Multi-scale noise-guided progressive network for image splicing detection and localization

Dengyong Zhang[a,b], Ningjing Jiang[a,b], Feng Li[a,b], Jiaxin Chen[a,b], Xin Liao[c,*], Gaobo Yang[c], Xiangling Ding[d]

[a]*Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China*

[b]*School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China*

[c]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*

[d]*School of Computer and Communication Engineering, Hunan University of Science and Technology, Xiangtan, 411201, Hunan, China.*

Abstract

Image splicing can easily be used in illegal activities, which may have negative impact on society. Therefore, detecting splicing images and precisely localizing tampered regions are challenging tasks. In this work, we propose a dual-branch Multi-Scale Noise-Guided Progressive Network (MSNP-Net). The multi-resolution branch extracts deep semantic features of images while suppressing redundant noise. A multi-scale noise-guided branch is designed to capture more subtle tampering artifacts and guide the network to strengthen the learning of spatial structure features. The two branches complement and restrict each other. The features of different scales are fused in a gradual mechanism, and the feature expression is aggregated through the spatial channel feature aggregation module (SCAM). A large number of experiments show that the MSNP-Net proposed in this paper has better completeness and lower false alarm rate for the detection and localization of spliced images.It can achieve more refined detection results, and has good stability and robustness. Significantly outperforms other state-of-the-art methods. Our source code is available at https://github.com/Swag-Jiang/MSNP-Net.

Keywords: image splicing localization, progressive network, multi-scale noise-guided, feature aggregation

1. Introduction

With the widespread use of powerful digital image editing tools, image content can be re-edited and modified very easily. While image editing technology brings convenience to people, the problems it brings are challenging to ignore. Especially in this Internet era, once manipulated images start to spread on the Internet, the spread range is so extensive, and the speed of spread is so fast that it can almost reach the trend of getting out of control. In recent years, more and more manipulated images have been misused in various fields, such as scientific research, news media, judicial evidence collection, and the economy, causing severe adverse effects [1].

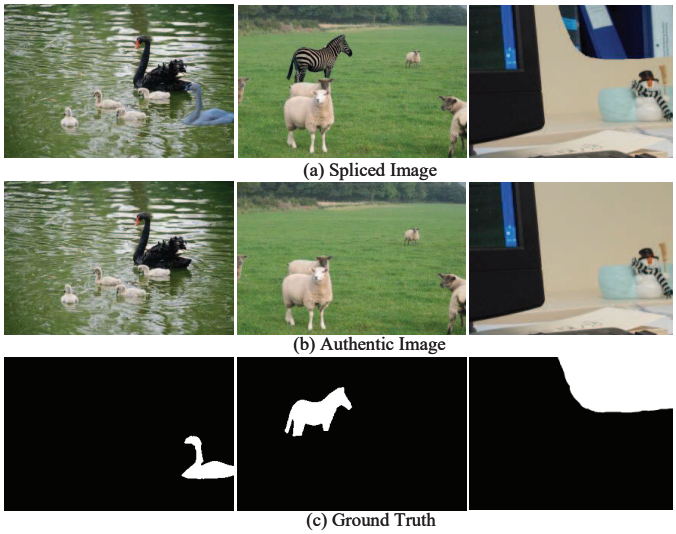
Image splicing is one of the most common means of image manipulation. It copies a specific region from a source image and pastes it on the host image [2], as shown in Fig. 1. Sometimes geometric transformations such as flipping and scaling are performed on the tampered regions before pasting to make the spliced image more realistic. Image splicing changes the semantic content of the image and destroys the authenticity, integrity, and originality of the image content. Therefore, image splicing forgery detection and localization has received extensive attention and research.

*Corresponding author

Email addresses: zhdy@csust.edu.cn (Dengyong Zhang), 21208051580@stu.csust.edu.cn (Ningjing Jiang), lif@csust.edu.cn (Feng Li), jxchen@csust.edu.cn (Jiaxin Chen), xinliao@hnu.edu.cn (Xin Liao), yanggaobo@hnu.edu.cn (Gaobo Yang), xianglingding@163.com (Xiangling Ding)

All images will form inherent attributes of the image after processing and post-processing by imaging equipment, such as lighting [3], shadows, noise patterns [4], camera features [5], etc. These intrinsic attribute features are independent of the image content and are called image fingerprints. In splicing images, since the tampered regions and the untampered regions come from different source images, the attribute difference between the tampered regions and the untampered regions can be used to detect splicing images and locate the spliced regions. Researchers have proposed many feature extraction methods in the past few years to explore the differences among image attributes as clues to detect splicing forgery images. It can be roughly divided into traditional feature extraction methods [3–7] and convolutional neural network-based methods [8–11]. The traditional forensics method first extracts features through manual design and then conducts a comparative analysis of the features or analyzes the statistical characteristics of the image to determine whether it has been manipulated and find out the spliced regions. However, the traditional image forensics method has certain defects. Most of the features based on manual design have limitations and lack of representativeness. The extracted features are only for a specific image attribute. The tampered image undergoes a series of post-processing, leading to the loss of this feature. The detection performance of traditional forensics methods will be significantly reduced and may even fail.

With the successful application of deep learning in various fields [12–14], since 2016, more and more researchers have tried to apply deep learning methods to the field of image foren-



(a) Spliced Image

(b) Authentic Image

(c) Ground Truth

Fig. 1. Example of image splicing forgery. (a) The spliced image; (b) The authentic image; (c) The ground truth.

sics [15–17]. Conventional computer vision tasks learn more about the semantic features of images, which runs counter to the task of passive image forensics. Image forgery forensics, especially image splicing forgery forensics, needs to pay attention to the subtle contrast changes of tampered edges and some tampering artifacts left by post-processing techniques, according to these tampering artifacts and statistical features [18–20] to locate the tampered regions. In order to avoid paying too much attention to the semantic content of the image and ignoring the subtle tampering artifacts, some researchers proposed a method of extracting features at the image block level [17], dividing the image into blocks of uniform size, and then feed into the feature extraction network to extract the features of each block. However, this method of using image blocks as network input loses the spatial information and context information of the image, which quickly leads to prediction errors. In order to obtain more global features, more and more methods obtain a large enough receptive field through continuous downsampling and stacking convolutional layers. However, too many downsampling operations will prevent a lot of detailed features of the image from being lost. The deepening of the network depth will also lead to the problem of gradient degradation. These works use a single kind of feature as input, which cannot take into account both semantic features and spatial structure features, and do not use subtle tampering artifacts as forensic clues, resulting in inaccurate localization of tampered regions.

Therefore, we design a two-branch network. In RGB domain, We use the high-resolution network (HR-Net) [21] as the backbone network, which maintains high-resolution representations throughout the process through parallel multi-resolution branches, and high-resolution feature maps can better preserve image subtle tampering artifacts, cross-fusion between feature maps of different resolutions can better model multi-scale changes. At the same time, we designed a multi-scale branch of the noise domain to perform multi-resolution feature information interaction in the spatial domain and the noise domain so that the network can pay attention to semantic information and spatial structure information simultaneously and better adapt to multi-scale changes. Using the characteristics of different sources of the tampered region and the untampered region of the spliced image, the statistical characteristics of the two parts show differences, and this inconsistency is evident in the noise image. Therefore, some researchers use the Spatial Rich Model (SRM) [22] to extract the noise of the image and send the noise and the original image to the subsequent network. However, since the weight of SRM is fixed, it cannot be learned and updated, so some works use constrained convolution [23] instead of SRM, but constrained convolution still has the problem of instability. Therefore, we use the Improved Bayar Convolution [24] to extract the noise image. Compared to previous works, the noise images we extracted do not go through the deep convolutional neural network because these shallow features have few features left after passing through the deep network, leaving more semantic features instead, which is inconsistent with our original intention. We send the extracted noise domain features into the noise-guided module (NGM), which is designed to enhance the feature representation in the noise domain, guide the network to learn the spatial structure information of the image and enhance feature differences between tampered and untampered regions through residual connections.

