

Received 28 February 2023, accepted 24 March 2023, date of publication 3 April 2023, date of current version 6 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264014

RESEARCH ARTICLE

TransU²-Net: A Hybrid Transformer Architecture for Image Splicing Forgery Detection

CAIPING YAN¹, (Member, IEEE), SHUYUAN LI¹, AND HONG LI², (Member, IEEE)

¹Department of Computer Science, Hangzhou Normal University, Hangzhou 311121, China

²Hangzhou InsVision Technology Company Ltd., Hangzhou 311121, China

Corresponding author: Hong Li (hong.li.leon@connect.um.edu.mo)


This work was supported in part by the National Natural Science Foundation of China under Grant 61902102, and in part by the Natural Science Foundation of Zhejiang Province under Grant LQ19F020004.

ABSTRACT In recent years, various convolutional neural network (CNN) based frameworks have been presented to detect forged regions in images. However, most of the existing models can not obtain satisfactory performance due to tampered areas with various sizes, especially for objects with large-scale. In order to obtain an accurate object-level forgery localization result, we propose a novel hybrid transformer architecture, which exhibits both advantages of spatial dependencies and contextual information from different scales, namely, TransU²-Net. Specifically, long-range semantic dependencies are captured by the last block of encoder to locate large-scale tampered areas more completely. Meanwhile, non-semantic features are filtered out by enhancing low-level features under the guidance of high-level semantic information in the skip connections to achieve more refined spatial recovery. Therefore, our hybrid model can locate spliced forgeries with various sizes without requiring large data set pre-training. Experimental results on the Casia2.0 and Columbia datasets show that our framework achieves better performance over state-of-the-art methods. On the Casia 2.0 dataset, F-measure improve by 8.4% compared to the previous method.

INDEX TERMS Image splicing forgery detection, tampered region localization, convolutional neural network, self-attention, cross-attention.

I. INTRODUCTION

Digital image generation and transmission has become very easy due to the rapid development of modern mobile devices. Meanwhile, the operation of image editing software is simple, which makes it easy for anyone to modify the picture. Generally speaking, people modify the image for the purpose of beautification and entertainment. However, some forged images may be abused maliciously, causing negative impact on the society and the country [1]. Therefore, it has become increasingly important to detect image manipulations. Among different image manipulations, splicing is regarded as copying a part of an image and pasting into another image to form a new image. As shown in Fig. 1, an example of image splicing forgery is given, including tampered image, authentic image and ground truth, where white area in ground-truth image is tampered area. There exist

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar .

intrinsic image discrepancies between the authentic and the tampered regions named image fingerprints, which can be used to determine whether an image has been tampered or not and locate the tampered areas in the forged image. During the last decade, two main categories of splicing forgery detection methods have been proposed: traditional feature extraction based methods and CNN-based detection methods.

Many traditional detection methods extract the particular image fingerprints, such as color filter array interpolation [2], sensor noise [3], particular photo-response non-uniformity [4], [5], [6], etc. Some traditional detection methods have problems in detecting tampered regions with rich complex textures. In addition, traditional detection methods can only have a certain effect on specific image fingerprints, but when the specific fingerprints do not exist in the image, the detection results will be poor. Moreover, the specific image fingerprint will be affected by attack effects such as Gaussian noise attacks, JPEG compression, resize attacks, which means that the robustness of traditional detection methods is poor.

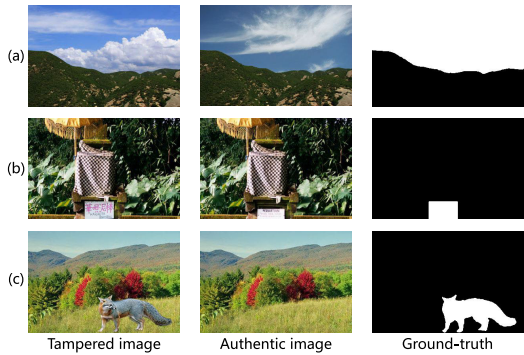


FIGURE 1. Examples of image splicing forgery. (a) Modified the sky background, (b) modified the object in the middle, (c) added animals to the image.

Based on locality and translation invariance, CNN has achieved tremendous success in computer vision, such as classification [7], [8], [9], object detection [10] and semantic segmentation [11]. As a result, many splicing detection methods based on CNN have been proposed and demonstrate better performance than traditional ones. These CNN-based tamper detection methods can extract a variety of image fingerprints simultaneously. By a huge margin, traditional methods can only get a single image attribute and lacks of generalization. For example, Zhou et al. [12] introduced Faster Region-CNN using both RGB stream and noise stream to detect the tampered regions of a manipulated image. They extract noise features from the forged images using an SRM filter layer that can capture noise inconsistencies between tampered areas and real areas. Nonetheless, the limitation of this method is that it can only achieve region-level tamper localization result. In order to achieve pixel-level result, a possible solution is to use fully convolutional networks. Bappy et al. [13] presented a hybrid CNN-LSTM model to learn the spatial structure between those tampered and non-tampered regions in the shared boundary. Chen et al. [14] proposed an encoding and decoding framework that utilizes both hybrid features and semantic reinforcement network for image forgery detection. Particularly, resampled features with long-short term memory is utilized to capture traces from the image patches for finding manipulating artifacts. However, all the above-mentioned methods [12], [13], [14] perform image tamper detection by processing pictures in patch first, which means they focus only on local areas and ignore relationships between sub-regions, thus their localization performance depends on the size of the patches.

