

Received 7 April 2024, accepted 5 June 2024, date of publication 13 June 2024, date of current version 8 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3413709

## RESEARCH ARTICLE

# Dual-Stream Intermediate Fusion Network for Image Forgery Localization

CAIPING YAN<sup>1</sup>, (Member, IEEE), RENHAI LIU<sup>1</sup>, HONG LI<sup>1,2</sup>, JINGHUI WU<sup>1</sup>, AND HAOJIE PAN<sup>1</sup>

<sup>1</sup>School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China

<sup>2</sup>Hangzhou 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** Nowadays, powerful image editing applications not only simplify image processing significantly but also enhance the realism of processed digital images. However, this convenience has presented unprecedented challenges in verifying the authenticity of images. Although existing methods have achieved significant results in image forgery localization, most of them struggle to obtain satisfactory performance when dealing with tampered areas of various sizes, especially for large-scale tampered regions. To enhance the localization performance for various types and sizes of tampered regions, we propose a novel dual-stream intermediate fusion network for image forgery localization, named DIF-Net. This network adopts an encoder-decoder architecture composed of an adaptive convolutional pyramid and dual-stream intermediate fusion modules. Specifically, the former extracts multi-scale information from different depths by utilizing two depth-wise strip convolutions instead of standard large-kernel convolutions. Moreover, during feature fusion, learnable parameters are employed to dynamically allocate weights to each feature scale, so that the network can adaptively select the most relevant features at the target scale. The latter effectively reduces category information differences between the two feature streams by utilizing two learnable intermediate representations to model channel and spatial consistency in the dual-stream features. Compared to traditional and previous deep learning methods, the DIF-Net can generate high-quality prediction masks with fewer parameters. Through extensive experimental validation, our DIF-Net demonstrates outstanding performance on various datasets, surpassing the state-of-the-art forgery localization methods currently available. On the commonly used CASIA2 dataset, our DIF-Net achieves an improvement of 3.3% in F1 and 2.4% in AUC compared to previous methods.

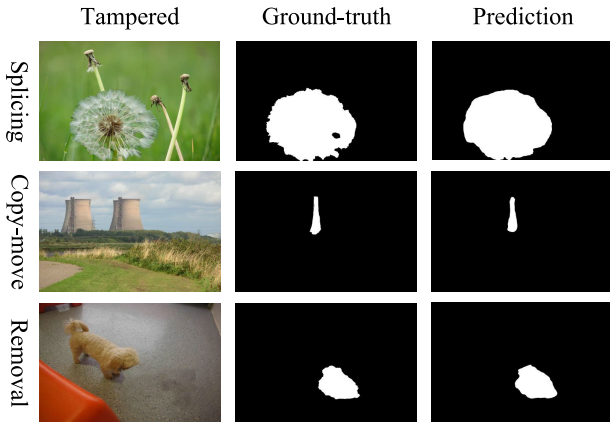
**INDEX TERMS** Information security, image forensics, image forgery localization, dual-stream network, feature fusion.

## I. INTRODUCTION

With the rapid advancement of digital technology, a large number of user-friendly image editing applications with simple operations and excellent effects have been widely adopted by the general public. These editing tools allow users to easily and effectively generate realistic images without requiring professional knowledge. When the general public uses these editing tools, they typically only alter the contrast

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

and colors of photos, thereby enhancing and beautifying them without changing the content of the images. However, malicious actors manipulate the content of images and use manipulated visuals for malicious purposes, such as creating fake news, spreading rumors, and fabricating false evidence. These actions can have a significant impact on social security, information credibility, and legal integrity, among other factors. There are three main techniques commonly used in image manipulation [1], [2], [3], [4]: (1) splicing, (2) copy-move, and (3) removal. As shown in Fig. 1, splicing involves pasting a region from one image onto another,



**FIGURE 1.** Three common examples of tampered images and the localization results of the proposed method.

copy-move refers to the operation of copying and pasting a specific region in an image to another location, and removal erases a region from the image and fills it based on the surrounding environment. In the real world, distinguishing between the tampered areas of a forged image and the original areas using only the naked eye and traditional techniques is difficult. This is because after an image is tampered with, it undergoes various post-processing operations to conceal alteration traces, such as scaling, contrast adjustment, rotation, and blurring. Therefore, the development of a robust and efficient pixel-level image forgery localization method to accurately identify the modified portions from suspected tampered images has become an urgent need in the current context.

Over the years, a variety of methods have emerged in the field of image tampering localization and detection, including traditional approaches such as Discrete Cosine Transform (DCT) [5], Color Filter Array (CFA) [6], and Steganalysis Rich Model (SRM) [7]. These methods manually acquire local differences between tampered and untampered regions such that their detection capabilities are limited to specific types. In the face of complex types of image manipulation in the real world, these methods are not suitable for practical applications. In recent years, deep learning has gradually become an important tool for image tampering detection with its strong feature learning ability and end-to-end advantages. Among them, the methods [8], [9], [10] using the traditional encoder-decoder structure have achieved remarkable success in image tampering detection. Still, these methods ignored the importance of multi-scale features, thus limiting the further improvement of model performance. Other researchers have recognized the significance of multi-scale feature fusion. Their proposed methods [11], [12], [13], [14], [15] have effectively integrated shallow and deep features to achieve interaction among multi-scale features, which enabling the model to better comprehend images. While these methods obtain multi-scale features across different levels, they do not extract multi-scale features from each level. In addition,

some methods [13], [16], [17], [18] attempt to combine multi-modal features to extract more useful information, especially from RGB invisible traces. However, these methods may overlook feature disparities during feature fusion, potentially impacting the model's performance.