Overall, the main contributions of this work can be summarized as follows:

1. We propose a dual-branch network MSNP-Net. The noise branch can effectively capture low-level spatial structure features and guide the network to learn rich tampering artifacts. The multi-resolution branch captures deep semantic context, suppresses irrelevant redundant information, and generates effective features for good completeness and low false positives.

2. We design a new progressive mechanism to fuse multi-scale features from two branches, and incorporate several S-CAMs into the network, to recalibrate the fused features to aggregate feature representation.

3. Extensive experimental results on three publicly available image splicing datasets demonstrate that our network outperforms other state-of-the-art methods regarding detection and localization performance.

2. Related work

Over the years, researchers have proposed various visual forensics methods to identify various manipulation. In the field of image splicing forgery detection, the earliest detection methods are based on traditional feature extraction methods. People mainly use specific image fingerprints or intrinsic statistical features as forensic clues to detect tampered images, including Color Filter Array (CFA) interpolation mode [25, 26], image compression properties [27, 28], etc. Farid et al. [19] used the method of multi-scale wavelet decomposition to image for high-order statistical modeling. Shi et al. [20] used the multi-scale discrete cosine transform (MBDCT) and Markov model to capture local abnormal features, and then input the transition probability moments and MBDCT low-order moments of

the Markov model into SVM for image classification. However, these traditional methods rely too much on a specific image attribute. When the tampered image undergoes post-processing operations such as overall smoothing, blurring, and compression, it may cause some specific image fingerprints to fail, which in turn leads to the failure of the detection method.

With the development of deep learning, researchers have found that the architecture of the convolutional neural network (CNN) matches the tasks in computer vision very well, so many methods based on convolutional neural network have emerged in the field of image forgery detection. Ying Zhang et al. [29] first applied deep learning methods to image passive forensics in 2016. They proposed a block-level coarse localization method for spliced images based on wavelet features. Salloum et al. [30] applied fully convolutional neural networks to image splicing forensics tasks, and an MFCN framework for dual-branch multi-task learning was proposed, which achieved pixel-level fine-grained segmentation for the first time. Due to the smooth operation in the MFCN framework, the network often ignores some small objects. Bi et al. [31] added a ringed residual structure (RRU-Net) to U-Net and designed a residual propagation and residual feedback module to enhance feature expression. However, they only pay attention to the characteristic information of the RGB domain, and the localization of the tampered regions is not accurate enough. Xiao et al. [32] proposed a coarse-to-fine network (C2R-Net) to roughly and finely locate the tampered regions. Then they use an adaptive clustering algorithm to determine the final and accurate tampered regions. This approach is not end-to-end and requires clustering and filling algorithms for post-processing operations. Minyoung Huh et al. [33] innovatively integrated a self-learning method into image splicing forgery detection. The model uses EXIF metadata in JPEG compression as a forensic clue to predict tampered region masks based on the similarity between different image patches. Kwon et al. [34] combined deep learning methods and traditional features to design an end-to-end fully convolutional neural network (CAT-Net) with RGB and DCT streams to jointly learn the forensics of JPEG compression artifacts in RGB and DCT domains feature. Although these methods use convolutional neural networks to extract features, they still have a certain dependence on specific image fingerprints. Edge and noise features are significantly different between tampered and untampered regions of spliced images, and there are many methods to use them as forensic clues. Zhou et al. [35] fed the noise features extracted by SRM and features extracted from input images into Faster R-CNN [36] and trained them end-to-end to detect tampered regions of images. Zhang et al. [37] propose a multi-task Squeeze and Excite Network (SE-Network) for splicing localization, which utilizes the edges of the spliced images, the ground-truth mask, and the edge of the mask to guide the learning of the label mask, The global edge information can be fully utilized to provide more comprehensive supervision for the localization of spliced regions.

3. Proposed method

We propose a model named MSNP-Net, as shown in Fig. 2. MSNP-Net consists of two parallel branches that process different feature information: 1) Multi-resolution feature extraction branch (MREB), which inputs RGB images into the multi-resolution branch with HR-Net as the backbone to obtain different resolutions of RGB features f_i; 2) Multi-scale noise-guided branch (MSNB), which uses the Improved Bayar Convolution to extract the noise map of the picture and inputs it to the noise-guided module (NGM) to obtain noise features of three scales N_i. Then the same resolution RGB domain features and noise domain features obtained by the two branches are simultaneously input into our proposed dual-domain progressive fusion module (DPFM), and the features of the two parts are fused from low-scale to high-scale in a progressive manner to obtain the final prediction mask.

3.1. Multi-resolution feature extraction branch

Most deep convolutional neural networks obtain high resolutions by reducing the resolution and then increasing the resolution. For example, U-Net, SegNet, and DeconvNet are essentially this structure. These existing methods transfer input through the network, usually connecting feature maps of different resolutions in series, ignoring the interaction between features of different scales, inaccurately grasping global features, and losing high-resolution features. In this process, the pooling operation is used to reduce the resolution so that many details are lost. Recent research has shown that pooling is not advisable for tasks requiring subtle signals because it enhances content and suppresses noise-like signals [38]; it is undesirable for forensics tasks. Especially in image splicing forgery forensics, the noise domain contains many detailed features, which is a valuable forensic clue.

Based on this, we choose the HR-Net [21] as the backbone network. It is able to maintain a high-resolution representation throughout the process, starting with a high-resolution subnetwork as the first stage and gradually adding high-resolution to low-resolution sub-networks one by one, forming more stages. The multi-resolution sub-networks are connected in parallel, and repeated multi-scale fusion is performed by repeatedly exchanging information between the parallel multi-resolution subnetworks throughout the process. This fits well with the main task of our network, which can fully utilize global and local features, refine the localization of spliced edges and irregular spliced regions, and capture the overall structure without losing the fine details required for forensic investigations.