To get around of this issue, some end-to-end networks have been proposed, which regard the problem of pixel-level forgery localization as an image segmentation or object detection task. Bi et al. [15] presented an end-to-end deep neural network that uses the residual-propagation and residual-feedback modules to capture discriminative features between manipulated and non-manipulated regions. However, the performance of the method is not yet satisfactory especially for large-scale tampered areas, which may be

caused by the insufficient consideration of global correlations. Wu et al. [16] presented an end-to-end deep neural architecture (ManTra-Net), which utilizes two structures to classify forgery images and detect different forgeries such as enhancement, splicing, copy-move, and even unknown types. However, the performance of Mantra-Net is also poor particularly on the dataset of Casia. Bi et al. [17] employed an image forgery detection method based on dual-encoder U-Net. This method utilizes a fixed encoder and an unfixed encoder. The unfixed encoder learns the forensic features to distinguish between the tampered and non-tampered areas. And the fixed encoder uses DWT (Haar discrete wavelet transform) to extract the direction information of the boundary of the tampered area. In addition, a spatial pyramid global-feature extraction module is designed to get global feature. The results show that the method can effectively improve the detection accuracy without pre-training or training on a large dataset. Bi et al. [18] proposed a multitask wavelet corrected network (MWC-Net) which can generate more comprehensive and representative features for image splicing detection and localization. However, the method of MWC-Net still does not take into account global features, which leads to probably poor localization performance for large-scale forgeries. Myung-Joon Kwon et al. [19] presented CAT-Net, which is also an end-to-end fully convolutional neural network including a RGB stream and a DCT stream. CAT-Net effectively learn the forensic features remaining in each domain through RGB stream and DCT stream. Each stream takes different resolutions to handle tampered regions of various sizes and shapes.

With the popularity of attention mechanism, some methods add attention mechanism to CNN and achieve better performance. Rao et al. [20] proposed a novel network that incorporates with multi-semantic CRF-based attention model. The attention map generated by multi-semantic CRF-based attention model can suppress noise and highlight informative regions to guide the network to extract more representative features around forged boundaries. Liu et al. [21] developed a PSCC-Net for image forgery detection and localization. PSCC-Net adopt HR-Net as backbone, and generate manipulated regions in a coarse-to-fine fashion. Meanwhile, a spatial-channel correlation module is proposed to perform channel-wise attention and spatial attention on extracted features, which can improve the detection accuracy and the robustness of the network.

However, most prior works ignore that the size of the tampered area is variant and encounter difficulties in locating tampered regions of different sizes. Due to the intrinsic locality of convolution operation, it is difficult for CNN-based methods to learn explicit global semantic information dependencies and have difficulties in leveraging local and global features jointly. Therefore, most CNN-based approaches can only deal with limited scale variation. In addition, these methods may cause incomplete localization or high false detection rate in locating large-scale tampered areas.

In order to accurately locate forged areas with various sizes, we introduce the TransU²-Net, a U-shape hybrid Transformer Network, which takes the advantages of convolution and attention for image splicing forgery detection and localization. Firstly, in the process of encoding, we employ mixed receptive fields to extract the features of tampered images. Secondly, the self-attention module explicitly models the complete context information by using the global interaction between the semantic features at the end of the encoder. In addition, cross-attention is introduced in skip connections to filter out non semantic features, so as to achieve fine spatial recovery in TransU²-Net decoder, and promote correctness of prediction results. Finally, we input the feature maps after self-attention into the decoder for decoding. In the decoding stage, we use the learned features to estimate the final manipulation mask. Due to this design, the final manipulation mask can obtain both local and global feature.

The contributions of this work can be summarized as follows:

- We propose a novel TransU²-Net, a U-shape hybrid Transformer Network, which integrates both self- and cross-attention into U²-Net. It is able to capture more contextual information and spatial dependencies from different scales.
- We design a new cross-attention module in skip connection to filter out non semantic features, so as to enhance the low-level feature maps that are passed through the skip connections under the guidance of high-level semantic information, and achieve fine spatial recovery in decoder, so as to finally promote correctness of prediction results.
- We introduce self-attention at the last block of encoder to combine the strength from both self-attention mechanism and convolution. Therefore, the TransU²-Net can avoid to rely heavily on large-scale pre-training, and it has the ability of Transformer to learn explicit long-range semantic information dependencies.
- The proposed TransU²-Net, which fuses convolution and Transformer together, can locate spliced forgeries with various sizes, thus achieve new state-of-the-art performance on public datasets.

The remainder of this paper is organized as follows. The second section presents the implementation details of TransU²-Net model, including residual U-blocks, self-attention mechanism and cross-attention module. The third section describes the experimental results of the proposed TransU²-Net and other detection methods for comparison on different datasets. The conclusion of this paper is in the fourth section.

II. PROPOSED DETECTION METHOD

In this section, we present the details of our proposed TransU²-Net, which aims to locate the tampered regions at pixel level. Fig. 2 gives the general framework of our

TransU²-Net. First, considering the powerful ability of Transformer, we combine self-attention into the last block of encoder to learn image fingerprints, which can better integrate the advantages of Transformer and encoder network, i.e., the ability of modeling global and long-range dependencies and feature learning ability. Then, cross-attention is used to enhance the low-level feature maps with the guidance of high-level semantic information, filter out non-semantic features between the encoding and decoding networks, thus achieves finer spatial recovery in the decoder and promotes correctness of prediction results. Finally, our decoder learns to fusion all the feature maps from low- to high-resolution and predicts the tampered area at pixel-level. Next, we introduce our TransU²-Net in the following three parts: encoding network, cross-attention module and decoding network.

A. ENCODING NETWORK

Most deep learning models for image splicing forgery localization use traditional encoder-decoder [22] and no-pooling structures [23] to extract features. However, most structures ignore that the size of the tampered area is variant, and thus can only solve limited scale variation. A significant improvement of the U²-Net model [24] compared with the traditional model is that the Residual U-Block (RSU) is able to mix receptive fields of different sizes. That is to say, we can utilize this advantage to better deal with the problem of tampered areas with arbitrary size. Therefore, we use the model of U²-Net as the backbone of our proposed TransU²-Net. In addition, to address the issue that CNN-based detection networks lack the ability to explicitly model global information, we introduce self-attention modules in the last block of encoder to capture long-range semantic information interaction.

The TransU²-Net is a two-level nested U-structure, as shown in Fig. 2. The symmetrical structure can be used to learn how to extract and encode multi-scale context information. By using the different layers of RSU, the network can capture multi-scale features within the stage, and reduces the loss of detail caused by large-scale direct upsampling. In addition, the nested U-structure allows the network to obtain higher resolution feature maps, thus providing multi-level deep features.

As shown in Fig. 2, TransU²-Net consists of an encoder with six stages and a decoder with five stages. In the bottom-up path, we use the RSU-7, RSU-6, RSU-5, and RSU-4, respectively, where “7”, “6”, “5” and “4” denote the height (L) of the RSU block. The L is configured based on the spatial resolution of the input feature maps. Compared with **En_4**, the pooling and upsampling operations are replaced by dilated convolutions in the stage of **En_5**, and we name this block as RSU-4F (see Fig. 3). That means all feature maps in RSU-4F have the same resolution. In addition, in stage of **En_6**, we added self-attention module to the original RSU-4F to form self-attention U-block.