To address the problems in the aforementioned works, we propose DIF-Net for pixel-level image forgery localization. DIF-Net consists of a dual-stream adaptive convolution pyramid encoder and an intermediate channel spatial fusion decoder. In [19], [20], and [21] large kernel convolutions are widely used to increase the effective receptive field of the model and have achieved excellent results in multiple visual tasks. To this end, in DIF-Net, we propose the Adaptive Convolution Pyramid Module, which extracts multi-scale features using convolution kernels of different sizes. Additionally, in the feature fusion stage, we introduce a self-learning weight parameter to enable the network to automatically select scale features suitable for the target. Finally, we employ strip convolutions instead of traditional large kernel convolutions, enhancing the receptive field while reducing computational complexity. For the overall network, the pyramid module in the encoder facilitates multi-scale feature extraction at each level, and skip connections aid in the effective interaction of multi-scale information across hierarchies, thereby fully leveraging multi-scale information. Most existing works train models using multi-modal features, such as noise domain, frequency domain, and color space. However, these methods tend to simplify the fusion of multiple features, overlooking the differences in category information among various features. In our study, we employ images in two color spaces, RGB and HSV, and progressively fuse features through an intermediate spatial fusion module to reduce feature category information disparities.

For DIF-Net, we initially input images in both RGB and HSV color spaces and separately extract multi-scale dual-stream features using two adaptive convolution pyramid encoders with non-shared weights. Subsequently, the shallow-level detail features from the encoder are fed through skip connections into the intermediate channel spatial fusion module, where they are fused with deep semantic features. Drawing inspiration from [22]'s small-sample segmentation, CBAM [23] and SENet [24] attention mechanism, we incorporate two sets of learnable parameters as intermediate prototypes in each fusion module of the decoder. These parameters model the consistency in both channel and spatial dimensions of the dual-stream features to reduce their category information differences and establish long-range dependencies. Thus, it further enhances the network's localization performance. DIF-Net does not require any pre-processing or post-processing and can be trained end-to-end directly. We conducted experiments on four prominent image forgery datasets, including CASIA2 [1], Columbia [2], NIST16 [3], and IMD2020 [4]. The experimental results demonstrate that this method outperforms the state-of-the-art forgery localization methods.

Our primary contributions are as follows:

- We propose DIF-Net, an end-to-end encoder-decoder architecture for image manipulation localization. It combines RGB and HSV features to identify tampered regions, achieving superior image manipulation localization performance on standard datasets compared to state-of-the-art methods.
- We propose an adaptive convolution pyramid module with a multi-branch depth-wise strip convolution. Strip convolution replaces conventional large kernel convolution for multi-scale feature extraction, tailored to tamper regions of varying sizes. Additionally, self-learning weights are used for selecting multi-scale features, enhancing the network's capability to leverage these features effectively.
- To reduce the feature category disparities resulting from the direct fusion of dual-stream features, we propose a novel intermediate channel spatial fusion module. Within each fusion module, two sets of learnable parameters are separately employed as channel and spatial intermediate prototypes, thereby establishing channel and spatial consistency.

The remaining sections of this paper are organized as follows: Sec. II comprises a review of relevant research work in the field of image tampering detection. Sec. III presents the detailed design of the proposed method. Sec. IV demonstrates the experimental results and performance analysis. Finally, Sec. V provides a summary of the work presented in this paper.

## II. RELATED WORK

### A. IMAGE MANIPULATION LOCALIZATION

Most previous work has primarily focused on the detection or localization of a single type of manipulation, searching for different traces of manipulation in the image based on different tampering types to determine the authenticity of the image. Examples of such methods include Color Filter Array (CFA) [6] local noise inconsistencies [25], [26], [27], [28], JPEG compression artifacts [5], [29], [30], [31], etc. These methods often rely on manual feature extraction and perform well for specific tampering features. If post-processing has hidden or disrupted certain specific traces in the image, it will lead to a significant decrease in the performance of these methods.

At present, the main work uses deep learning for image tampering localization, which not only can achieve better localization results, but also has better robustness. Salloum et al. introduced a multi-task fully convolutional network (MFCN) [10], which accomplishes manipulation localization by learning to splice masks and mask edges, but it overlooks multi-scale information. Zhou et al. [16] proposed a dual-stream network (RGBN) for image tampering detection, which employs both noise stream and RGB stream. This algorithm uses steganalysis rich model (SRM) to extract noise information, indicating that the participation of noise stream

can better reflect tampering traces. However, the localization results are only coarse bounding boxes instead of accurate pixel-wise masks. Inspired by the recall and consolidation mechanism of the human brain, Bi et al. [12] proposed the Circular Residual U-Net [32] (RRU-Net) to enhance the CNN learning method through the process of residual propagation and feedback. The problem of localizing image manipulation was addressed as a local anomaly detection task in the study conducted by Wu et al., resulting in the development of ManTra-Net [9], which utilizes VGG [33] and Z-pooling techniques for precise localization of anomalous regions. Building upon ManTra-Net, SPAN [14] constructs a pyramid structure using self-attention blocks and dilated convolutions to model pixel-level representations at multiple scales. PSCC-Net [11] processes forged images through two paths, one from top to bottom and the other from bottom to top. The former extracts features at different scales, while the latter generates prediction masks at four scales, progressively from coarse to fine. To better distinguish the feature differences between tampered and untampered regions in images, CFL-Net [34] utilizes both contrastive loss and cross-entropy loss simultaneously. Zhang et al. [15] extract image and label edges and use global edge information to guide the network in learning label masks, thereby enhancing the network's localization results. To address the issue of limited training data in image forgery detection, Zhou et al. [35] proposed an adversarial training strategy and used self-attention mechanisms to locate tampered regions. To capture subtle manipulation traces that are no longer visible in the RGB domain, Wang et al. proposed Objectformer [18]. It combines high-frequency features from the image with RGB features to create multimodal patch embeddings. Das et al. [17] proposed a novel gated context attention network (GCA-Net) for forgery localization. It extracts multimodal features and utilizes non-local attention along with gating mechanisms to capture finer image differences.