Instead of using the last feature fusion unit of HR-Net, we make full use of the features of each scale. In order to achieve a balance between performance and calculation, in the first resolution branch, a convolution unit is used to reduce the resolution of the original image by four times. Then the resolution is reduced step by step, and each resolution is kept until the end. Get the feature maps of the four scales of the RGB domain $f_1 \in \mathbb{R}^{C_1 \times \frac{H}{4} \times \frac{W}{4}}$, $f_2 \in \mathbb{R}^{C_2 \times \frac{H}{8} \times \frac{W}{8}}$, $f_3 \in \mathbb{R}^{C_3 \times \frac{H}{16} \times \frac{W}{16}}$, $f_4 \in \mathbb{R}^{C_4 \times \frac{H}{32} \times \frac{W}{32}}$. Among them, C_1, C_2, C_3, and C_4 are 48, 96, 192, and 384 respectively, which is the same in the following text. The follow-

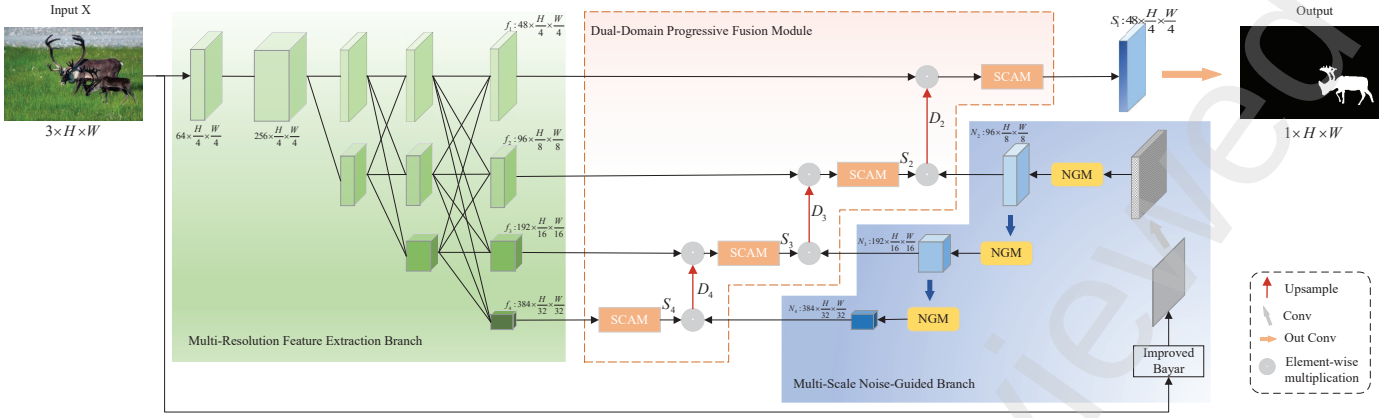


Fig. 2. The network architecture of MSNP-Net.

up will be combined with the noise domain feature map of the same resolution to progressively generate the final feature map from bottom to top and get the final prediction mask.

3.2. Multi-scale noise-guided branch

In the RGB domain, the convolutional neural network tends to focus on the semantic features of the image rather than the spatial features, which will cause the network to locate the tampered region of the spliced image inaccurately, and the edge of the tampered region is blurred. Studies have shown that these spatial structure features are better expressed in the noise domain, so we designed a multi-scale noise-guided branch (MSNB) using the Improved Bayar Convolution to extract noise images from RGB images as the input of the branch. Our proposed noise-guided block further enhances the spatial structure feature expression of tampered regions and edges in the noise domain. In order to match the multi-scale changes in the spatial domain and thoroughly combine the global and local features. Correspondingly, we extract the multi-scale features in the noise domain $N_1 \in \mathbb{R}^{C_1 \times \frac{H}{4} \times \frac{W}{4}}$, $N_2 \in \mathbb{R}^{C_2 \times \frac{H}{8} \times \frac{W}{8}}$, $N_3 \in \mathbb{R}^{C_3 \times \frac{H}{16} \times \frac{W}{16}}$, $N_4 \in \mathbb{R}^{C_4 \times \frac{H}{32} \times \frac{W}{32}}$ to guide the network to balance better the deep semantic features of the network and the shallow space structural features, which are more stable against tampering at different scales.

3.2.1. Noise extractor

Some initial works began to try to use SRM [22] to obtain the noise domain representation of the image and use the noise image as the input of the model. However, the weight of SRM is predefined and cannot be learned and updated. Some studies propose using constrained convolution [23] instead of SRM. The constraint rules of constrained convolution are as follows:

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

Where $w_k(c,c)$ represents the weight of the center position of the kth convolution kernel, and (m,n) represents the non-center position coordinates. Constrained convolution imposes constraints after the kernel function updates the weights. After the

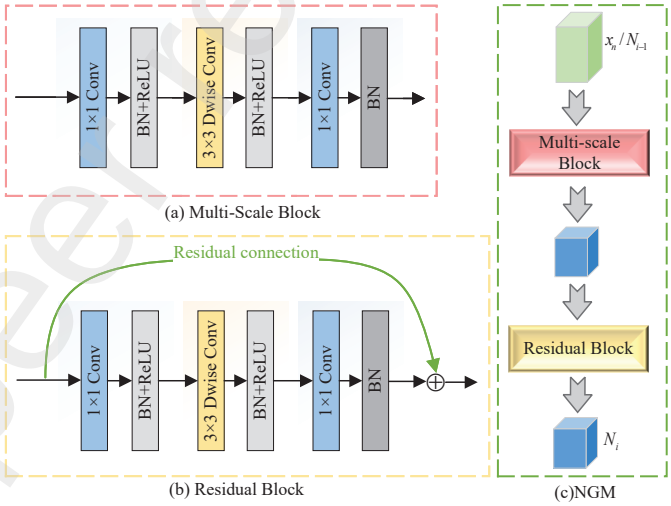


Fig. 3. The noise-guided module consists of a multi-scale block and a residual block. (a) Multi-Scale block; (b) Residual Block; (c) Noise-Guided Block.

constraints, the weight of the central position of the convolution kernel is −1, and the sum of the weights of other positions is 1. According to this constraint rule, if the weights of the convolution kernel after backpropagation are mixed with positive and negative weights, the absolute value of their sum may be minimal, and dividing the weights of all non-central positions by this sum will cause the weights to be amplified many times, or their sum is a negative value, the weight of the non-central position divided by a negative value will cause the positive and negative to be reversed, resulting in a massive change in the input of the model. These unstable factors in the training process will cause fluctuations, resulting in poor training results.

The Improved Bayar Convolution [24] has made the following improvements on the basis of constrained convolution:

1. Divide the weights of the non-central locations by the absolute value of the sum of these weights. $w_k(m,n) = w_k(m,n) \div S_k$, where $S_k = \sum_{m,n \neq c} |w_k(m,n)|$.

2. If the weight of the non-central position is $w_k(m,n) \leq 0.001$, set its value to 0.001.

3. Set the center position's weight to the inverse of the sum

of the other positions' weights. $w_k(c,c) = -S_k$.

Due to the above reasons, we choose to use the Improved Bayar Convolution as our noise extractor. These improvements can effectively smooth the fluctuations in the training process while ensuring that the weights can be learned. Subsequent experiments prove that the performance of choosing the Improved Bayar Convolution as the noise extractor is optimal.

3.2.2. Noise-guided module

The noise-guided module (NGM) takes noise features of different scales as input, and enhances the feature representation in the noise domain through NGM. In the process of combining with multi-resolution semantic features, the network is gradually guided to pay attention to detail tampering artifacts, so as to balance semantic and spatial structure information. This makes the localization of spliced regions more precise.