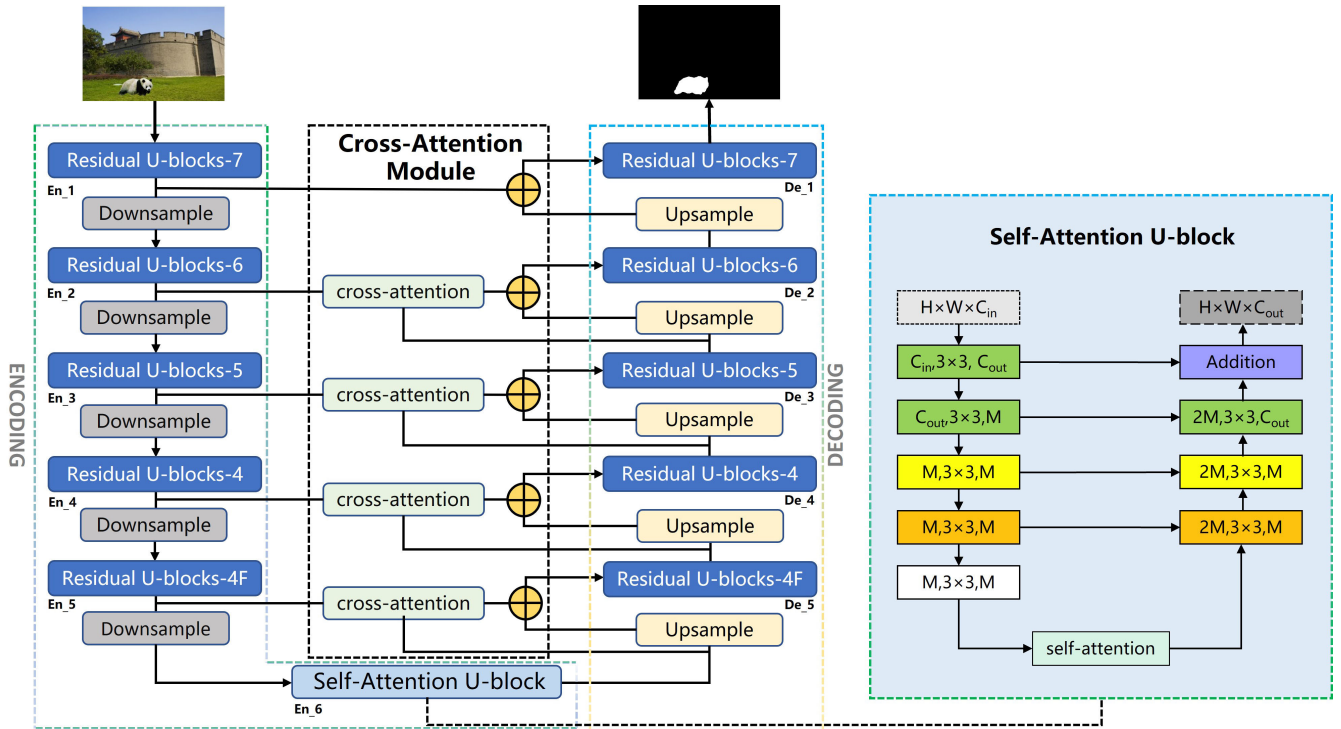


FIGURE 2. The network architecture of TransU²-Net for image forgery detection. The picture on the right is illustration of our proposed self-attention U-block.

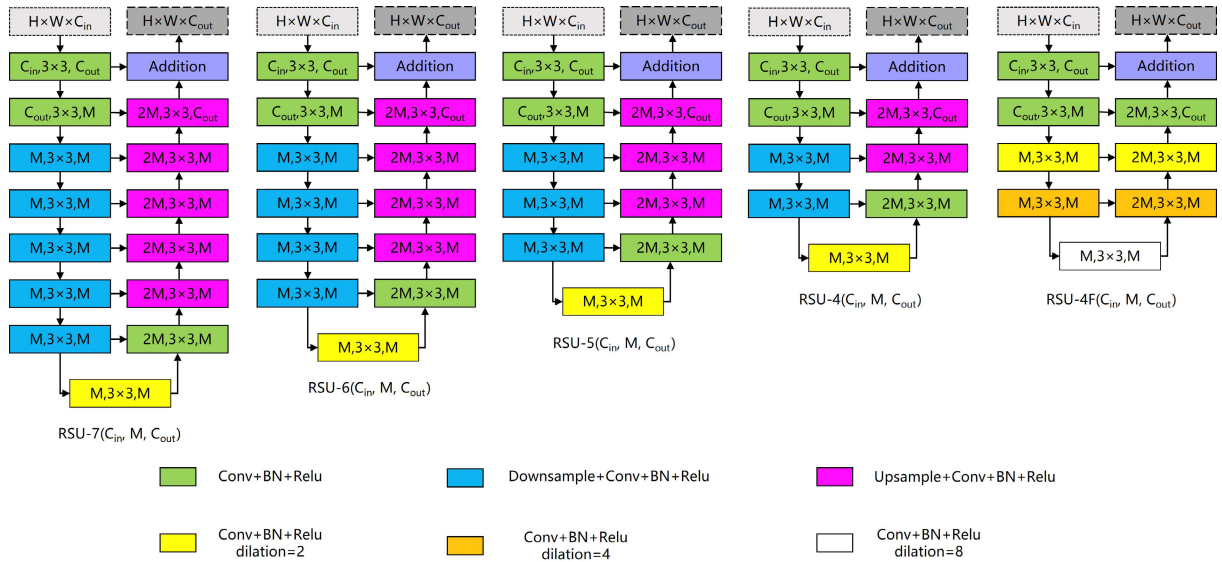


FIGURE 3. Illustration of residual U-block RSU.