In this work, firstly, we employ images of two color spaces as input, and use adaptive convolution pyramids in the encoder to extract features with different scales in each feature map and perform adaptive aggregation. Secondly, we input the shallow features of the dual branches into the fusion module of the decoder through skip connections. In each fusion module, two sets of learnable parameters are used as intermediate prototypes to model the channel and space consistency respectively to reduce the feature difference and generate channel and space weights. Finally, the weights are used to optimize the fused two-stream shallow features in turn, and the deep features of the decoder and the optimized ultra-shallow features are fused and output.

### B. ATTENTION MECHANISM

It is well known that attention plays a crucial role in the human visual system. In order to make rational use of the limited visual resources, humans will select the salient part when paying attention to things and then focus on it.

By imitating this mechanism, researchers have introduced an attention mechanism into deep learning to suppress useless information and highlight important information. Attention has been widely used in the visual field, which is mainly divided into channel attention, spatial attention, and mixed attention according to the field of use. Many recent works have proposed multiple kinds of attention, and channel attention is introduced in SENet [24], which focuses more on important features by assigning different weights to different channels and supports plug-and-play. SENet ignores the information interaction in the feature graph space. CBAM [23] adds a kind of spatial attention on the basis of it, and forms a comprehensive feature attention method by combining channel and spatial attention. In order to solve the inefficiency and optimization difficulty of convolution and recursion operations in capturing long-range dependencies. In deep neural networks, they [36], [37] employ a simple and efficient non-local operation to capture long-range dependencies. Aiming at the problem of the high computational complexity of non-local operations, a cross-attention method is proposed in [38], which obtains the information interaction of each pixel for all pixels of the feature map by stacking two attention blocks. Fu et al. [39] by combining channel and spatial self-attention, a comprehensive dual attention is proposed. Inspired by attention SENet [24] and CBAM [23], we generate channel and spatial weights after using intermediate prototypes to model the consistency of channel and space, respectively, to optimize the fused features for channel and space in turn.

### III. PROPOSED METHOD

#### A. OVERVIEW

In this section, we will provide a detailed explanation of the proposed image forgery localization method. The architecture of this network is illustrated in Fig. 2, comprising the Adaptive Convolution Pyramid module and the Intermediate Channel Spatial Fusion module, which together form a complete encoder-decoder structure. Due to the post-processing applied to tampered images to hide forgery traces, which can be challenging to detect in RGB images, some prior research [40], [41], [42], has proposed using the HSV color space for image forgery detection. Therefore, we also introduce the HSV color space to provide supplementary clues for forgery localization.

In DIF-Net, we first convert the given RGB manipulated image into the HSV color space image. Next, we input these two color space images into two adaptive convolution pyramid encoders with non-shared weight parameters to obtain dual-stream features. In the encoder, we extract features from different scales and adaptively aggregate them to enhance the representation capability. Subsequently, we progressively fuse the dual-stream shallow details from the encoder with the deep semantics from the decoder through the Intermediate Channel Spatial Fusion module. It is worth noting that we do not perform feature fusion in the final

layer. Instead, we directly use the convolutional output head to output pixel-level prediction masks. Through experiments, we have found that this design not only greatly reduces the model parameters but also enables the model to predict the tampered regions in the image more accurately at the pixel level.

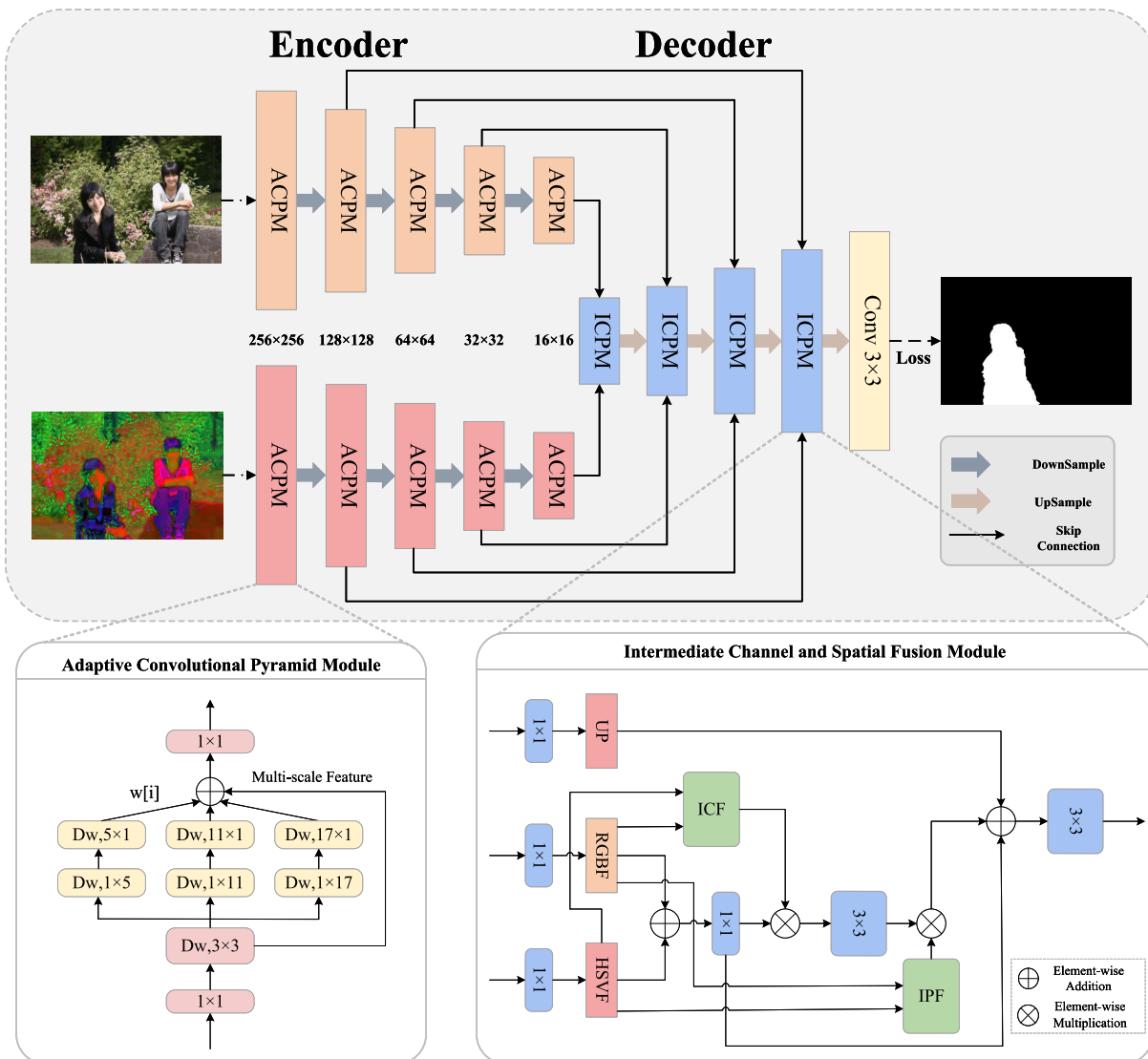
#### B. ADAPTIVE CONVOLUTION PYRAMID ENCODER