NGM consists of a multi-scale block (MSB) and a residual block (RB) concatenated. As shown in Fig. 3. Among them, the multi-scale block is used to adapt to the multi-scale changes of the network, as shown in Fig. 3(a). The residual block strengthens feature learning and further enhances the feature expression of the tampered regions and edge, as shown in Fig. 3(b). In order to make full use of the feature information while reducing the amount of computation, we first increase the number of channels of the features by a factor of six using 1×1 convolutions, then use 3×3 depthwise convolution instead of standard convolution. Finally, the channel is reduced to the required size by a 1×1 convolution. The main structure of the residual block is the same as that of the multi-scale block. However, the number of channels of features remains the same before and after two 1×1 convolutions, and a residual map is constructed to enhance the feature difference of the essential properties of the image. The final output of the noise enhancement block is defined as:

$$\begin{cases} N_i = RB(MSB(N_{i-1})), i = 2,3,4 \\ N_i = RB(MSB(x_n)), i = 1 \end{cases} \tag{2}$$

Where $N_i \in R^{C_i \times H/2^{i-1} \times W/2^{i-1}}$, N_{i-1} is the output of the previous noise-guided module and represents the input of the module. We use the noise map extracted by the Improved Bayar Convolution as the original input, send it to the NGM to get the output, and then use the output as a new input to the NGM of the next stage. Finally, four noise feature maps of different scales are obtained, and the noise domain feature maps of these four scales are respectively matched with the feature maps of the four stages in the RGB domain.

3.3. Dual-domain progressive fusion module

In order to better combine the deep and shallow features of the network, semantic features and spatial structure features, global and local features, and make the localization of spliced regions more complete and the shape more accurate, we designed a dual-domain progressive fusion module. Perform a similarity calculation on the feature f_i from the RGB domain and the feature N_i from the noise domain, and then send it to SCAM to obtain the aggregated features S_i. Then the low-resolution synthetic features are used as compensation for high-resolution features, and the final feature map S_1 is generated from coarse to fine in a progressive manner. where $S_1 \in \mathbb{R}^{48 \times \frac{H}{4} \times \frac{W}{4}}$. Then a 1×1 convolution is used to convert the 48-channel feature map to 2-channel logits, using parameter-free bilinear upsampling. Finally, an inter-channel softmax function to convert the logits to a full-sized single-channel prediction mask $\hat{y}$. The specific process is as follows:

$$S_i = F_{SCAM}^{(i)}\left(f_i \odot Up\left(D_{i+1}\right)\right), i = 1,2,3 \tag{3}$$

where

$$D_{i+1} = S_{i+1} \odot N_{i+1} \tag{4}$$

$F_{SCAM}^{(i)}$ represents the SCAM operation of the ith scale, f_i is the four scale features extracted from the RGB stream, Up represents the upsampling operation, N_{i+1} is the multi-scale noise feature generated by the MSNB, and $\odot$ represents the element-wise multiplication of two feature maps. Since the output f_4 of the deepest layer of the network has reached 1/32 the size of the original image, a lot of detailed information has been lost. In order to reduce feature redundancy and save calculation, we directly send it to SCAM for spatial and channel feature aggregation. So S_4 can be directly expressed as $S_4 = F_{SCAM}^{(4)}(f_4)$, for the features of the other three scales, the features of the previous scale are used as a priori to generate the feature map of the current scale.

The structure of SCAM is shown in Fig. 4. The attention mechanism is usually used to give different degrees of attention to information of different importance. Spatial attention (SA) can transform various deformation data in space and automatically capture important regional features. The essence of channel attention mechanism (CA) is to model the importance of each feature channel and selectively enhance or suppress different channels. We propose to combine spatial and channel attention, add the output of SA and CA after 1×1 convolution, aggregate the effective features from the two branches in space and channel, selectively consolidate the tampered regions features, and enhance the contrast of tampered and untampered regions. Furthermore, to capture the strong correlation between spatial and channel features, SA and CA branches apply the same linear embedding, which also helps to control the number of network parameters. In the process of dual-domain progressive feature fusion, in order to more intuitively understand the interaction of the two features and the gradual process of generating masks from coarse to fine, we visualize feature maps S_i, N_i and D_i at three scales at different stages in the middle of the network. As shown in Fig. 5, the first to third rows in (a)-(c) represent the smallest sizes scale4 to scale2. (d) represents the tampered image, the groundtruth, and the predicted mask from top to bottom. In Fig. 5(a), the features S_4, S_3, and S_2 after SCAM are shown from top to bottom. It can be seen that in the RGB domain, the deeper the network, the more attention is paid to the semantic features of the image, the overall grasp of the spliced region is not accurate, and the localization of the edge is blurred. As shown in Fig. 5(b), our proposed noise-guided

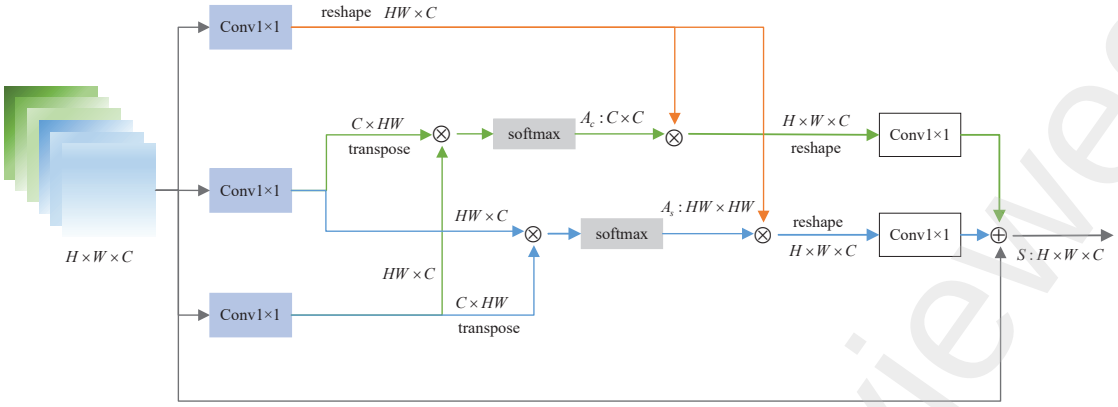


Fig. 4. The structure of the SCAM. where ⊗ denotes matrix multiplication and ⊕ denotes element-wise addition; A_c and A_s denote channel and spatial attention.

module performs very well in capturing low-level edge patterns, and the noise response is concentrated in the tampered region and the edge region of the object. Fig. 5(c) shows the features after we fuse these two parts of features scale by scale. It can be seen that the localization of the spliced region is gradually complete, the edge shape is precise, and the edge interference information of other objects is gradually weakened. Using the low-level noise domain spatial structure features to guide the deep semantic features of the network can better detect the precise location and edge shape of the spliced regions. At the same time, the semantic features of the RGB domain can suppress irrelevant redundant features in the noise domain, significantly reducing the false alarm rate. Finally, the features are continuously supplemented by combining multi-scale global and local information from coarse to fine to achieve a more accurate localization effect.

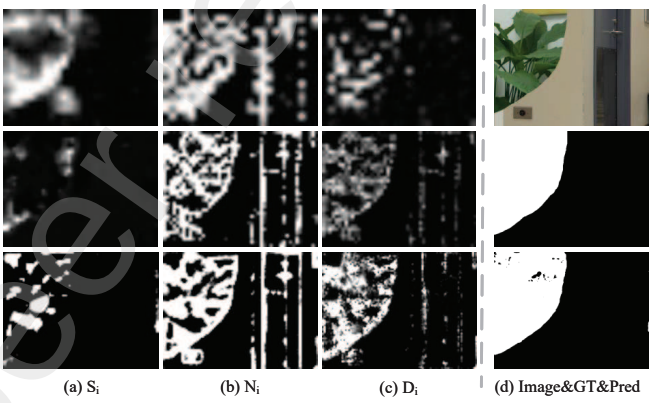


Fig. 5. Feature map visualization during the progressive fusion of two domains. (a) shows the feature map after SCAM; (b) is the extracted multi-scale noise map; (c) represents the feature map after the combination of (a) and (b).

