in \mathbb{R}^{(H\times W\times 3)}$ is mapped to a binary mask $\tilde{M} \in \mathbb{R}^{(H\times W\times 1)}$ by decoder.

B. RGB Feature Extractor

In neural networks, shallow networks tend to preserve more image attributes, such as abnormal edge artifacts and discordant texture information, through the extraction of low-level features. Conversely, deep networks capture richer semantic information in the high-level features [6]. Therefore, the integration of low-level and high-level features holds significant importance. In our method, RFE leverages ResNet34 as the backbone network and appends N Basic Blocks at the end to enhance network depth, facilitating the extraction of profound semantic features. The features $\{R_1, R_2, R_3\} \in \mathbb{R}^{(H\times W\times C)}$ are extracted in the first, second, and fourth layers of the network, respectively. These features are subsequently fed into EAM for refinement. In contrast to concatenating features layer-by-layer, which can introduce redundancy and potentially impact mask detection by causing false positives, our approach diverges from DenseNet [10] by employing element-wise summation across different layers. In this case, the skip connection not only offers the possibility to combine features from various layers but also continually enhances features through element-wise summation. Finally, the deep semantic feature map $R_4 \in \mathbb{R}^{(H\times W\times C)}$ is connected to R_3 and fed into the FFM as the presentation of RGB feature.

$$RFE(X) \rightarrow \{R_1, R_2, R_3, R_4\}, \tag{2}$$

where $\{R_1, R_2, R_3, R_4\}$ are the RGB features of the different layers extracted by RFE.

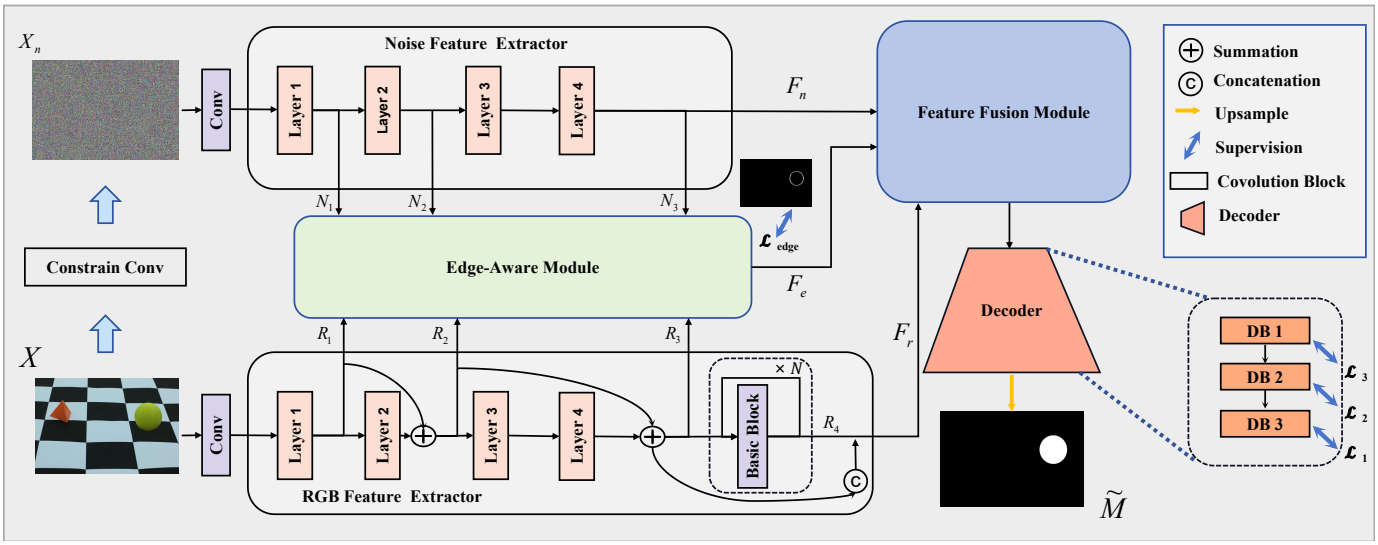

Fig. 2. An overview of EAN. The input is a suspicious image $X \in \mathbb{R}^{(H\times W\times 3)}$, we first leverage color and frequency features extractor to extract features. and the output is a predicted mask $M \in \mathbb{R}^{(H\times W\times 1)}$, which localizes the manipulation regions.

C. Noise Feature Extractor

It is undeniable that current tampering techniques have advanced to the point where they can effectively conceal manipulation traces, rendering them imperceptible to the human eye. Consequently, detecting subtle traces of manipulation within the RGB domain can be challenging. Following image manipulation, post-processing is typically carried out to weaken the traces of manipulation. Hence, extracting features to distinguish the anomalous region in the frequency domain while maintaining the robustness is crucial. We design the NFE to extract noise feature and implicit edge feature from the frequency domain to offer supplementary hints. Similarly, the backbone of NFE is also ResNet34 used to extract noise features from different layers. When processing the input image X, the initial step involves using Constrain Conv [38] to convert it from the RGB domain to the frequency domain:

$$X_n = \mathcal{C}(X), \tag{3}$$

where $X_n \in \mathbb{R}^{(H\times W\times 3)}$ is the frequency domain representation of the image and $\mathcal{C}$ denotes Constrain Conv. Subsequently, X_n is input into the backbone net in order to extract the frequency features of different layers, namely $\{N_1, N_2, N_3\} \in \mathbb{R}^{(H\times W\times C)}$. The outputs of the first layer N_1, second layer N_2, and fourth layer N_3 are then fed into the EAM. Additionally, N_3 serving as a representative of noise feature F_n, is fed into the FFM. The output of the NFE is expressed as

$$NFE(X_n) \rightarrow \{N_1, N_2, N_3\}. \tag{4}$$

D. Edge-Aware Module

EAM is tasked with the refinement of the explicit and implicit features. And we aim for the proposed network architecture to prioritize edge detection, facilitating the learning process to discern distinct characteristics on either side of the edge more effectively. As shown in Fig. 3, with receiving the