In the encoder stage, we use both RGB and HSV images as input. Compared to the RGB color space, the HSV color space exhibits better color separation properties, which facilitates more precise identification and analysis of different color characteristics. When images are tampered with, differences may exist in the hue, color saturation, and brightness between the tampered and untampered parts. Therefore, by leveraging these characteristics of the HSV color space, we can capture the features of image tampering more accurately. All the time, designing multi-scale networks has been a prominent area of research in computer vision [43], [44], [45], [46]. In tampered images, the size of the manipulated regions can vary significantly. For instance, in the image size is  $256 \times 256$ , small manipulated regions may consist of just a few dozen pixels, whereas large manipulated regions may encompass tens of thousands of pixels. This poses higher demands on pixel-level localization methods. To address varying sizes of manipulated regions, we propose a multi-scale architecture constructed using convolution operations of different sizes, known as the Adaptive Convolution Pyramid Module (ACPM), to obtain multi-scale features at each level. Simultaneously, to tackle the issue of excessive computational complexity associated with using large-kernel convolutions, we have employed depth-wise strip convolutions in each branch as a substitute for standard convolutions. As depicted in Fig. 2, we illustrate the structure of the dual-branch encoder, which consists of the ACPM and max-pooling. The specific structure of ACPM is shown in the bottom left corner of Fig. 2. Initially, we increase the network's depth through a  $1 \times 1$  convolution. Subsequently, we employ a  $3 \times 3$  convolution to extract local information and capture information at different scales using multi-branch depth-wise strip convolutions. Finally, to better fuse multi-scale features, we introduce learnable weights to dynamically allocate weights to different scale features. After completing the adaptive fusion of multi-scale features, a  $1 \times 1$  convolution is employed to model relationships between different channels and adjust the output channels. The ACPM can be described as follows:

$$F_{out} = Conv_{1 \times 1} \left( \sum_{i=0}^3 w_i \cdot Branch_i(DwC(Conv_{1 \times 1}(F))) \right) \quad (1)$$

where  $F$  represents the feature of the previous stage after downsampling. Meanwhile,  $DwC$  stands for depth-wise strip convolution, where depth-wise convolution can reduce the number of model parameters.  $Branch_i$  and  $w_i$  respectively denote the  $i$ -th branch and its corresponding weight.



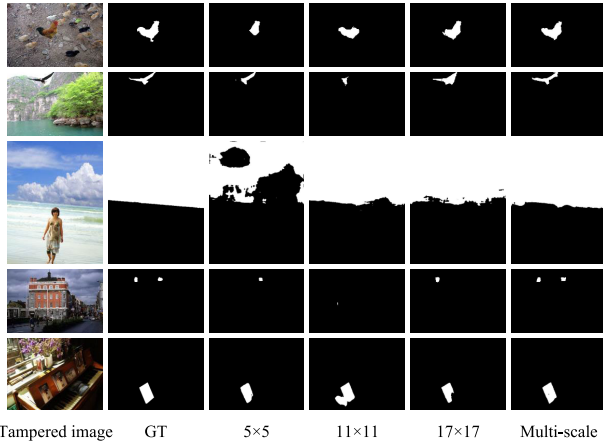


**FIGURE 2.** The proposed architecture of DIF-Net. Each encoder network contains five ACPM blocks and four  $2 \times 2$  Max pooling layers. The decoder network contains four ICPM, four bilinear interpolation upsampling layers, and a  $3 \times 3$  convolution for predicting the output mask. UP, RGBF, and HSVF respectively represent the upsampled feature, RGB stream feature, and HSV stream feature after the number of channels is reduced to  $1/r$  after  $1 \times 1$  convolution. Each  $1 \times 1$  convolution in the figure is followed by batch normalization and GELU.  $\otimes$  and  $\oplus$  respectively denote element-wise multiplication and addition of the matrix.

Here,  $Branch_3$  denotes the identity connection operation. To address the issue of excessive computational burden that may arise from using large-kernel convolutions, we utilize two stacked strip convolutions in each branch as a replacement for standard convolution. Since the size of the feature map is  $16 \times 16$  after four downsampling, the maximum size of the large kernel convolution is set to 17. For each branch, we set the kernel size to be 5, 11, and 17, respectively. In Fig. 3, we visualized the model’s results using multi-scale convolutions and single-scale convolutions in the encoder. From the visual results, it can be observed that multi-scale convolutions greatly improved the completeness of forged region localization. Following the  $3 \times 3$  convolution and the  $1 \times 1$  convolution for feature fusion, we applied batch normalization and the GELU.