4. Experiments

In order to evaluate the performance of MSNP-Net proposed in this paper, we compared the method with other image splicing forgery detection methods, conducted a series of ablation experiments to demonstrate the effectiveness of the main components in MSNP-Net, and in order to verify whether our network has stable performance against various attacks, we conducted robustness experiment.

4.1. Experimental setup

4.1.1. Datasets

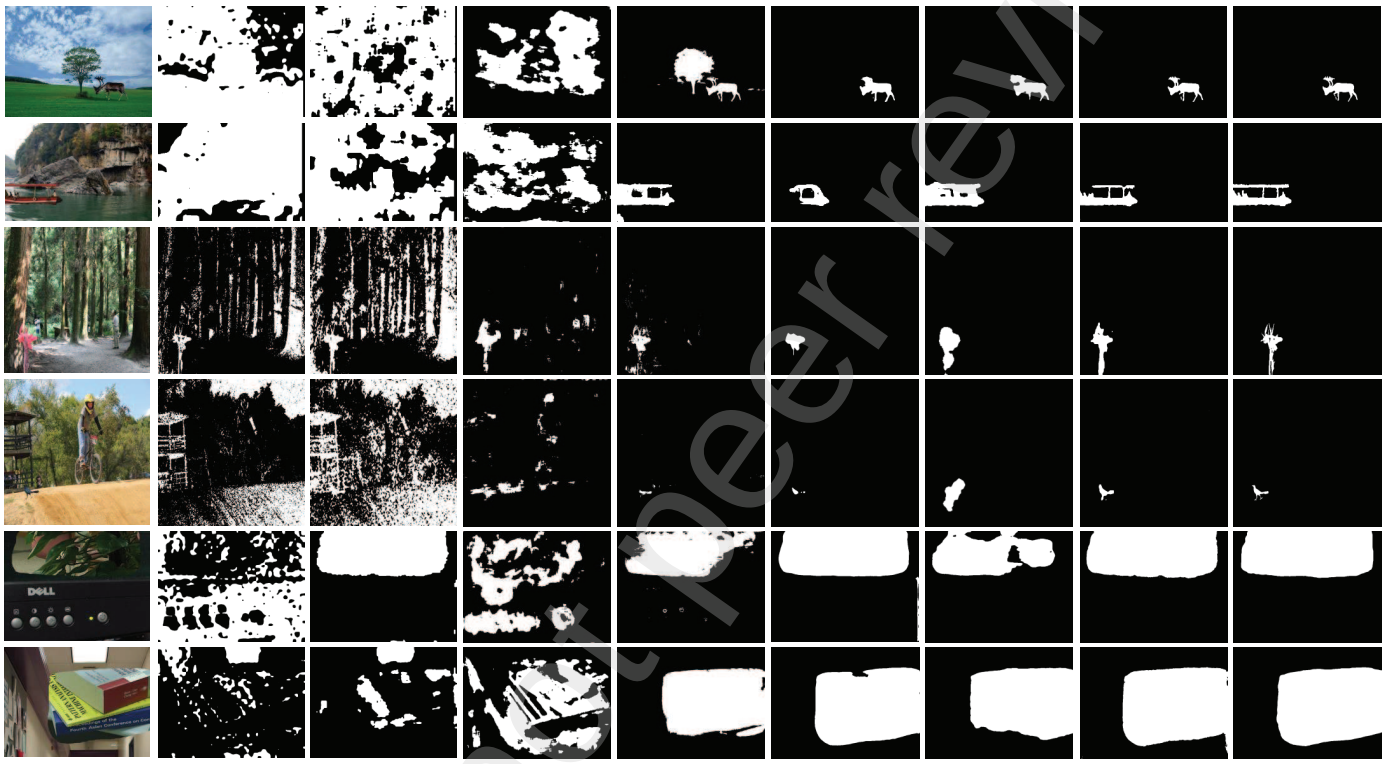
In this paper, we will analyze and evaluate our proposed method on public image manipulation datasets CASIA v2.0 [39], NIST'16 [40], and COLUMBIA [41]. In CASIA v2.0, the splicing region is small and delicate objects; on COLUMBIA, the splicing regions are some large and simple areas, and the detection difficulty is low. The fake images in the NIST'16 dataset have been post-processed to hide any visible traces of manipulation. In order to better train MSNP-Net, we use random flipping, random Gaussian noise, and JPEG compression to enhance the training set and validation set in CASIA v2.0 to form a dataset CASIA v2. 0+ with four times the capacity of the original. Table 1 summarizes the details of the datasets used in the experiments.

Table 1
The setup of training, validation and testing sets on CASIA v2.0+, NIST'16 and COLUMBIA datasets.

Sets	CASIA v2.0+					NIST'16	COLUMBIA
	Plain	Flip Operation	JPEG Compression	Gaussian Noise	Total		
Training Set	1300	1300	1300	1300	5200	190	125
Validation Set	50	50	50	50	200	18	5
Testing Set	200	-	-	-	200	80	50
All Images	-	-	-	-	5600	288	180

Table 2
Detection and localization performance of compared methods on different datasets (%).

Method	Year	Datasets								
		CASIA v2.0+			NIST'16			COLUMBIA		
		F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision
NOI[42]	2009	18.90	13.02	76.79	12.42	10.89	62.17	36.10	46.54	32.94
CFA[26]	2012	23.60	18.04	68.05	14.83	13.74	67.38	51.74	62.13	52.41
ManTra-Net[43]	2019	30.99	57.68	27.53	27.65	41.93	30.14	52.01	52.27	53.98
RRU-Net[31]	2019	84.22	**92.74**	81.67	77.98	76.12	85.16	77.50	76.07	81.67
C2R-Net[32]	2020	67.58	80.80	58.10	55.00	66.60	46.80	69.50	61.20	80.40
CAT-Net[34]	2021	80.12	75.05	87.54	90.19	88.28	93.77	97.44	97.55	97.84
PSCC-Net[44]	2022	79.78	86.73	78.15	76.28	93.70	68.41	93.53	90.92	96.90
MSNP-Net	2023	**88.42**	85.32	**92.79**	**95.50**	**95.00**	**96.26**	**98.63**	**99.04**	**98.25**



(a) Image (b) NOI (c) CFA (d) Mantra-Net (e) RRU-Net (f) CAT-Net (g) PSCC-Net (h) Ours (i) Ground-Truth

Fig. 6. Localization results of MSNP-Net and other methods on different datasets. Images from top to bottom are from CASIA v2.0+, NIST16, and COLUMBIA. Each dataset has two images.

4.1.2. Evaluation metrics

For image splicing forgery detection, the pixel-level localization performance of tampered regions is an important evaluation criterion. The evaluation metrics are the number of correctly detected tampered pixels (TP), the number of falsely detected tampered pixels (FP), and the number of falsely detected untampered pixels (FN). In the following experiments, we evaluate the performance of the proposed splicing forgery detection method using Precision, Recall and F-measure. The Precision is defined as Eq. (5), which indicates the proportion of correctly predicted pixels in the detected regions. Recall refers to the ratio of correctly detected pixels to the groundtruth, defined as Eq. (6). F-measure is a comprehensive evaluation index combining Precision and Recall, as shown in Eq. (7).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

4.1.3. Implementation details and loss function

Our model is implemented in PyTorch and trained using a GeForce GTX 1080Ti. We initialize the network backbone with ImageNet pre-trained weights and optimize the entire model by stochastic gradient descent with a momentum of 0.9. The batch size is set to 4, and the initial learning rate is 0.005, which decays exponentially. The input size of the network training images is 512×512. In order to better train MSNP-Net, we use the weighted binary cross-entropy loss to supervise the network, where the weight parameter is set to (0.5, 2.5).