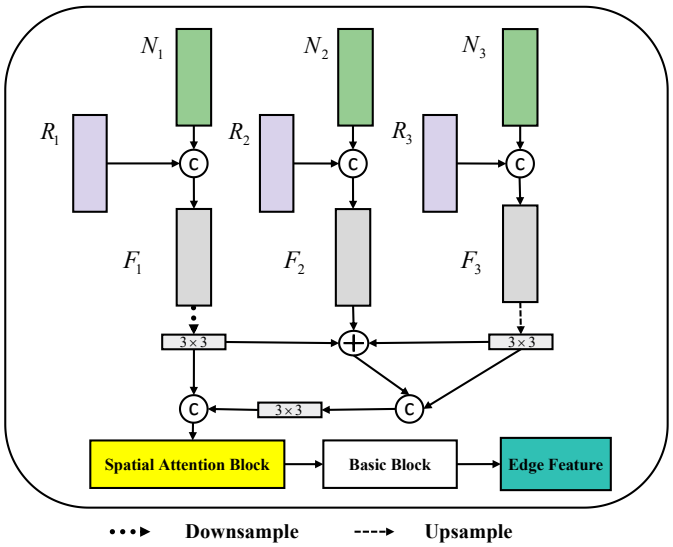

Fig. 3. Edge-Aware Module Graph. Implicit and explicit edge features are refined.

feature input from NFE (N_1, N_2, N_3) and RFE (R_1, R_2, R_3), the integrity of the edge features is ensured by concatenating the feature maps at each level. Then, inspired by [7], the features are reinforced and enhanced by the interaction of adjacent scale features. The procedure can be described by the following equations.

$$\{F_1, F_2, F_3\} = \{N_1 \copyright R_1, N_2 \copyright R_2, N_3 \copyright R_3\}, \tag{5}$$

$$F_4 = interact(down(F_1), F_2, up(F_3)), \tag{6}$$

$$F_e = conv(SAB(F_4)). \tag{7}$$

Where $\{F_1, F_2, F_3\}$ denote the concatenated feature map, $\copyright$ is the concaenate operation of feature maps. F_4 is the output of the interaction among adjacent features. $up(\cdot)$ is upsampling operation and $down(\cdot)$ is downsampling operation. The meaning of $interact(\cdot)$ operation include features map merging

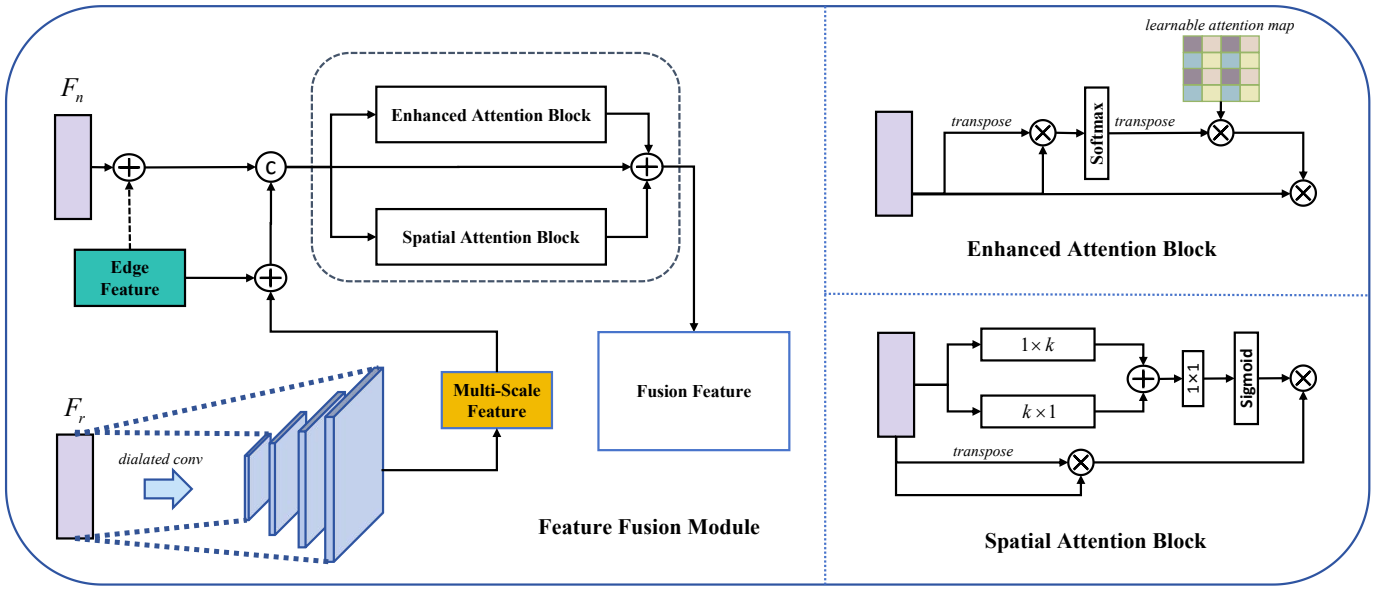

Fig. 4. Diagram of Feature Fusion Module.