1) RESIDUAL U-BLOCKS

Multi-scale features cannot be obtained by only using ordinary convolution blocks and the shallow feature maps only contain local features. Although the dilated convolution can extract both local features and non-local multi-scale features by expanding the receiving field, multiple convolutions on the feature maps with original resolutions require too much computational and memory resources. In order to solve the

above-mentioned problems, U²-Net introduces a novel RSU to capture the multi-scale features in each stage. The structures of five residual U-block RSU-L (C_{in} , M , C_{out}) with different value L are shown in Fig. 3, where L represents the height of the RSU block, C_{in} , C_{out} are input and output channels, respectively. M represents the number of channels in the intermediate layers of RSU. In general, the RSU mainly consists of three components:

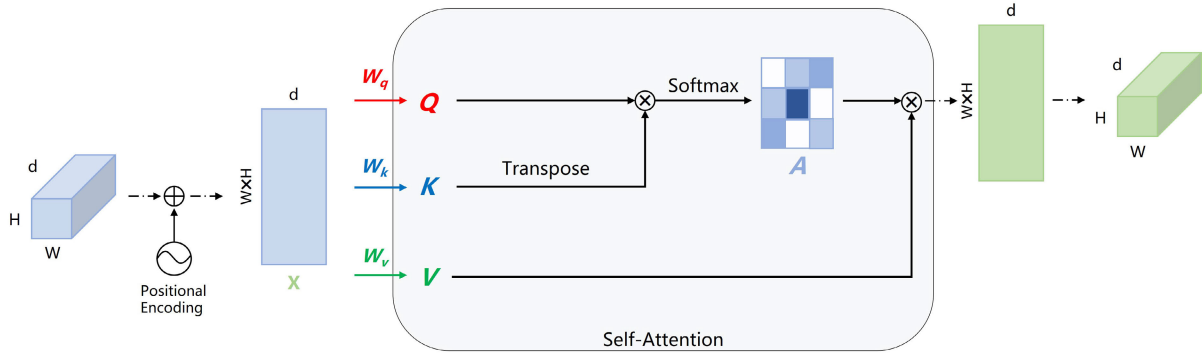


FIGURE 4. Illustration of self-attention module.

1) A 3×3 convolution layer that transforms the feature map \mathbf{x} ($H \times W \times C_{in}$) to obtain an internal feature map $F_1(x)$ with the channel of C_{out} . The convolutional layer is used to extract the local features of the current feature map.

2) A symmetric encoder-decoder structure with a height of L , takes the internal feature map $F_1(x)$ as input, and extracts multi-scale feature $U(F_1(x))$. This component encodes multi-scale features into high-resolution feature maps through upsampling, concatenation and convolution, which reduces the feature loss and improves the accuracy of tamper localization.

3) A residual connection which mixes local and multi-scale information by the summation: $F_1(x) + U(F_1(x))$.

2) SELF-ATTENTION U-BLOCK

For image splicing detection and detection, both local and global contextual information are very important. Self-attention U-block is based on the self-attention module [25], which can be used to model global semantic information interaction from images. Specifically, the self-attention u-block is constructed by integrating self-attention to the bottom of the residual U-block, as shown in the right of Fig. 2, which connects each element in the highest-level feature map to access a global receptive field. That is to say, the decision of a particular pixel may be affected by all input pixels. To take into account the absolute context information, positional encoding is added on the highest-level feature map $\mathbf{X} \in \mathbf{R}^{d \times H \times W}$ shown in Fig. 4, where H and W represent the height and width of the feature map, respectively, and d is the number of channels. The positional encoding is particularly suitable for capturing the absolute and relative positions between the tampered areas in self-attention. The feature map \mathbf{X} is then reshaped to 2D feature map with size $n \times d$, where $n = H \times W$. And \mathbf{X} performs linear transform to obtain $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbf{R}^{d \times n}$ for query, key, value embedding. Embedded matrices are represented as $\mathbf{W}_q, \mathbf{W}_k$, and \mathbf{W}_v . The final output is a scaled dot-product (1):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} = \mathbf{A}\mathbf{V}. \quad (1)$$

Note that $\mathbf{A} \in \mathbf{R}^{n \times n}$ is called attention matrix, or similarity matrix. The attention matrix can be regarded as weights to account for the feature interactions between the keys and the queries, as shown in Fig. 4. In this way, this module can intrinsically achieve the global receptive field, and can explicitly model long-range semantic information interaction.

B. CROSS-ATTENTION MODULE

In order to make the low-level feature maps that are passed through the skip connections more expressive for better spatial recovery in decoder, we introduce cross-attention into each skip connections of our TransU²-Net, as shown in Fig. 2. Note that this module is not suitable for original resolution feature map for heavy computational cost. In order to ensure high-resolution information from skip connections, we utilize the cross-attention module to enhance the low-level feature maps under the guidance of high-level semantic information. The specific operations of our cross-attention module are shown in Fig. 5. We first add positional encoding on the low-level feature map $\mathbf{U} \in \mathbf{R}^{d \times 2H \times 2W}$, where H, W are the height and width of space and d is the number of channels. The feature map \mathbf{U} is then downsampled and used as the key of the cross-attention block. The high-level feature map $\mathbf{N} \in \mathbf{R}^{2d \times H \times W}$ after adding positional encoding and a 1×1 convolution serve as the query and key. Then the cross affinity matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$ is obtained by matrix multiplication and a Softmax function, where $n = H \times W$. Then we use the Relu activation function to re-scale the calculated weight value. The result is shown as \mathbf{S} in Fig. 5, which can be regarded as a filter, where the low magnitude element represents the irrelevant area that need to be reduced. The noise and non-semantic features is filtered out by the Hadamard product $\mathbf{U} \cdot \mathbf{S}$. Finally, the filtered features and high level features are concatenated. Through the cross-attention module, more detailed information can be retained than that of ordinary skip connections, thus improving the detection performance.