**C. INTERMEDIATE FUSION DECODER**

Fig. 2 bottom right corner illustrates the overall structure of ICPM. We input the multi-scale shallow features from the encoder and the deep upsampled features from the decoder into ICPM through skip connections. The sizes of these three input features are  $H \times W \times C$ . To reduce computational complexity while establishing channel and spatial consistency, three  $1 \times 1$  convolutions are used to reduce the number of channels in the input feature map, obtaining feature maps RGBF, HSVF, and UP with sizes of  $H \times W \times C_1$  (where  $C_1 = C/r$ ,  $r$  is the scaling factor) (i.e.,  $F_r$ ,  $F_h$ , and  $F_u$ ). Next, we separately input  $F_r$  and  $F_h$  into the Intermediate Channel Fusion module (ICF) and Intermediate Spatial Fusion module (IPF) to generate channel attention weight matrix  $A_c$  and spatial attention



**FIGURE 3.** For the visualization results using single-scale and multi-scale convolutions.

weight matrix  $A_p$ . At the same time, we sum  $F_r$  and  $F_h$ , and perform information interaction and feature fusion between channels using a  $1 \times 1$  convolution, resulting in feature map  $Y$ . To reduce the differences between features and generate fine-grained features, we element-wise multiply the generated intermediate channel attention matrix  $A_c$  and intermediate spatial attention matrix  $A_p$  with the fused feature map  $Y$ . It is noteworthy that in the process of applying attention mechanisms, we do not employ parallel channel attention and spatial attention. Instead, we first utilize channel attention, followed by feature fusion through a  $3 \times 3$  convolution. Finally, we merge the upsampled feature  $F_u$ , feature  $Y$ , and the feature post-channel attention through a  $3 \times 3$  convolution for the final output fusion. The whole process can be described as:

$$Out = Conv_{3 \times 3}(A_p \cdot Conv_{3 \times 3}(A_c \cdot Y) + Y + F_u) \quad (2)$$

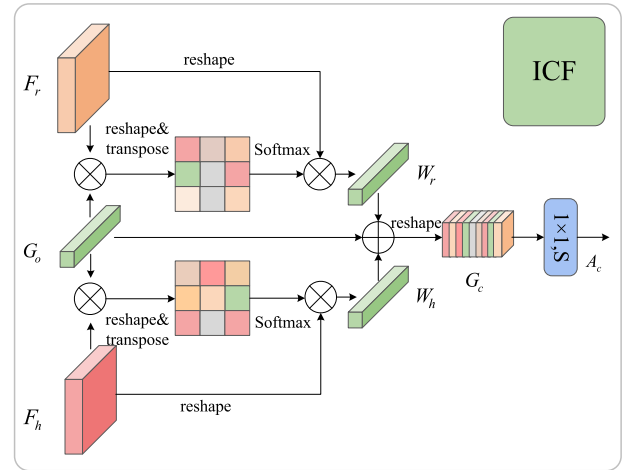
where  $(\cdot)$  denotes element-wise multiplication. For each  $3 \times 3$  convolution, both batch normalization and the GELU are applied afterward.

### 1) INTERMEDIATE CHANNEL FUSION

Each channel map of the high-level features can be considered as distinct semantic features, and these different semantic features are interrelated with each other. By exploiting the interdependence among channels, we can strengthen the interdependent feature mappings and reduce the disparities among feature categories. To achieve this, we introduce the Intermediate Channel Fusion module (ICF), which utilizes a set of learnable parameters

$G_{o1} \in R^{C \times 1}$  as intermediate prototypes for channels. It weights and adjusts the features based on the relationships between different channels to model the interdependence among channels and the consistency of feature categories, generating the channel attention matrix.

The ICF, as shown in Fig. 4, we first reshape the input features  $F_r$  and  $F_h$  into  $F'_r$  and  $F'_h$  and transpose them to  $F'^T_r$  and  $F'^T_h$ , in particular,  $\{F'_r, F'_h\} \in R^{C \times 1 \times HW}$ . Then, a matrix



**FIGURE 4.** The structure of ICF. Here  $\otimes$  represents the matrix multiplication and  $\oplus$  the element-wise addition.  $S$  denotes the Sigmoid function.

multiplication is performed between  $G_{o1}$  and  $F'^T_r$ , followed by the application of the Softmax function to compute the channel attention map. Subsequently, another matrix multiplication is applied between  $F_r$  and the channel attention map, resulting in the matrix  $W_r \in R^{C \times 1 \times 1}$ . Simultaneously, similar operations are performed for  $G_{o1}$ ,  $F_r$ , and  $F_h$  to obtain  $W_h \in R^{C \times 1 \times 1}$ . Finally, the  $W_r \in R^{C \times 1 \times 1}$ ,  $W_h \in R^{C \times 1 \times 1}$ , and  $G_{o1}$  summation results are reshaped into  $R^{C \times 1 \times 1}$  and the channel attention matrix  $A_c$  is generated using  $1 \times 1$  convolution, batch normalization, and Sigmoid functions. The can be described as follows:

$$W_i = F'_i \text{Softmax}(F'^T_i G_{o1}) \quad i \in \{r, h\} \quad (3)$$

$$A_c = \text{Sig}(\text{BN}(\text{Conv}_{1 \times 1}(\text{reshape}(W_r + W_h + G_{o1})))) \quad (4)$$

where  $\text{Sig}$  and  $\text{Softmax}$  represents Sigmoid and Softmax functions respectively, and  $\text{BN}$  is batch normalization.  $\text{Conv}_{1 \times 1}$  denotes  $1 \times 1$  convolution.