Table 3
Different configurations of ablation study and their experimental results on the CASIA v2.0+ dataset (%).

Model	Components						Metric		
	Noise Extractor			NGM		SCAM			
	bayar+	srm	mul_noise	MSB	MSB+RB		F-measure	Recall	Precision
HR							70.86	66.48	80.44
SB						✓	70.46	65.65	78.81
DB	✓					✓	87.06	84.08	91.25
DB+NGM-	✓			✓		✓	87.01	82.51	93.47
DB_srm+NGM		✓			✓	✓	87.09	82.73	93.28
DB_mul+NGM			✓		✓	✓	87.31	85.11	90.42
DB_bayar+NGM(w/o SCAM)	✓				✓		87.11	81.94	**94.21**
DB_bayar+NGM	✓				✓	✓	**88.42**	**85.32**	92.79

4.2. Comparison with the state of the art

To evaluate the performance of MSNP-Net, we selected two traditional methods and five advanced deep learning methods as baseline methods, namely NOI [42], CFA [26], ManTra-Net [43], RRU-Net [31], C2R-Net [32], CAT-Net [34] and PSCC-Net [44]. Table 2 lists the detection performance of these methods on CASIA v2.0+, NIST'16 and COLUMBIA datasets respectively. As can be seen from Table 2, the method proposed in this paper has an F-measure of 88.42% and a Precision of 92.79% on the CASIA v2.0+, both of which have achieved the best performance, but its Recall is 85.32%, which is lower than the 92.74% of RRU-Net and 86.73% of PSCC-Net. This is because the Recall calculates the proportion of the number of correctly predicted positive samples to the total number of positive samples, the proportion of the groundtruth that is correctly detected. From the visualization results of the network in Fig. 6, it can be seen that the false positive rate of RRU-Net is high. The pixels around the tampered object are all judged as positive samples, and the localization of the tampered regions is not accurate enough. PCSS-Net is inaccurate in locating tampered edges and has insufficient grasp of details. Our method achieves the best performance on the COLUMBIA dataset. It is worth noting that the method proposed in this paper has an F-measure of 95.5% on the NIST'16 dataset, and the overall performance of other indicators is far superior to other methods. This is a challenging dataset in which fake images have undergone a series of post-processing to hide any visible traces of tampering. As can be seen in the third and fourth lines of Fig. 6, RRU-Net and CAT-Net cannot completely locate the tampered regions at all, and PSCC-Net can only perform rough localization of tampered regions. This demonstrates the superiority of our network in capturing subtle image tampering artifacts and discovering learnable features.

We randomly select two tampered pictures in each of the three datasets and analyze the qualitative results of the compared baseline methods, as shown in Fig. 6. The experimental results show that our method is significantly better than the baseline method. MSNP-Net's spatial and noise dual-branch structure can not only grasp the semantic information of the image and describe the overall shape of the tampered object but also capture the detailed features of the image structure and clearly locate the edge of the tampered regions. Our multi-scale progressive method can adapt to different scales of tampering. It can accurately locate large-scale and small-scale tampering, and the detection effect on different datasets is stable.

4.3. Ablation study

To better verify the effectiveness of each module in our proposed network, we gradually add each component to the MREB for training and compare the results. We conducted a series of experiments on the CASIAv2.0+ dataset, and the designed ablation model is as follows:

- HR: Use HR-Net as the basic model.
- SB: A multi-resolution feature extraction branch is built based of HR-Net, and a SCAM module is added as a fusion method.
- DB: Add multi-scale noise-guided branches builds a two-branch network, using the Improved Bayar Convolution as the noise extractor.
- DB+NGM-: A noise-guided module is added to the dual-branch network, and only multi-scale blocks are included in the noise-guided module.
- DB_srm+NGM: Use SRM as the noise extractor in the multi-scale noise-guided branch of a dual-branch network.
- DB_mul+NE: Use SRM and the Improved Bayar Convolution simultaneously as the noise extractors in the multi-scale noise-guided branch of a dual-branch network.
- DB_bayar+NGM (w/o SCAM): Use the Improved Bayar Convolution as the noise extractor. SCAM is not used for feature aggregation in dual-branch networks.
- DB_bayar+NGM: A noise-guided module is added to the dual-branch network, the Improved Bayar Convolution is used as the noise extractor, and a SCAM module is added for feature aggregation, which is our final model.

The setup of the ablation experiment and its results are shown in Table 3. HR, SB, and DB prove the role of our proposed multi-scale noise-guided branch, and the dual-branch network greatly improves the detection accuracy. Comparing the results of DB, DB+NGM- and DB_bayar+NGM shows that the noise-guided module combined with the multi-scale block and

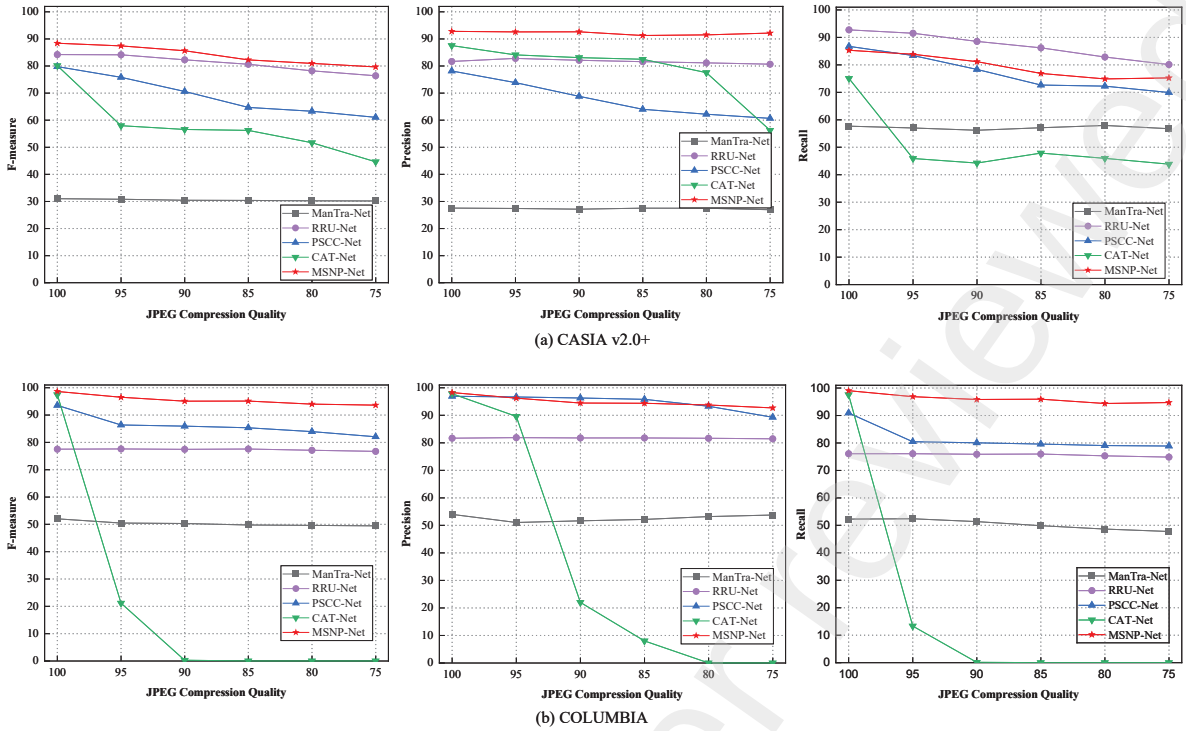


Fig. 7. Comparison results under JPEG compression attack. The three columns represent the F-measure, Precision and Recall. (a)represent the experiment results on CASIA v2.0+; (b) represent the experiment results on COLUMBA.