C. DECODING NETWORK

The structure of the decoder is similar to the encoder. For example, in De_5, the dilated RSU-4F is deployed and has

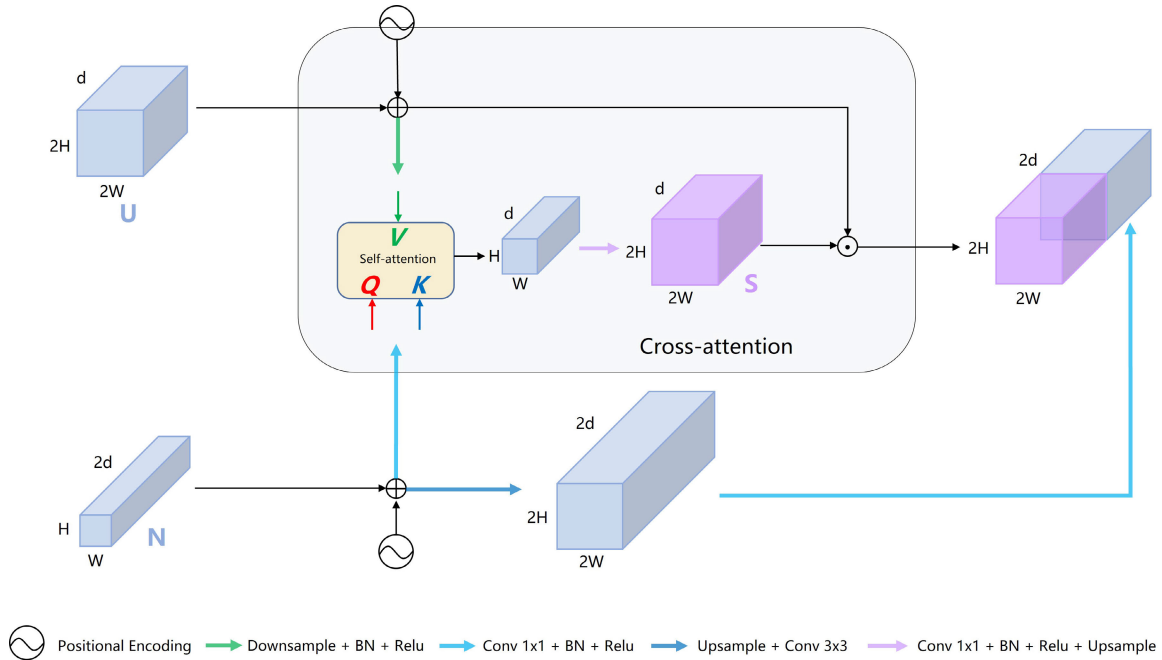


FIGURE 5. Illustration of cross-attention module.

a similar structure with En_5 in the encoder. The input of the decoder stage is the concatenation of the upsampled feature maps generated by its previous stage and the output feature maps of the current cross-attention module, as shown in Fig. 2. The final mask of tampered area is generated by a 3×3 convolution and a sigmoid operation at the end of the network.

In summary, our TransU²-Net design can capture rich multi-scale features, so as to locate tampered areas of different scales.

III. EXPERIMENT

This section mainly introduces various comparative experiments. Subsection III-A describes the datasets used in the experiment, the details of the experiment, and the evaluation metrics. Subsection III-B describes ablation experiments to verify the validity of self-attention and cross-attention. Subsection III-C describes the comparison results between TransU²-Net and several other existing methods for detecting image splicing forgery. Subsection III-D analyzes the robustness of TransU²-Net and other detection methods.

A. EXPERIMENTAL DETAILS

In this section, we present experimental settings. Specifically, subsection III-A1 presents details about used datasets. Subsection III-A2 presents details about implementation and subsection III-A3 shows employed evaluation metrics.

1) DATASETS

In this paper, we conduct experiments on two public datasets to analyze and evaluate image splicing forgery detection

TABLE 1. Precision, Recall and F-measure obtained by different U²-Net variants, bold values represent best results.

Methods	Precision	Recall	F-measure
U ² -Net	0.6031	0.6634	0.5932
U ² -Net-base-SA-1	0.6148	0.6312	0.5755
U ² -Net-base-SA-2	0.6136	0.5751	0.5532
U ² -Net-base-SA-3	0.6093	0.6225	0.5793
U ² -Net-base-SA-4	0.7258	0.7340	0.6946
U ² -Net-base-CA	0.6823	0.6260	0.6082
TransU²-Net(ours)	0.8086	0.7460	0.7351

methods, which are Casia [32], Columbia [33]. Casia is a popular dataset of image forgery localization, including images from multiple sources. Casia includes two types of tampering: splicing, copy-move. The tampered regions of image in Casia dataset are carefully manipulated and post-processed through filtering and blurring, which makes it more challenging. Since the ground truth mask is not officially available, we used a third-party mask [34] in our experiments. Images on the Casia dataset have a resolution of 384×256 and 640×480 . Columbia dataset only consists of splicing forgery, and has large but simple tampered areas. Images in the Columbia dataset have a typical resolution of 757×568 and 1152×768 . The corresponding ground truth mask is provided. We selected 1802 sets of splicing tampered images and randomly divided them into 1622 training data and 180 test data. Similarly, 180 groups of splicing tampered images were selected from the Columbia dataset and randomly assigned to 160 training data and 20 test data.

2) IMPLEMENTATION DETAILS

During the training process, our TransU²-Net is trained by using Adam optimizer [35]. The epoch of two stages are

TABLE 2. Comparison of splicing forgery detection results between this method and other detection methods.

Method	Detection Result					
	CASIA			COLUMBIA		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CFA [26]	0.144	0.697	0.202	0.747	0.599	0.584
ELA [27]	0.086	0.975	0.158	0.316	0.961	0.475
NOI [28]	0.149	0.992	0.258	0.422	0.997	0.593
ADQ [29]	0.402	0.585	0.476	0.367	0.998	0.536
NADQ [30]	0.278	0.455	0.285	0.329	0.225	0.238
U-Net [31]	0.734	0.629	0.626	0.820	0.816	0.779
U ² -Net [24]	0.603	0.663	0.593	0.915	0.803	0.808
U ² -Net-base-SA-1	0.615	0.631	0.576	0.859	0.801	0.804
U ² -Net-base-CA	0.682	0.626	0.608	0.876	0.858	0.857
RRU-Net [15]	0.678	0.586	0.586	0.836	0.904	0.855
ManTra-Net [16]	0.631	0.673	0.651	0.716	0.549	0.621
TransU²-Net(ours)	0.809	0.746	0.735	0.885	0.878	0.872

100 and 150, respectively. The initial learning rate for the first stage is set to 0.001 and that of the second stage is set to 0.0001. The weight decay is set to 0, and the batch size is set to 8. During the experiments, the proportion of validation sets is 0.1. We use Pytorch 1.8.1 as our training framework. Training and testing are conducted on RTX 3060 GPU (12GB memory).