### 2) INTERMEDIATE SPATIAL FUSION

In order to establish distant contextual information to differentiate feature representations of forged regions from original regions and prevent overfitting to specific features during training, we introduce the Intermediate Spatial Fusion module (IPF). This module encodes contextual information from different branches into local features, enhancing their representational capacity. Simultaneously, IPF dynamically updates the spatial positions' weights by weighting and aggregating information from all positions of dual-stream features. This helps suppress redundant information, enabling the network to selectively focus on tampered regions. As shown in Fig. 5. Similar to the Intermediate Channel Fusion module, in the Intermediate Spatial Fusion module, we also employ a set of learnable parameters  $G_{o2} \in R^{1 \times HW}$  as spatial prototypes. The process of generating the spatial attention matrix is similar to that of the channel attention

**TABLE 1.** Summary of the dataset we used (✓ and ✗ indicate where or not the manipulation type is involved).

Dataset	Train	Val	Test	Total	Splicing	Copy-move	Removal
CASIA2	4096	512	512	5120	✓	✓	✗
Columbia	144	18	18	180	✓	✗	✗
NIST16	451	57	56	564	✓	✓	✓
IMD2020	1608	201	201	2010	✓	✗	✗

matrix and can be concisely described as follows:

$$W_i = F_i' \text{Softmax}(G_{o2} F_i'^T) \quad i \in \{r, h\} \quad (5)$$

$$A_p = \text{Sig}(\text{Conv}_{3 \times 3}(\text{reshape}(W_r + W_h + G_{o2}))) \quad (6)$$

where *Sig* and *Softmax* represents Sigmoid and Softmax functions respectively,  $\text{Conv}_{3 \times 3}$  denotes  $3 \times 3$  convolution.

#### D. LOSS FUNCTION

In pixel-level forgery localization, the area of the manipulated region is often much smaller than the original region, leading to class imbalance. To address this issue, we use both dice loss [47] and binary cross-entropy loss as the final loss function for the model. The dice loss is represented as follows for the given predicted mask *P* and ground truth mask *M*:

$$\mathcal{L}_{dice} = 1 - \frac{2 \times \sum (M_{(i,j)} \cdot P_{(i,j)}) + \epsilon}{\sum M_{(i,j)} + \sum P_{(i,j)} + \epsilon} \quad (7)$$

where  $M_{(i,j)}$  and  $P_{(i,j)}$  represent the pixel values of the label image and the model's prediction result at position  $(i, j)$ , respectively.  $\epsilon$  is a small constant (e.g.,  $1e-8$ ). The binary cross-entropy loss can be described as follows:

$$\mathcal{L}_{bce} = - \sum_{j=1}^W \sum_{i=1}^H M_{(i,j)} \log P_{(i,j)} + (1 - M_{(i,j)}) \log(1 - P_{(i,j)}) \quad (8)$$

Finally, the loss function can be represented as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{dice} + \mathcal{L}_{bce} \quad (9)$$

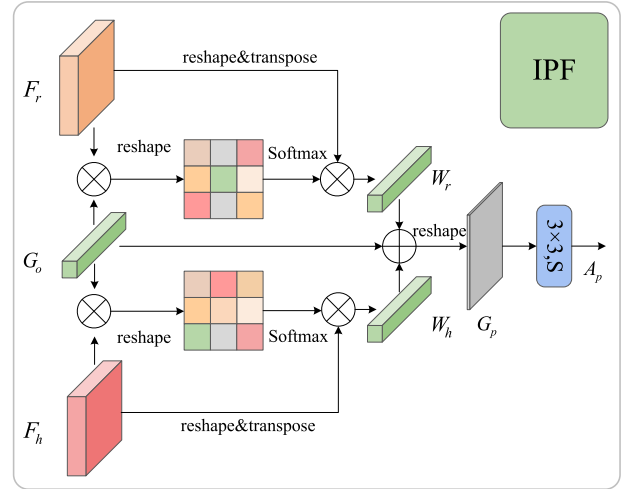
## IV. EXPERIMENTS

We conducted various experiments to assess the performance of the proposed method. Additionally, we compared this method with several other image forgery localization methods on different datasets.

### A. EXPERIMENTAL SETUP

#### 1) DATASETS

We analyzed and evaluated the DIF-Net on four publicly available datasets, including CASIA2 [1], Columbia [2], NIST16 [3], and IMD2020 [4]. For each dataset, we adopted an 8:1:1 training-validation-testing split ratio, and specific image partition details are provided in Table 1. The CASIA2 dataset is a widely used and challenging dataset, comprising two types of operations: splicing and copy-move. It consists of 5120 images with a resolution mostly at  $384 \times 256$ . In this dataset, all tampered images undergo post-processing steps

**FIGURE 5.** The structure of IPF.

such as filtering and blurring. Columbia dataset contains only spliced images with 180 uncompressed images, each having a resolution of  $757 \times 568$ . All manipulated images in this dataset have not undergone any post-processing. IMD2020 dataset consists of 2010 real-world operation images collected from the internet. NIST16 is a challenging dataset encompassing three types of operations: splicing, copy-move, and object removal. It includes 564 high-resolution images with a resolution of  $3888 \times 2592$ . Post-processing operations have also been used on the images in NIST16 to conceal visible operation traces.

#### 2) EVALUATION METRICS

In our research, we utilized pixel-level F1 score, Area Under the Curve (AUC), and Intersection over Union (IOU) as the evaluation metrics for performance comparison. By employing these comprehensive evaluation metrics, we could objectively assess the performance of our proposed method in the task of image forgery localization and compare it with other methods. Firstly, we employed the pixel-level F1 score to measure the precision and recall of the model. The F1 score is used to evaluate the model's localization accuracy at the pixel level. To assess the true performance of the model, we set the threshold to 0.5. Secondly, we used IOU to evaluate the model's localization accuracy. IOU is calculated by measuring the degree of overlap between the predicted boundaries and the true boundaries. Finally, we employed the AUC to measure the model's performance.