residual block proposed by us can better enhance the feature expression, provide more tampering clues, and improve the detection effect of the network. In order to verify the optimality of the noise extractor we used, we tested the effect of using only SRM, only using the Improved Bayar, and using two noise extractors at the same time. From the results of DB_srm+NGM, DB_mul+NGM, and DB_bayar+NGM, it can be found that the performance of using the Improved Bayar Convolution as the noise extractor is optimal. After adding SCAM, although the precision dropped by 1.42%, the recall increased by 3.38% and the comprehensive evaluation F-measure increased by 1.31%. The results of DB_bayar+NGM (w/o SCAM) and DB_bayar+NGM show that SCAM can better aggregate the spatial channel features from the two branches, strengthen the feature learning of tampered regions, and better distinguish tampered pixels from untampered pixels. This proves the effectiveness of SCAM in feature fusion.

4.4. Robustness evaluation

To further verify the effectiveness and robustness of the proposed method, we evaluate the performance of the detection method under various attacks, including JPEG compression and noise attacks, on the CASIA v2.0+ and COLUMBIA datasets. Since the traditional methods NOI and CFA almost detect the entire image as a tampered region, only four deep learning methods are selected here for comparison.

The comparative experimental results under the JPEG compression attack are shown in Fig. 7. The F-measure, Precision, and Recall of the experimental results are compared from left to right in Fig. 7. It can be seen that the F-measure and Precision of MSNP-Net on CASIA v2.0+ are better than the other four detection methods, while the Recall is slightly lower than RRU-Net. As mentioned above, there are many misjudgments in the detection of RRU-Net, and the localization of the tampered regions is general but not precise. Similarly, the metrics of MSNP-Net on COLUMBIA are all the better than the other four detection methods. We observe that CAT-Net performs abnormally robustly on the COLUMBIA dataset. This is due to the fact that CAT-Net relies on JPEG compression artifacts as forensic clues and is pretrained using a large dataset with diverse quantization tables containing various compressions (50-99). However, the number of pictures in the COLUMBIA dataset is small, and these pictures are all in non-JPEG format. Therefore, when adding compression with different quality factors, the features learned by the DCT stream are useless, seriously affecting the performance of forensics. Through experiments, it can be found that under the JPEG compression attack of the two datasets, the detection result of MSNP-Net is better than other detection methods, and it has high robustness.

Fig. 8 shows the comparative experimental results under the Gaussian noise attack on the CASIA v2.0+ and COLUMBIA datasets. On CASIA v2.0+, the F-measure and Precision of MSNP-Net are better than the other four detection methods, and the Recall is slightly lower than RRU-Net. On the COLUMBIA dataset, the F-measure, Precision, and Recall of MSNP-Net are better than the other four detection methods. From the above analysis, it can be seen that MSNP-Net shows more stable performance under the noise attack.

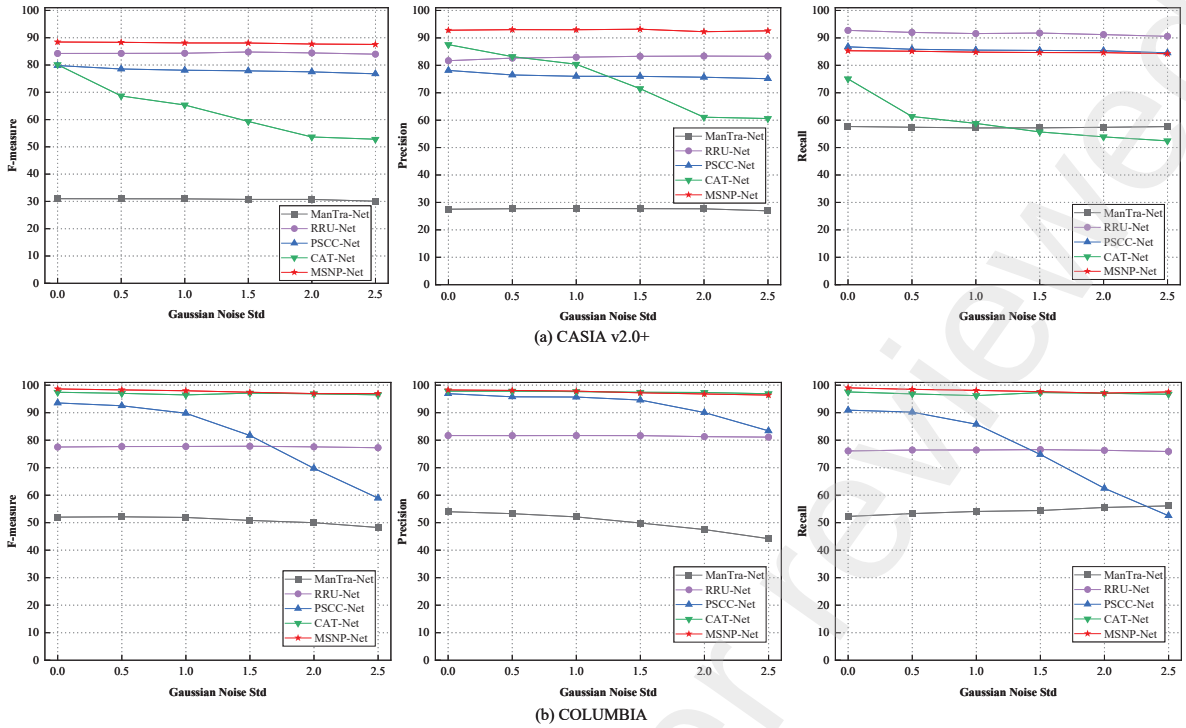


Fig. 8. Comparison results under noise corruption attack. The three columns represent the F-measure, Precision and Recall. (a)represent the experiment results on CASIA v2.0+; (b) represent the experiment results on COLUMBA.

5. Conclusion

This paper proposes a multi-scale noise-guided progressive network for image splicing forgery detection and localization. The dual-branch structure can fully use deep semantic and shallow spatial features, and the noise-guided module is used to enhance the expression of spatial structure features further. At the same time, multi-scale RGB domain features and noise domain features are combined, and a progressive fusion mechanism is used to generate the final prediction mask. Focusing on global and local features enables our model to better adapt to different sizes of tampering and maintain stable detection performance under different splicing forgery situations. Extensive experiments show that MSNP-Net outperforms the state-of-the-art methods in detection and localization and exhibits better robustness.

Although our method is superior to the existing splicing localization methods, its localization performance on small, shapeless, and irregular spliced regions without semantic features is not ideal. In order to overcome this limitation, in future work, we will add the edge features of images as forensic clues, and use diversified and multi-dimension feature information to achieve precise localization of various types of spliced images.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China under grant 62172059, 62272160, and U22A2030; Scientific Research Fund of Hunan Provincial Education Department of China under Grant 22A0200.