and convolution, the specific interaction process is illustrated in Fig. 3. To boost the encoder's focus on edge artifacts, F_4 serves as the input of the spatial attention block (SAB). Different from the channel attention mechanism [39] that adjusts the weights at each channel, the spatial attention mechanism works by adjusting the weights of the corresponding locations of the feature maps. Therefore, the spatial attention mechanism is more appropriate for edge-aware task. To reduce the computational complexity, we employ two complementary and asymmetric convolution kernels $\{1 \times k, k \times 1\}$ to SAB which is based on the self-attention mechanism. Then, the final convolutional block serves to optimize the edge feature representation further and fine-tune the scale of the edge feature. The whole process can be simplified as follows:

$$F_e = EAM(R_1, R_2, R_3, N_1, N_2, N_3). \tag{8}$$

Where F_e denotes that the final edge feature is a combination of implicit and explicit edge features.

E. Feature Fusion Module

The FFM is responsible for merging the acquired edge feature, RGB feature, and noise feature, thereby amplifying the features crucial for detecting manipulation artifacts. Firstly, the RGB feature is transformed into a multi-scale representation using dilated convolution [40] to determine the spatial relationships at different scales. Then, incorporate F_e into multi-scale feature and noise feature. Given the weight-sharing nature of convolutional neural networks, traditional channel-attention and spatial-attention mechanisms encounter challenges in enhancing features from a global perspective. As distinct from convolutional and recursive operations, the self-attention mechanism capture long-range dependencies directly by computing interactions between any two positions, irrespective of their positional separation. Moreover, it can be easily combined with other operations and the self-attention mechanism maintains variable input sizes. Since the edge artifacts are represented in the feature map longitudinal direction, the spatial attention mechanism will be more suitable compared to the channel attention mechanism.

For the better fusion of different features and enhancement of significant features, we propose an improved attention which integrates the self-attention mechanism and spatial attention mechanism. As shown in Fig. 4, the attention mechanism is realized by the enhanced self-attention block (ESAB) and spatial attention block (SAB). ESAB utilizes the self-attention mechanism to associate the forgery artifacts with the fusion feature and selectively updates the learnable matrix. This adaptive update of parameters enables a focus on boundary artifacts during backpropagation. Meanwhile, SAB gathers information on manipulation artifacts at all positions. A pair of asymmetric complementary convolutions is used in SAB to reduce computational complexity. Therefore, the features extracted by EAN will focus on edge and forged regions due to the attention mechanism of FFM. Subsequently, the outputs of SAB and ESAB are aggregated to combine with the input feature, which can be expressed as follows:

$$F_f = F_e + F_n + dialtedconv(F_r), \tag{9}$$

$$FusionFeature = ESAB(F_f) + SAB(F_f) + F_f. \tag{10}$$

Where F_f is the primary fusion feature, the $FusionFeature$ denotes the final fusion feature. Ultimately, the final fusion feature serves as the input to the decoder for predicting the mask. Decoder consists of three decoder blocks (DB), each of which includes upsampling operation and convolution.

F. Loss Function

Let M be the ground-truth mask and $\tilde{M}$ be the predicted mask. Considering that the manipulation localization task resembles a binary classification challenge and faces issues with positive and negative sample inconsistencies in masks. The

TABLE I
THE DETAILS AND EXPERIMENTAL SETUP FOR STANDARD DATASETS.

Datasets	Experimental setup			Involved manipulation technique			Post-processed
	Training set	Testing set	Total	Splicing	Copy-move	Removal	
CASIAV2	5023	-	5023	✓	✓	×	✓
CASIAV1	-	920	920	✓	✓	×	✓
NIST16	414	150	564	✓	✓	✓	✓
Columbia	130	50	180	✓	×	×	×

most image manipulation localization methods employ cross-entropy loss [3] for training. The property of cross-entropy loss operating at the pixel level to evaluate the correctness of each position emphasizes pixel precision, but it falls short in leveraging the spatial location and structural information of manipulation areas. This limitation renders cross-entropy loss unsuitable for calculating edge loss in scenarios with small effective regions and substantial gaps between positive and negative samples. Dice loss [41] is effective in handling class imbalance, which is common in segmentation tasks where the background class dominates the foreground class. It helps address the issue of small effective regions and large gaps between positive and negative samples. So we introduce cross entropy ($\mathcal{L}_{ce}$) of binary classification and Dice loss ($\mathcal{L}_{dice}$).

$$\mathcal{L}_{ce} = -\sum_{i,j} M_{i,j} log(\tilde{M}_{i,j}) - \sum_{i,j} (1 - M_{i,j}) log(1 - \tilde{M}_{i,j}), \tag{11}$$

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{i,j} M_{i,j} \cdot \tilde{M}_{i,j}}{\sum_{i,j} M_{i,j}^2 + \tilde{M}_{i,j}^2}. \tag{12}$$

Where i and j indicate the pixel location in the mask. Therefore, the loss function for the predicted mask is denoted as:

$$\mathcal{L}_{mask} = \mathcal{L}_{ce} + \mathcal{L}_{dice}. \tag{13}$$

And the loss function for edge predicted mask is expressed as:

$$\mathcal{L}_{edge} = \mathcal{L}_{dice}. \tag{14}$$

Considering mask losses at three scales, a total loss can be calculated by:

$$\mathcal{L}_{total} = \mathcal{L}_{edge} + \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3, \tag{15}$$

where $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ denote the loss functions of different decoder blocks in decoder.