3) EVALUATION METRICS

In order to quantitatively evaluate the pixel level performance of image splicing forgery detection methods, we use the precision, recall rate and F-measure as evaluation indicators.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

where TP denotes the numbers of correctly detected and FP denotes the numbers of erroneously detected tampered pixels and FN is the numbers of falsely missed pixels.

B. ABLATION EXPERIMENT

In order to verify the influence of attention, we evaluate the performance of the U²-Net in varied setups with the attention added progressively. All results in this section are based on experiments by using Casia dataset. As shown in Table 1, ‘U²-Net’ means the original U²-Net without attention. ‘U²-Net-base-SA-1’ means to replace the last RSU-4F of U²-Net encoder with self-attention U-block. ‘U²-Net-base-SA-2’ means to replace the all RSU-4F of U²-Net with self-attention U-block. ‘U²-Net-base-SA-3’ refers to adding a self-attention module after the last RSU-4F of U²-Net encoder. ‘U²-Net-base-SA-4’ refers to adding self-attention modules to the last two layers of all RSU encoders in U²-Net. ‘U²-Net-base-CA’ indicates U²-Net adding four cross-attention modules. ‘TransU²-Net’ means U²-Net combining self-attention and cross-attention. Comparing U²-Net-base-SA and U²-Net-base-CA and the original U²-Net, the performance is

gradually improved, which clearly shows the effectiveness of self-attention and cross-attention. And TransU²-Net has achieved the highest precision, recall, and F-measure.

C. COMPARATIVE EXPERIMENTS AND ANALYSIS

This subsection compares the performance of our proposed TransU²-Net and other methods.

To further evaluate the effectiveness of our TransU²-Net effectiveness, we selected five traditional detection methods ADQ [29], CFA [26], ELA [27], NOI [28], NADQ [30] and six detection methods RRU-Net, U-Net [31], ManTra-Net [16], U²-Net [24], U²-Net-base-CA, U²-Net-base-SA-1 based on deep learning to compare with the TransU²-Net proposed in this paper.

From the experimental results presented in Table 2, it can be seen that compared with deep learning based methods, the traditional methods have poor precision and F-measure value. NOI and ELA achieved very high recall value because they treated entire image as a tampered area. TransU²-Net achieved the highest precision and F-measure values on both Casia and Columbia datasets. In addition, the number of parameters of the proposed network, U-Net, and RRU-Net, U²-Net are 1,397,421, 13,395,329, 4,097,249, and 1,131,181 respectively, and U²-Net uses a smaller version. In summary, it can be seen that the proposed TransU²-Net has relatively small number of parameters and achieved high performance.

As shown in Fig. 6, six groups of data are selected from the test dataset as examples, of which the first to fourth groups of data are from Casia dataset, the fifth and the sixth groups of data are from Colombia dataset. The first row and the second row are the tampered image and the ground truth, respectively. All the remained rows are the detection results of other nine methods and TransU²-Net. From the experimental results, ADQ algorithm has a certain detection effect on the splicing tampering of Casia dataset, while the splicing tampering detection effect of Columbia dataset is very poor. CFA algorithm can better locate the tampered area in Columbia dataset, but it is difficult to locate the tampered area in caisa dataset image. Compared with traditional methods, the deep learning methods can effectively locate the