AUC is commonly used to evaluate the accuracy of binary classification models. In our research, we transformed the image forgery localization task into a binary classification problem. We calculated the area under the ROC curve as the AUC value. A higher AUC value (closer to 1) indicates better classification accuracy at different thresholds and, thus, better model performance. The corresponding formula is as follows:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (10)$$

$$IOU = \frac{TP}{TP + FP + FN} \quad (11)$$

$$AUC = \int_0^1 TPRd(FPR) \quad (12)$$

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \quad (13)$$

where TP and FN respectively represent the numbers of true positive and false negative classified forged pixels. TN and FP represent the numbers of true negative and false positive classified original pixels. TPR stands for True Positive Rate, which corresponds to the proportion of correctly classified forged pixels among all forged pixels. FPR stands for False Positive Rate, indicating the proportion of incorrectly classified original pixels among all original pixels.

### 3) IMPLEMENTATION DETAILS

DIF-Net is an end-to-end lightweight model implemented using PyTorch. We trained it using an NVIDIA GeForce GTX TITAN GPU. In comparison to ManTra-Net with 3.8 million parameters and RRU-Net with 4 million parameters, DIF-Net has a significantly smaller parameter count, only 1.4 million parameters. The input image size is  $256 \times 256$ . We use AdamW for optimization with an initial learning rate of 0.003. The learning rate decay strategy is Reduction on Plateau, where the reduction factor is 0.5. For DIF-Net, we trained it for 200 epochs with a batch size of 16. The model weights that achieved the highest F1 score on the validation set were utilized for testing. Data augmentation techniques were employed during training, including flipping, random rotation, Gaussian noise, and Gaussian blur.

## B. COMPARISON WITH EXISTING METHODS

### 1) COMPARATIVE METHODS

To evaluate the pixel-level forgery localization performance of our proposed DIF-Net, we compared it with several other methods on different datasets. The selected methods include both handcrafted traditional methods and deep learning-based methods.

The traditional methods consist of Error Level Analysis (ELA) [31], Noise Residual Analysis (NOI) [25], and Color Filter Array (CFA) [6]. ELA is an error level analysis method aimed at finding compression error differences between tampered and original regions by using different JPEG compression qualities. The CFA pattern estimation method approximates the camera filter array pattern using neighboring pixels and then generates tampering probabilities

for each pixel. NOI is a local noise modeling method based on noise inconsistency of high-pass wavelet coefficients. For deep learning-based methods, we chose the following four: RRU-Net [12], ManTra-Net [9], CFL-Net [34], MTSN [15], SATFL [35], and GCA-Net [17]. RRU-Net, based on the UNet [32] structure, uses a CNN with residual feedback to enhance the visibility of image attribute differences between untampered and tampered regions. ManTra-Net captures manipulation traces using a feature extractor and performs tampering localization through local anomaly detection. CFL-Net learns a mapping to a feature space using contrastive loss to differentiate untampered and tampered regions for each image. MTSN locates the tamper region by utilizing both the image edge and the mask edge. SATFL utilizes self-adversarial training and self-attention mechanisms to achieve tampering localization. GCA-Net utilizes gate mechanisms and non-local attention to identify forged regions. We retrained RRU-Net, MTSN, SATFL, and CFL-Net using their available complete codes and tested them. ManTra-Net did not provide complete code, but they offered pre-trained weights, so we used the provided pre-trained weights for testing. GCA-Net did not publicly release their code, so we performed comparisons using the available data from the paper.

### 2) COMPARISON RESULT

We evaluated our method on four datasets: CASIA2, Columbia, NIST16, and IMD2020. Table 2 presents the localization performance of various methods on these datasets. From the table, it can be observed that we achieved the best localization performance on most of the datasets. On the CASIA2 dataset, our F1 and AUC outperformed GCA-Net, which used pre-training and fine-tuning on a comprehensive dataset, by 3.3% and 2.4%, respectively. This indicates that we obtained promising localization results even without using pre-training. However, on the IMD2020 dataset, our F1 lagged behind GCA-Net by 1.1%, but the AUC remained comparable. We believe that two key reasons contribute to the occurrence of this situation. On one hand, IMD2020 is composed of images obtained from the real world, which makes it highly challenging. On the other hand, our model's training data is significantly smaller compared to GCA-Net. On the NIST16 dataset, our AUC was the same as the second-ranked RRU-Net, but our F1 and IOU improved by 1.1% and 1.6%, respectively. On the Columbia dataset, our F1, AUC, and IOU significantly outperformed the compared methods. Visualizations of the partial results of various methods on these datasets are shown in Fig. 6.