V. EXPERIMENTS

A. Experimental Setup

Datasets. We compare our method with current state-of-the-art methods on NIST Nimble 2016 [42] (NIST16), CASIA [43] and Columbia [44] dataset. We pre-train our model on the dataset from [45]. To fine-tune EAN, we follow the same training and testing split on NIST16, CASIA, and Columbia as in [5], [3]and [45] for fair comparisons. The information for each dataset is summarized in Table I.

- **NIST16 [42].** NIST16 is a dataset containing three manipulation techniques, the dataset's images are post-processed with high image clarity and provide authentic images with masks of manipulated iamges.
- **CASIA [43].** The CASIA dataset provides images for two main types of operations: splicing and copy-and-move. Similarly, some post-processing techniques such as filtering and blurring are applied. And the ground-truth mask can be obtained by thresholding the difference between the processed image and the authentic image.
- **Columbia [44].** The Colombia dataset contains 180 manipulated image data with a high degree of clarity and more visible manipulation marks.

Implementation. EAN is implemented in PyTorch and trained on an NVIDIA A800 GPU. The size of input images is 512×512. The backbone ResNet34 used in EAN is initialized with ImageNet-pretrained counterparts. We use the SGD optimizer with a learning rate that periodically decays from 10^{-3} to 10^{-7}.

Evaluation Metrics. We evaluated the performance of the proposed EAN on image manipulation localization tasks. We analyzed the manipulation masks using the Area Under the Curve (AUC) and F_1 score at pixel level.

B. Comparison with State-of-the-art

To demonstrate the overall advantages of our proposed EAN, we first compared it with several state-of-the-art baseline methods. In order to make a fair comparison, F_1 and AUC metrics are used to measure the differences between the ground truth mask and the predicted mask. The reported results of all compared methods are sourced either from their original papers or obtained by running their publicly available codes. All experiments were tested over three datasets: NIST16 [42], CASIA [43], and Columbia [44].

Baseline Models. We compare our EAN method with various baseline models as described below:

- RGB-N [3]: RGB-N combines the RGB and noise branches to detect manipulations and noise inconsistencies in images.
- MantraNet [4]: The image manipulation detection problem is defined as a local anomaly detection problem, and an evaluation Z-Score is designed to describe local anomalies.
- SPAN [5]: SPAN simulates the relationship between image blocks at various scales by constructing a pyramid of local self-attention blocks. It also introduces a new

TABLE II
COMPARISON AGAINST EXISTING METHODS OVER THREE STANDARD DATASETS, NIST16, CASIA AND COLUMBIA. [THE * REPRESENTS A MODEL THAT HAS NOT BEEN PRE-TRAINED]

Methods	Datasets							
	NIST16		CASIA		COLUMBIA		Average	
	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1
RGB-N [3]	0.937	0.722	0.613	0.408	0.581	0.612	0.710	0.581
MantraNet [4]	0.795	-	0.817	-	0.824	-	0.812	-
SPAN [5]	0.836	0.582	0.814	0.382	0.936	0.815	0.862	0.593
MVSSNET [6]	0.628	0.737	0.534	0.452	0.719	0.703	0.627	0.631
TDA-Net [46]	0.948	0.756	0.831	0.582	0.892	0.735	0.890	0.691
ImageForensicsOSN [47]	0.783	0.332	0.873	0.509	0.862	0.707	0.839	0.516
SATFL [48]	0.937	0.613	0.762	0.359	**0.999**	0.983	0.899	0.652
MFFAES [7]	0.954	0.854	0.701	0.456	0.962	0.833	0.872	0.714
EMT-Net [49]	0.98	0.825	0.856	0.459	0.900	0.561	0.914	0.615
DS-UNet [50]	0.828	0.772	0.749	**0.610**	0.977	0.940	0.851	0.774
ERMPC* [51]	0.895	0.836	**0.876**	0.586	-	-	0.886	0.711
$Ours^*$	0.980	0.819	0.832	0.521	**0.999**	0.975	0.937	0.772
$Ours$	**0.991**	**0.858**	0.855	0.528	**0.999**	**0.987**	**0.948**	**0.791**

position projection based on the transformer to encode the spatial position of the patch.

- MVSS-Net [6]: MVSS-Net uses multi-view feature learning and multi-scale monitoring for manipulation detection.
- TDA-Net [46]: TDA-Net consists of three convolutional network branches to extract three types of features from the spatial and frequency domains: visual perception, resampling, and local inconsistency.
- ImageForensicsOSN [47]: ImageForensicsOSN exploits various frequency feature from Online Social Networks (OSNs) to improve the proposed model's robustness.
- SATFL [48]: For localizing forged areas in suspicious images, SATFL uses the Channel-Wise High-Pass Filter Block (CW-HPF) self-attention mechanism.
- MFFAES [7]: An novel manipulation localization network architecture has been developed to efficiently segment the manipulated regions from a suspicious image.
- EMT-Net [49]: A novel network is proposed for the learning and enhancement of multiple manipulation traces (EMT-Net), including noise distribution and visual artifacts.
- DS-UNet [50]: A dual-stream U-network that combines the advantages of U-networks and residual networks is proposed for pixel-level localisation in image manipulation detection tasks.
- ERMPC [51]: A two-step Edge-aware Regional Message Passing Controlling strategy is proposed to address the image manipulation localization.

Table II shows the localization performance of different methods using pixel-level AUC and F_1 on the three standard datasets, where '-' represents an unreported value.