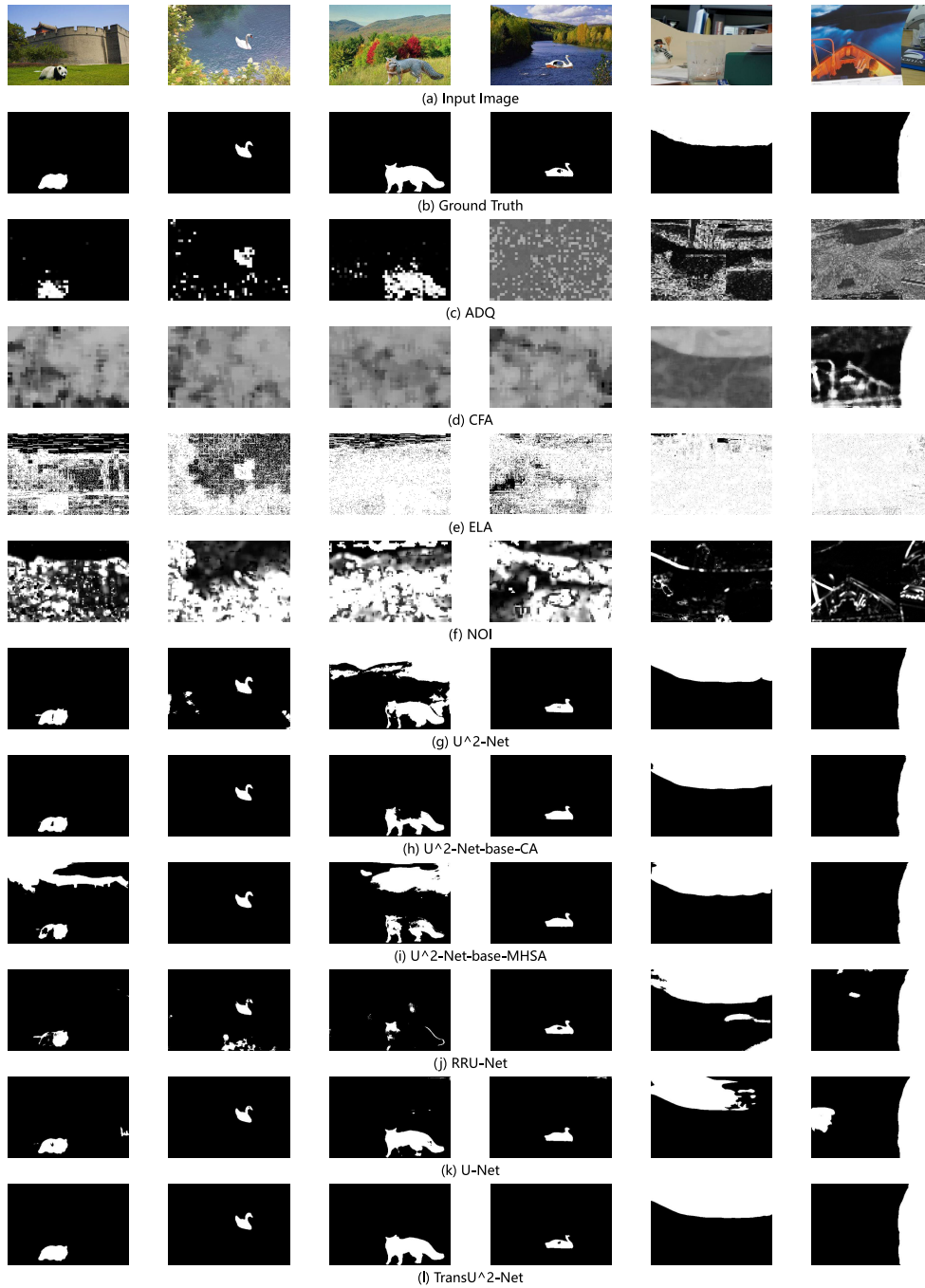


FIGURE 6. Comparison of splicing forgery detection results between the proposed TransU²-Net and other detection methods.

tampered area for Casia dataset and Columbia dataset. However, U-Net and RRU-Net still have many problems such as wrong segmentation, excessive segmentation and incomplete localization. Finally, our proposed TransU²-Net can more accurately locate the tampered areas of different sizes.

D. ROBUSTNESS ANALYSIS

To further evaluate the robustness of TransU²-Net, we compare the performance of the TransU²-Net and other detection

methods in Casia and Columbia datasets under two types of attacks: Gaussian noise attack and JPEG compression.

1) EXPERIMENTAL RESULTS OF ROBUSTNESS AGAINST GAUSSIAN NOISE ATTACK

This section discusses and evaluates the robustness of TransU²-Net and other detection methods against Gaussian noise attack. As shown in Fig. 7, we compare different evaluation metrics of the Casia dataset images under different

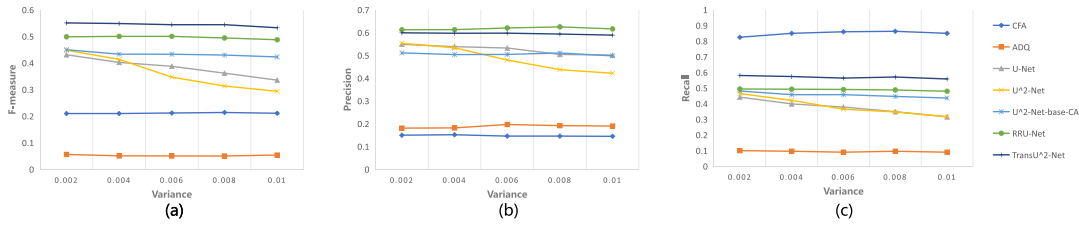


FIGURE 7. Comparison of experimental results of different methods for adding Gaussian noise to Casia dataset.

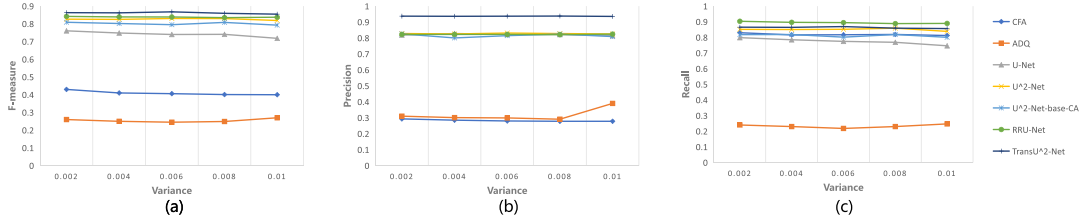


FIGURE 8. Comparison of experimental results of different methods for adding Gaussian noise to Columbia dataset.

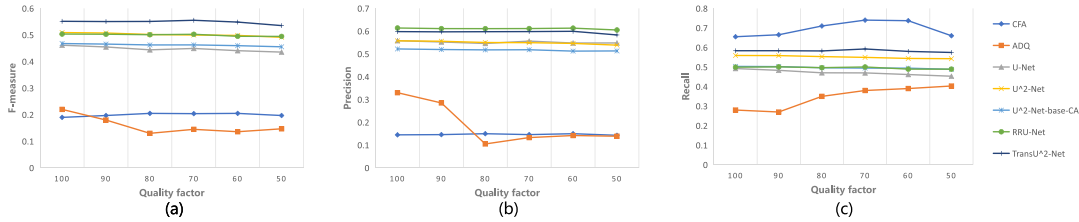


FIGURE 9. Comparison of experimental results of JPEG compression attacks by different methods in Casia dataset.

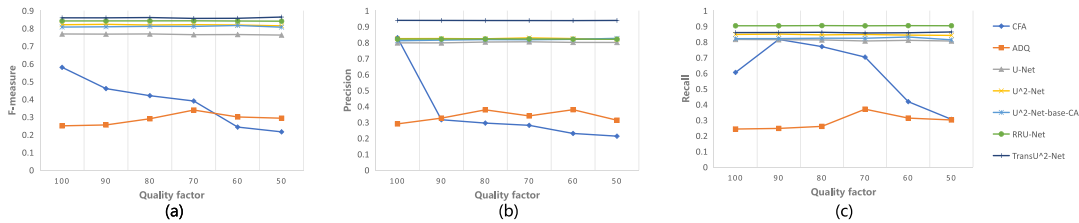


FIGURE 10. Comparison of experimental results of JPEG compression attacks by different methods in Columbia dataset.

variance of Gaussian noise attacks. Also as shown in Fig. 8, we compare different evaluation metrics of Columbia dataset images with different variance of Gaussian noise attacks. From the experimental results, it can be seen that with the increase of variance of Gaussian noise, each evaluation index of different detection methods has a certain degree of reduction. As shown in the Fig. 7 and Fig. 8, with the increase of variance of Gaussian noise, the Gaussian noise attack has some impact on the deep learning methods, but its performance is still better than traditional detection methods. On Casia dataset and Columbia dataset, the precision and F-measure of TransU²-Net are far better than those of other six detection methods. The recall value of the TransU²-Net is slightly lower than the RRU-Net. In addition, with the increase of variance of Gaussian noise, the detection index

of TransU²-Net is least affected. This experiment proves that the proposed TransU²-Net is robust against noise attacks on all two datasets.