References

[1] D. A. T. Thomson, P. Dootson, Seeing no longer means believing. `https://indaily.com.au/opinion/2020/11/04/seeing-should-not-mean-believing/`, (accessed 7 july 2023) (2020).

[2] L. Verdoliva, Media forensics and deepfakes: an overview, IEEE Journal of Selected Topics in Signal Processing 14 (5) (2020) 910–932.

[3] M. K. Johnson, H. Farid, Exposing digital forgeries in complex lighting environments, IEEE Transactions on Information Forensics and Security 2 (3) (2007) 450–461.

[4] B. Mahdian, S. Saic, Detection of resampling supplemented with noise inconsistencies analysis for image forensics, in: 2008 International Conference on Computational Sciences and Its Applications, IEEE, 2008, pp. 546–556.

[5] Y.-F. Hsu, S.-F. Chang, Detecting image splicing using geometry invariants and camera characteristics consistency, in: 2006 IEEE International Conference on Multimedia and Expo, IEEE, 2006, pp. 549–552.

[6] S. Ye, Q. Sun, E.-C. Chang, Detecting digital image forgeries by measuring inconsistencies of blocking artifact, in: 2007 IEEE International Conference on Multimedia and Expo, Ieee, 2007, pp. 12–15.

[7] Z. Fang, S. Wang, X. Zhang, Image splicing detection using color edge inconsistency, in: 2010 International Conference on Multimedia Information Networking and Security, IEEE, 2010, pp. 923–926.

[8] A. Ghosh, Z. Zhong, T. E. Boult, M. Singh, Spliceradar: A learned method for blind image forensics., in: CVPR Workshops, 2019, pp. 72–79.

[9] X. Cun, C.-M. Pun, Image splicing localization via semi-global network and fully connected conditional random fields, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.

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

[11] C. Dong, X. Chen, R. Hu, J. Cao, X. Li, Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection, IEEE Trans-

actions on Pattern Analysis and Machine Intelligence 45 (3) (2022) 3539–3553.
[12] A. M. Ismael, A. Şengür, Deep learning approaches for covid-19 detection based on chest x-ray images, Expert Systems with Applications 164 (2021) 114054.
[13] T. Mahmood, S. W. Cho, K. R. Park, Dsrd-net: Dual-stream residual dense network for semantic segmentation of instruments in robot-assisted surgery, Expert Systems with Applications 202 (2022) 117420.
[14] B. Gülmez, Stock price prediction with optimized deep lstm network with artificial rabbits optimization algorithm, Expert Systems with Applications 227 (2023) 120346.
[15] J. H. Bappy, C. Simons, L. Nataraj, B. Manjunath, A. K. Roy-Chowdhury, Hybrid lstm and encoder–decoder architecture for detection of image forgeries, IEEE Transactions on Image Processing 28 (7) (2019) 3286–3300.
[16] N. Huang, J. He, N. Zhu, A novel method for detecting image forgery based on convolutional neural network, in: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), IEEE, 2018, pp. 1702–1705.
[17] Y. Rao, J. Ni, A deep learning approach to detection of splicing and copy-move forgeries in images, in: 2016 IEEE international workshop on information forensics and security (WIFS), IEEE, 2016, pp. 1–6.
[18] W. Chen, Y. Q. Shi, G. Xuan, Identifying computer graphics using hsv color model and statistical moments of characteristic functions, in: 2007 ieee international conference on multimedia and expo, IEEE, 2007, pp. 1123–1126.
[19] H. Farid, S. Lyu, Higher-order wavelet statistics and their application to digital forensics, in: 2003 Conference on computer vision and pattern recognition workshop, Vol. 8, IEEE, 2003, pp. 94–94.
[20] Y. Q. Shi, C. Chen, W. Chen, A natural image model approach to splicing detection, in: Proceedings of the 9th workshop on Multimedia & security, 2007, pp. 51–62.
[21] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE transactions on pattern analysis and machine intelligence 43 (10) (2020) 3349–3364.
[22] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Transactions on information Forensics and Security 7 (3) (2012) 868–882.
[23] B. Bayar, M. C. Stamm, Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection, IEEE Transactions on Information Forensics and Security 13 (11) (2018) 2691–2706.
[24] Z. Zhang, Y. Qian, Y. Zhao, L. Zhu, J. Wang, Noise and edge based dual branch image manipulation detection, arXiv preprint arXiv:2207.00724.
[25] A. E. Dirik, N. Memon, Image tamper detection based on demosaicing artifacts, in: 2009 16th IEEE International Conference on Image Processing (ICIP), IEEE, 2009, pp. 1497–1500.
[26] P. Ferrara, T. Bianchi, A. De Rosa, A. Piva, Image forgery localization via fine-grained analysis of cfa artifacts, IEEE Transactions on Information Forensics and Security 7 (5) (2012) 1566–1577.
[27] Z. Lin, J. He, X. Tang, C.-K. Tang, Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis, Pattern Recognition 42 (11) (2009) 2492–2501.
[28] T. Bianchi, A. De Rosa, 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.
[29] Y. Zhang, J. Goh, L. L. Win, V. L. Thing, Image region forgery detection: A deep learning approach., SG-CRC 2016 (2016) 1–11.
[30] R. Salloum, Y. Ren, C.-C. J. Kuo, Image splicing localization using a multi-task fully convolutional network (mfcn), Journal of Visual Communication and Image Representation 51 (2018) 201–209.
[31] X. Bi, Y. Wei, B. Xiao, W. Li, Rru-net: The ringed residual u-net for image splicing forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.
[32] B. Xiao, Y. Wei, X. Bi, W. Li, J. Ma, Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering, Information Sciences 511 (2020) 172–191.
[33] M. Huh, A. Liu, A. Owens, 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.
[34] M.-J. Kwon, I.-J. Yu, S.-H. Nam, H.-K. Lee, Cat-net: Compression artifact tracing network for detection and localization of image splicing, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 375–384.
[35] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, Learning rich features for image manipulation detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1053–1061.
[36] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28.
[37] Y. Zhang, G. Zhu, L. Wu, S. Kwong, H. Zhang, Y. Zhou, Multi-task senetwork for image splicing localization, IEEE Transactions on Circuits and Systems for Video Technology 32 (7) (2021) 4828–4840.
[38] M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images, IEEE Transactions on Information Forensics and Security 14 (5) (2018) 1181–1193.
[39] J. Dong, W. Wang, T. Tan, Casia image tampering detection evaluation database, in: 2013 IEEE China summit and international conference on signal and information processing, IEEE, 2013, pp. 422–426.
[40] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, J. Fiscus, Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, 2019, pp. 63–72.
[41] J. Hsu, S. Chang, Columbia uncompressed image splicing detection evaluation dataset, Columbia DVMM Research Lab 6.
[42] B. Mahdian, S. Saic, Using noise inconsistencies for blind image forensics, Image and vision computing 27 (10) (2009) 1497–1503.
[43] Y. Wu, W. AbdAlmageed, P. Natarajan, Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9543–9552.
[44] X. Liu, Y. Liu, J. Chen, X. Liu, Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization, IEEE Transactions on Circuits and Systems for Video Technology 32 (11) (2022) 7505–7517.