Model Without Pre-training. As deep learning models increase in complexity and scale, the significance of large-scale datasets for effective training has become even more evident. However, not everyone has the resources or infrastructure to pre-train models on such extensive datasets due to the significant computational requirements and associated costs. So we train and test standard datasets on models that have not been pre-trained on large datasets for reference. Table II shows that the model performs well even without pre-training on large-scale datasets. In particular, the best AUC score is obtained in the Columbia and NIST16 datasets, with **99.9%** and **98.0%**, respectively. A possible explanation is that the manipulation techniques used in the Columbia dataset are simplistic and lack post-processing, leading to easily detectable traces of manipulation. As for NIST16, good performance was achieved thanks to the model's excellent learning ability even if the model is not pre-trained.

Fine-tuned Model. The pre-trained model's network weights are utilized to initialize the fine-tuned models, which will undergo fine-tuning and testing on the NIST16, CASIA, and Columbia datasets. It can be observed that the performance of our method is the best on the whole. Specifically, our method significantly outperformed other methods over NIST16 and Columbia. Especially, EAN achieves 99.9% AUC on Columbia dataset. We argue that other methods can capture the manipulation artifacts, but they don't pay enough attention to edge information, which leads to low accuracy in locating the manipulated region. Moreover, the refined attention mechanism amplified the recognition degree of edge artifacts, ensuring the fitness of the predicted binary mask. So, recurrent attentional enhancement of specific features that accurately locate manipulated regions plays a crucial role in mitigating the negative effects of feature redundancy. But we fail to achieve the best performance in CASIA. This may be due to the fact that the current mainstream practice is to train with the CASIA2.0 version and test with the CASIA1.0, but there are many images with different manipulation methods. As a result, EAN fails to learn enough artifacts. TDA-Net, ERMPC, and DS-Unet perform excellently on CASIA dataset owning to pre-train their model with the synthetic dataset being made by themselves, which is similar to the CASIA dataset.

C. Robustness Evaluation

In this section, we conducted a series of experiments to assess the robustness of our method. Firstly, we applied

TABLE III
ROBUSTNESS EVALUATION USING AUC FOR MANTRANET [4], SPAN [5], MSFFAES [7] AND OUR METHOD. FOUR OPERATION ATTACK METHODS, RESIZE, GAUSSIAN BLUR, GAUSSIAN NOISE, JPEG COMPRESSION, ARE USED TO SIMULATE ATTACK MODELS, IN WHICH TWO SCALING FACTORS 0.78× AND 0.25× FOR RESIZE ATTACK, TWO KERNELS k = 3 AND 15 FOR GAUSSIAN BLUR ATTACK, TWO NOISE INTENSITY FACTORS σ = 3 AND 5 FOR GAUSSIAN NOISE ATTACK AND TWO QUALITY FACTORS QF = 100 AND 50 FOR JPEG COMPRESSION ATTACK. N DENOTES NIST16 DATASET AND C DENOTES COLUMBIA DATASET. [THE * MEANS THE MODEL TRAINED WITHOUT DATA ENHANCEMENT.]

Image Processing	Parameter	Method									
		MantraNet		SPAN		MSFFAES		*Ours**		*Ours*	
		N	C	N	C	N	C	N	C	N	C
w/o distortion	-	0.780	0.779	0.839	0.936	0.954	0.962	**0.9887**	**0.9994**	0.9673	0.9846
Resize	0.78×	0.774	0.690	0.832	0.899	0.931	0.920	**0.9881**	**0.9981**	0.9675	0.9841
Resize	0.25×	0.755	0.686	0.803	0.690	0.815	0.757	**0.9756**	0.9120	0.9650	**0.9708**
Gaussian Blur	k = 3	0.774	0.677	0.831	0.789	0.923	0.884	0.9538	0.9807	**0.9655**	**0.9814**
Gaussian Blur	k = 15	0.745	0.628	0.791	0.677	0.881	0.807	0.8547	0.8477	**0.9605**	**0.9762**
Gaussian Noise	σ=3	0.674	0.682	0.751	0.751	0.889	0.867	0.9584	0.9610	**0.9637**	**0.9853**
Gaussian Noise	σ=5	0.585	0.549	0.672	0.658	0.702	0.756	0.9426	0.9219	**0.9648**	**0.9842**
JPEG Compression	QF = 100	0.779	0.750	0.835	0.933	0.925	0.944	**0.9835**	**0.9984**	0.9664	0.9834
JPEG Compression	QF = 50	0.743	0.593	0.806	0.746	0.813	0.848	0.9426	**0.9838**	**0.9508**	0.9815

four different post-processing methods to the images in the Columbia dataset: (1) Scaling factors with different proportions, (2) Gaussian blur with varying kernel size k, (3) Gaussian noise with varying standard deviation σ, and (4) JPEG compression with varying compression quality QF. Subsequently, by utilizing AUC as an evaluation metric, we compared the detection results obtained by our method against other networks.