2) EXPERIMENTAL RESULTS OF ROBUSTNESS AGAINST JPEG COMPRESSION

This section discusses and evaluates the robustness of TransU²-Net and other detection methods against JPEG compression attack. In Fig. 9, we compare different evaluation metrics of Casia dataset images under distinct JPEG compression. In Fig. 10, we compare different evaluation metrics of Columbia dataset images under distinct JPEG compression. We can see from these figures that the JPEG compression has some influence on image tampering detection. When the quality factor gradually decreases, the precision, recall, and

F-measure values of CFA and ADQ is reduced to a very low value, but the impact on the method based on deep learning is very slight. The deep learning method has a strong resistance to compression attacks because it is a detection method based on image content features. Therefore, compression attacks have little impact on the deep learning method. In addition, as the quality factor is reduced from 100 to 50, the TransU²-Net presented in this paper is still superior to other methods and robust to both Casia and Colombia datasets.

IV. CONCLUSION

In this paper, we propose a novel hybrid Transformer architecture named TransU²-Net, which can locate image tampered regions. Firstly, TransU²-Net integrates both self-attention and cross-attention into U²-Net. Therefore, the TransU²-Net can avoid relying heavily on large-scale pre-training, and it has the ability to learn explicit long-range semantic information dependencies. Compared with previous methods, our model can locate forged areas that have different scales. Secondly, to enhance the low-level feature maps that are passed through the skip connections and achieve fine spatial recovery in TransU²-Net decoder, we designed the cross-attention module. Therefore, it is more general and effective for complex image forgery. Finally, Experiments on two standards image forgery datasets show that our method achieves better performance than state-of-the-art methods.

Although our method achieves better performance in terms of precision, recall rate and F-measure without using large-scale pre-training, there are still some problems to be solved, such as: 1) our model is not good at generalization for cross dataset testing; 2) The detection capability is relatively simple and only applicable to the detection of splicing images. On this basis, our future work will focus on addressing these issues. Specifically, 1) improve the universality of the algorithm model by training the model on high-quality data sets with multiple tampering methods; 2) Combining the idea of transfer learning, an effective detection model suitable for cross dataset training and detection can be proposed to improve the generalization ability of the model.

DATA AVAILABILITY STATEMENT

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

CONFLICTS OF INTERESTS/COMPETING INTERESTS

The authors have no competing interests to declare that are relevant to the content of this article.

REFERENCES

- [1] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," *Multimedia Tools Appl.*, vol. 76, no. 4, pp. 4801–4834, Feb. 2017.
- [2] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, Feb. 2005.
- [3] J. Lukáš, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006.
- [4] M. Chen, J. Fridrich, M. Goljan, and J. Lukáš, "Determining image origin and integrity using sensor noise," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 74–90, Mar. 2008.
- [5] X. Pan, X. Zhang, and S. Lyu, "Exposing image forgery with blind noise estimation," in *Proc. 13th ACM Multimedia Workshop Multimedia Secur.*, Sep. 2011, pp. 15–20.
- [6] P. Korus and J. Huang, "Multi-scale analysis strategies in PRNU-based tampering localization," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 4, pp. 809–824, Apr. 2016.
- [7] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 487–495.
- [8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [10] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [12] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1053–1061.
- [13] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4970–4979.
- [14] H. Chen, C. Chang, Z. Shi, and Y. Lyu, "Hybrid features and semantic reinforcement network for image forgery detection," *Multimedia Syst.*, vol. 28, pp. 363–374, May 2021.
- [15] X. Bi, Y. Wei, B. Xiao, and W. Li, "RRU-Net: The ringed residual U-Net for image splicing forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 30–39.
- [16] Y. Wu, W. Abdalmegeed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9543–9552.
- [17] B. Liu, R. Wu, X. Bi, B. Xiao, W. Li, G. Wang, and X. Gao, "D-UNet: A dual-encoder U-Net for image splicing forgery detection and localization," 2020, *arXiv:2012.01821*.
- [18] X. Bi, Z. Zhang, Y. Liu, B. Xiao, and W. Li, "Multi-task wavelet corrected network for image splicing forgery detection and localization," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.
- [19] M.-J. Kwon, I.-J. Yu, S.-H. Nam, and H.-K. Lee, "CAT-Net: Compression artifact tracing network for detection and localization of image splicing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 375–384.
- [20] Y. Rao, J. Ni, and H. Xie, "Multi-semantic CRF-based attention model for image forgery detection and localization," *Signal Process.*, vol. 183, Jun. 2021, Art. no. 108051.
- [21] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization," 2021, *arXiv:2103.10596*.
- [22] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.
- [23] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 312–328.

- [24] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U²-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
- [26] P. Ferrara, T. Bianchi, A. D. Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.
- [27] N. Krawetz and H. F. Solutions, "A picture's worth," *Hacker Factor Solutions*, vol. 6, no. 2, p. 2, 2007.
- [28] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1497–1503, Sep. 2009.
- [29] Z. Lin, J. He, X. Tang, and C.-K. Tang, "Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis," *Pattern Recognit.*, vol. 42, no. 11, pp. 2492–2501, 2009.
- [30] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.
- [31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2015, pp. 234–241.
- [32] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2013, pp. 422–426.
- [33] Y.-F. Hsu and S.-F. Chang, "Detecting image splicing using geometry invariants and camera characteristics consistency," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2006, pp. 549–552.
- [34] N. Pham, J.-W. Lee, G.-R. Kwon, and C.-S. Park, "Hybrid image-retrieval method for image-splicing validation," *Symmetry*, vol. 11, no. 1, p. 83, Jan. 2019.
- [35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.



CAIPING YAN (Member, IEEE) received the Ph.D. degree in computer science from the University of Macau, China, in 2017. She is currently with Hangzhou Normal University, Hangzhou, China. Her current research interests include digital forensics and image processing.



SHUYUAN LI received the B.S. degree from Hangzhou Dianzi University, Hangzhou, China, in 2020. He is currently pursuing the M.S. degree with the Department of Computer Science, Hangzhou Normal University, Hangzhou. His research interest includes digital forensics.



HONG LI (Member, IEEE) received the Ph.D. degree from the Department of Computer and Information Science, University of Macau, Macau, China, in 2017. His research interests include image processing, semi-supervised learning, and computer vision.

...