As presented in Table III, the table demonstrates that EAN exhibits superior robustness in various post-processing operations compared to other neural networks. In general, images are already resized when they are uploaded on social applications. For the NIST16 dataset, when the scale factor decreased from 0.75 to 0.25, the scores dropped by 1.25% for the model trained with data augmentation and by 0.25% for the model trained without it. It is worth noting that resizing has minimal impact on EAN is benefit from multi-scale feature extraction, which suggests that EAN is likely to excel in detecting real manipulated images on social networks. Besides, noise features commonly used in other networks are easily destroyed after post-processing operations. It indicates that the AUC score of the model before data enhancement training decreased faster. When Gaussian noise with σ values of 3 and 15 is introduced into the images from the NIST16 dataset, the AUC scores of EAN decrease by 3.31% and 13.40% respectively, whereas MASFFAES experiences a greater decline, meaning that its associated model obtains poorer robustness. Additionally, Gaussian blur results in a 4.8% reduction in performance for SPAN. Although EAN training without post-processing method has also declined, EAN training with post-processing only decreased by 0.05% and 0.5% for $k = 3$ and $k = 15$ respectively, hardly declined at all. Both the EAN trained without data enhancement and other networks experienced a decline in AUC scores after image post-processing. In contrast, the proposed model trained with data enhancement, despite starting with a lower initial score, exhibited only a marginal decrease in performance. This resilience to post-processing effects underscores the efficacy of our method in enhancing the model's robustness and adaptability. It can be analyzed from the results that the edge-aware mechanism is crucial for model's robustness, the proposed attention blocks plays a significant role in combating the post-processing and feature redundancy problems.

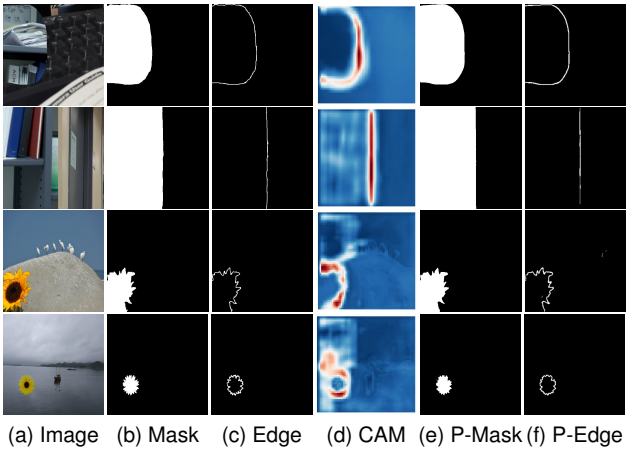

Fig. 5. Sample visual results of the proposed EAN. From left to right, we show manipulated images, ground-truth binary masks of manipulated image, ground-truth binary masks of boundary, CAM, predicted mask and boundary of the proposed EAN.

D. Visualization Results

We ran a series of experiments on different datasets, and some of the visualizations are shown in Fig. 6. Even in the face of a complex foreground, our model is able to accurately locate the manipulated area. It is observed that the predictions of others methods have a certain amount of false positives, like snowflakes. Thanks to the edge-aware mechanism, the prediction boundary of EAN is smooth and extremely accurate. In Fig. 5, we use gradient-weighted class activation maps (GradCAM) [52] to visualize the generated channel attention graph. After the attention module adaptively adjusts the weight distribution, the fused feature pays more attention to the edge and even extends to both sides of the boundary depending on the edge-aware mechanism and

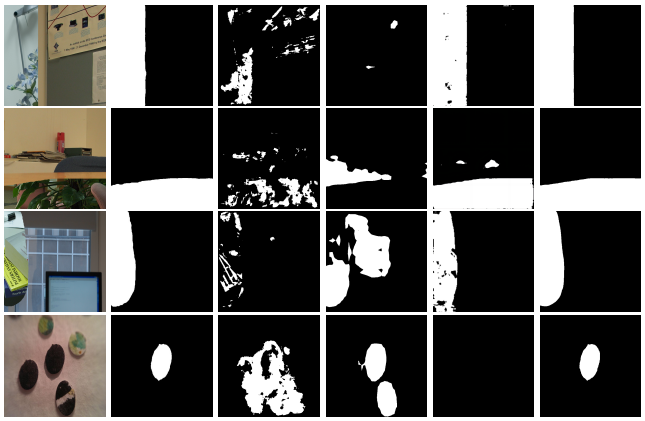

Fig. 6. Sample qualitative results of the proposed EAN compared with three SoTA methods. The images are displayed from left to right in the following order: manipulated images, ground-truth binary masks of manipulation, predictions of ManTra, MVSS, PSCC, and the proposed EAN.

feature enhancement. This enables the model to discern the distinctions between manipulated and authentic regions more effectively, thereby confirming the effectiveness of both the edge-aware module and feature fusion module.

E. Ablation Study

The NFE is designed to extract operational artifacts from the frequency domain, while the EAM captures the edge artifacts between manipulated and authentic regions by directing the encoder to focus on edge features. At last, the FFM integrates features from three components to generate a more precise mask. To reveal the influence of each component, we evaluate the performance of the proposed model in different setups where the components are added progressively on NIST16 and Columbia datasets.

TABLE IV
ABLATION STUDY.

Components	Columbia		NIST16	
	AUC	F_1	AUC	F_1
RFE	0.887	0.581	0.914	0.418
RFE+NFE	0.977	0.905	0.966	0.723
RFE+NFE+EAM	0.998	0.973	0.977	0.772
RFE+NFE+EAM+FFM	**0.999**	**0.975**	**0.980**	**0.819**

The quantitative results are listed in Table IV. We observe that with the introduction of NFE, the AUC scores and F_1 scores for Columbia increased by 9.0% and 32.4%, respectively. Similarly, for NIST16, there was an improvement of 6.3% in AUC and 38.8% in F_1 scores. This performance enhancement underscores the significance of frequency domain features in detecting forged regions.

On the other hand, the AUC scores with EAM for Columbia and NIST16 increased by 2.1% and 1.1% respectively, while F_1 scores improved by 6.8% and 4.9%. These improvements validate that EAM strengthens attention to boundaries, effectively boosting our model's performance. Furthermore, the FFM, derived from a comprehensive attention mechanism, demonstrates performance improvement, notably on the NIST16 dataset with a 0.3% increase in AUC and a 4.7% increase in F_1 scores. This further attests to the effectiveness of the integrated attention mechanism against feature redundancy.

F. Limitations and Future Work

Although our proposed model achieves excellent performance in the image manipulation localization task, there are still some problems. Firstly, the proposed method is to direct the network to focus on edge feature through the edge-aware mechanism, so the accuracy of the reference boundary mask is significant. Our boundary reference mask is computed using the Sobel operator [53]. However, the boundary masks generated by this approach may be discontinuous and relatively rough, potentially overlooking some subtle edges. In this situation, the boundary prediction mask generated by the EAN may resemble the reference mask, ultimately resulting in suboptimal performance. To improve this, we plan to find a novel approach to obtain an accurate boundary reference mask. Secondly, although edge awareness improves the result, the reference boundary mask contains too little information. We believe that extracting more artifacts from other domains can solve this problem. Therefore, We will explore better methods to obtain edge information and extract artifacts from other domains.

VI. CONCLUSION

We introduce an edge-aware network, a novel framework for image manipulation localization. EAN treats the image manipulation localization problem as a segmentation problem inside and outside the boundary. We first utilize convolutional neural networks to extract the RGB and noise features. Benefiting from the skip connection among different layers, we get features that are a fusion of features from different hierarchical levels. Then, implicit and explicit edge features from RFE and NFE interact with each other to form integrated edge features in the proposed edge-aware module, which is used to effectively capture edge artifacts and direct network focus on edge region. Thanks to the edge-aware mechanism and elaborate spatial attention block, the encoder not only pays more attention to the boundary regions but also reduces the risk of feature redundancy. Furthermore, three different features were fed into the feature fusion module, and further integrated by enhanced self-attention block and spatial attention block. Experimental result shows that the validity of our framework construction is verified and the best overall performance is achieved compared to existing state-of-the-art methods.

VII. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62172098, 62471141; in part by the Natural Science Foundation of Fujian Province under Grant 2020J01497.

REFERENCES

[1] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 101–117.

[2] C.-T. Li and Y. Li, "Color-decoupled photo response non-uniformity for digital image forensics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 2, pp. 260–271, 2011.

[3] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1053–1061.

[4] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9543–9552.

[5] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "Span: Spatial pyramid attention network for image manipulation localization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, Springer. Cham: Springer International Publishing, 2020, pp. 312–328.

[6] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3539–3553, 2022.

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

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Cham: Springer, 2015, pp. 234–241.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[11] T. Bianchi, A. De Rosa, and A. Piva, "Improved dct coefficient analysis for forgery localization in jpeg images," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2444–2447.

[12] X. Zhao, S. Wang, S. Li, and J. Li, "Passive image-splicing detection by a 2-d noncausal markov model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 2, pp. 185–199, 2015.

[13] D.-Y. Huang, C.-N. Huang, W.-C. Hu, and C.-H. Chou, "Robustness of copy-move forgery detection under high jpeg compression artifacts," *Multimedia Tools and Applications*, vol. 76, pp. 1509–1530, 2017.

[14] J. Zhao and J. Guo, "Passive forensics for copy-move image forgery using a method based on dct and svd," *Forensic science international*, vol. 233, no. 1-3, pp. 158–166, 2013.

[15] L. Li, S. Li, H. Zhu, S.-C. Chu, J. F. Roddick, and J.-S. Pan, "An efficient scheme for detecting copy-move forged images by local binary patterns." *J. Inf. Hiding Multim. Signal Process.*, vol. 4, no. 1, pp. 46–56, 2013.

[16] M. Bashar, K. Noda, N. Ohnishi, and K. Mori, "Exploring duplicated regions in natural images," *IEEE Transactions on Image Processing*, 2010.

[17] L. Su, C. Li, Y. Lai, and J. Yang, "A fast forgery detection algorithm based on exponential-fourier moments for video region duplication," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 825–840, 2017.

[18] A. Costanzo, I. Amerini, R. Caldelli, and M. Barni, "Forensic analysis of sift keypoint removal and injection," *IEEE transactions on information forensics and security*, vol. 9, no. 9, pp. 1450–1464, 2014.

[19] Y. Zhu, X. Shen, and H. Chen, "Copy-move forgery detection based on scaled orb," *Multimedia Tools and Applications*, vol. 75, pp. 3221–3233, 2016.

[20] Y. Zeng, B. Zhao, S. Qiu, T. Dai, and S.-T. Xia, "Toward effective image manipulation detection with proposal contrastive learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4703–4714, 2023.

[21] L. Bondi, S. Lameri, D. Guera, P. Bestagini, E. J. Delp, S. Tubaro *et al.*, "Tampering detection and localization through clustering of camera-based cnn features." in *CVPR Workshops*, vol. 2, 2017, p. 2.

[22] Y. Zhang, G. Zhu, L. Wu, S. Kwong, H. Zhang, and Y. Zhou, "Multi-task se-network for image splicing localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4828–4840, 2021.

[23] D. Cozzolino, G. Poggi, and L. Verdoliva, "Efficient dense-field copy–move forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[24] L. D'Amiano, D. Cozzolino, G. Poggi, and L. Verdoliva, "A patchmatch-based dense-field algorithm for video copy–move detection and localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 669–682, 2018.

[25] A. Islam, C. Long, A. Basharat, and A. Hoogs, "Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4676–4685.

[26] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "Coverage—a novel database for copy-move forgery detection," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 161–165.

[27] Y. Wu, W. Abd-Almageed, and P. Natarajan, "Busternet: Detecting copy-move image forgery with source/target localization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 168–184.

[28] H. Wu and J. Zhou, "Iid-net: Image inpainting detection network via neural architecture search and attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1172–1185, 2021.

[29] Q. Yang, D. Yu, Z. Zhang, Y. Yao, and L. Chen, "Spatiotemporal trident networks: Detection and localization of object removal tampering in video passive forensics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4131–4144, 2020.

[30] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun, "A deep learning approach to patch-based image inpainting forensics," *Signal Processing: Image Communication*, vol. 67, pp. 90–99, 2018.

[31] F. Li, H. Zhai, T. Liu, X. Zhang, and C. Qin, "Learning compressed artifact for jpeg manipulation localization using wide-receptive-field network," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[32] F. Li, H. Zhai, X. Zhang, and C. Qin, "Image manipulation localization using spatial-channel fusion excitation and fine-grained feature enhancement," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2364–2373.

[38] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, 2018.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[40] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[41] Q. Wei, X. Li, W. Yu, X. Zhang, and Y. Chen, "Learn to segment retinal lesions and beyond," in *25th International Conference on Pattern Recognition (ICPR)*, 2020.

[42] N. Nimble, "Datasets," 2016.

[43] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *2013 IEEE China summit and international conference on signal and information processing*. IEEE, 2013, pp. 422–426.

[44] T.-T. Ng, J. Hsu, and S.-F. Chang, "Columbia image splicing detection evaluation dataset," *DVMM lab. Columbia Univ CalPhotos Digit Libr*, 2009.

[45] X. Liu, Y. Liu, J. Chen, and X. Liu, "Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7505–7517, 2022.
[46] S. Li, S. Xu, W. Ma, and Q. Zong, "Image manipulation localization using attentional cross-domain cnn features," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
[47] H. Wu, J. Zhou, J. Tian, and J. Liu, "Robust image forgery detection over online social network shared images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 440–13 449.
[48] L. Zhuo, S. Tan, B. Li, and J. Huang, "Self-adversarial training incorporating forgery attention for image forgery localization," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 819–834, 2022.
[49] 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 Recognition*, vol. 133, p. 109026, 2023.
[50] Y. Huang, S. Bian, H. Li, C. Wang, and K. Li, "Ds-unet: A dual streams unet for refined image forgery localization," *Information Sciences*, vol. 610, pp. 73–89, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025522008866
[51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8222–8232.
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[53] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.

Yun Chen is currently pursuing a master's degree in the School of Computer Science and Big Data at Fuzhou University, His current research interests include multimedia security.

Hang Cheng received his BS and MS degrees in applied mathematics from Fuzhou University, Fuzhou, China, in 2002 and 2005, respectively, and PhD in signal and information processing with Shanghai University, Shanghai, China, in 2016. He is a Professor in the School of Mathematics and Statistics, Fuzhou University, Fuzhou, China. He is a research scholar in the School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. His current research interests include multimedia security, image processing, cryptography, and information hiding.

Haichou Wang is currently pursuing a master's degree in the School of Mathematics and Statistics at Fuzhou University. His current research interests include image manipulation localization.

Ximeng Liu (Senior Member, IEEE) received the B.E. degree with the Department of Electronic Engineering from Xidian University, Xi'an, China, in 2010 and Ph.D. degree with the Department of Telecommunication Engineering from Xidian University, Xi'an, China in 2015. He is a postdoctoral fellow with the Department of Information System, Singapore Management University, Singapore. And he is also a Professor in the School of Computer Science and Big Data, Fuzhou University, Fuzhou, China. His research interests include applied cryptography and big data security.

Fei Chen (Member, IEEE) received the Ph.D. degree in signal and information processing from Zhejiang University, Hangzhou, China, in 2013. He is currently an Associate Professor with the College of Computer Science and Big Data, Fuzhou University, Fuzhou, China. His current research interests include machine learning, computer vision, and deep learning techniques in image processing.

Fengyong Li received the M.S. degree from the School of Information and Engineering in 2010 from Zhengzhou University, and the Ph.D. degree in School of Communication and Information Engineering in 2014 from Shanghai University. He is currently an Associate Professor of Shanghai University of Electric Power. He worked as a visiting scholar in University of Victoria, B.C. Canada from 2018 to 2019. His research interests include multimedia security, information hiding, and machine learning. He has published more than 50 peer-reviewed papers.

Xinpeng Zhang (Senior Member, IEEE) received the B.S. degree in computational mathematics from Jilin University, China, in 1995, and the M.E. and Ph.D. degrees in communication and information system from Shanghai University, China, in 2001 and 2004, respectively. Since 2004, he has been with the faculty of the School of Communication and Information Engineering, Shanghai University, where he is currently a Professor. He is also with the faculty of the School of Computer Science, Fudan University. He was a visiting scholar with The State University of New York at Binghamton from 2010 to 2011, and also with Konstanz University, as an experienced Researcher, sponsored by the Alexander von Humboldt Foundation, from 2011 to 2012. His research interests include multimedia security, image processing, and digital forensics. He has published over 200 articles in these areas. He has served as an Associate Editor for the IEEE Transactions on Information Forensics and Security from 2014 to 2017.

Meiqing Wang received her BS and MS degrees in applied mathematics from Tsinghua University, Beijing, China, in 1987 and 1989, respectively, and PhD in Department of Computing, Xi'an Jiaotong University, China, in 2002. Now, she is a Professor in the School of Mathematics and Statistics, Fuzhou University, Fuzhou, China. Her current research interests include computing science, image processing, and computational finance.