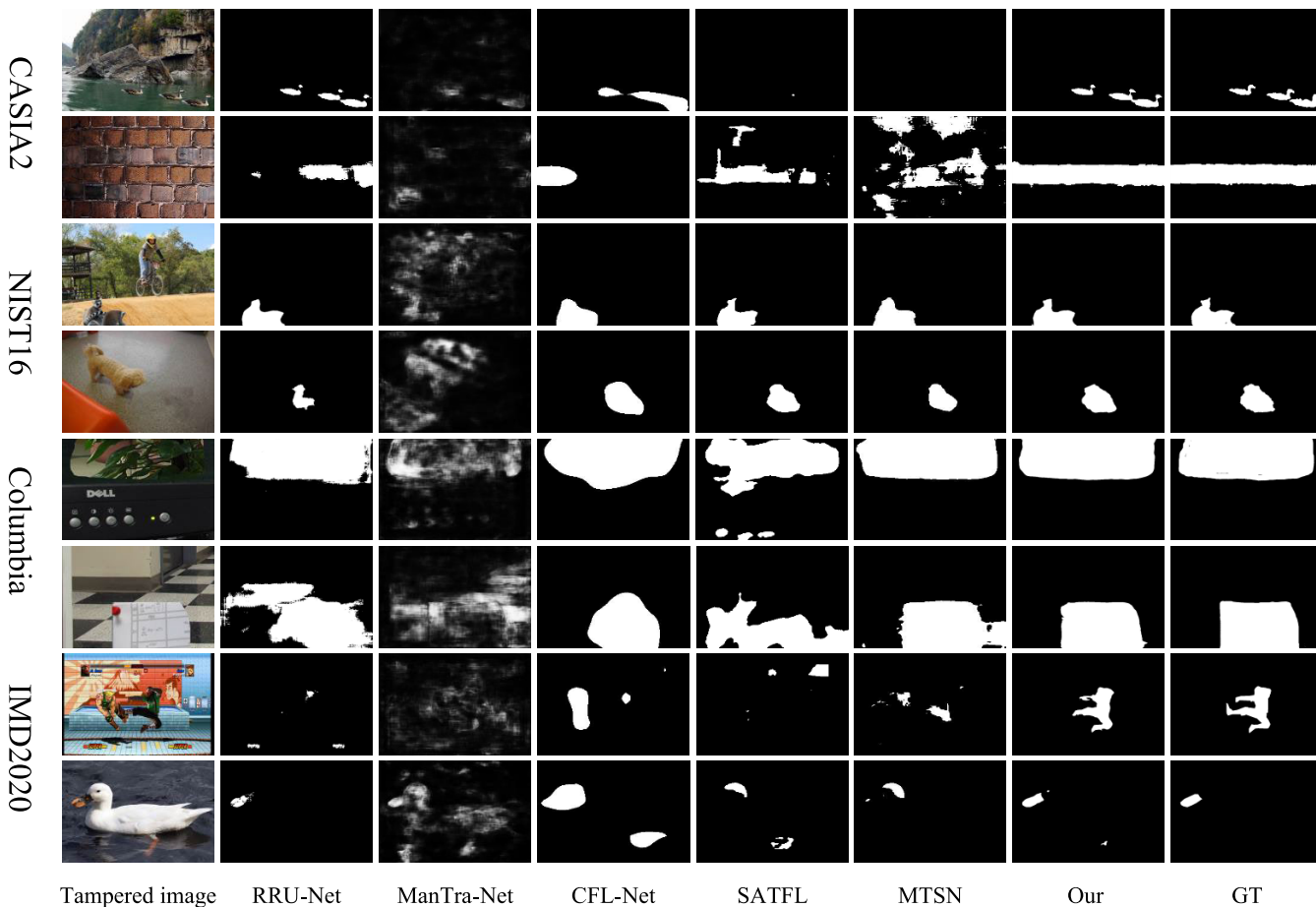
## C. ABLATION STUDY

We conducted extensive experiments to validate the impact of model parameter settings and important components on the model's final performance. All experiments were performed on the CASIA2 dataset, and the results are reported in terms of IOU, AUC, and pixel-level F1 score. Fig. 7 displays the Visualization results of some of the ablation experiments.



**TABLE 2.** The localization results of the compared methods on different datasets, the best results are highlighted in bold. - denotes the result is unavailable using the original method.

Method	CASIA2			NIST16			Columbia			IMD2020		
	F1	AUC	IOU	F1	AUC	IOU	F1	AUC	IOU	F1	AUC	IOU
ELA	21.4	61.3	-	23.6	42.9	-	47.0	58.1	-	-	-	-
NOI	26.3	61.2	-	28.5	48.7	-	57.4	54.6	-	-	58.6	-
CFA	20.7	52.2	-	17.4	50.1	-	46.7	72.0	-	-	48.7	-
RRU-Net	52.7	87.9	46.2	89.1	<b>99.7</b>	81.8	80.2	91.2	73.3	30.8	82.3	23.5
ManTra-Net	12.4	74.5	8.6	19.8	72.5	13.6	35.4	72.5	23.8	13.3	78.4	8.9
CFL-Net	40.7	87.6	47.7	74.1	99.1	69.8	84.9	97.0	74.9	33.3	85.2	24.7
SATFL	21.1	74.9	16.9	87.8	99.6	79.9	68.6	88.4	56.6	20.1	74.6	14.9
MTSN	16.7	71.6	13.3	76.7	99.2	68.0	87.3	96.2	79.8	13.1	77.4	9.2
GCA-Net	71.2	92.2	-	84.5	95.3	-	-	-	-	<b>42.6</b>	86.4	-
Our	<b>74.5</b>	<b>94.6</b>	<b>68.5</b>	<b>90.2</b>	<b>99.7</b>	<b>83.4</b>	<b>92.3</b>	<b>99.2</b>	<b>86.6</b>	41.5	<b>86.8</b>	<b>31.7</b>



**FIGURE 6.** Results of different methods on four publicly available datasets. From the first column to the eighth column, we show tampered images, RRU-Net, ManTra-Net, CFL-Net, and the prediction results of our proposed method, the GT mask of tampered images.

1) HSV COLOR SPACE

As shown in Table 3, we compared the results of RGB color space and HSV color space combinations to demonstrate the

effectiveness of the HSV color space. From the research findings, it can be observed that the HSV color space improves the model’s localization performance, F1/AUC/IOU have

**TABLE 3.** Localization results of HSV color space ablation study on CASIA2.

Method	F1	AUC	IOU
RGB + RGB	71.1	93.3	65.4
HSV + HSV	65.6	92.4	59.1
Ours	<b>74.5</b>	<b>94.6</b>	<b>68.5</b>

**TABLE 4.** Localization results of multi-scale ablation study in ACPM on CASIA2.

Method	F1	AUC	IOU
w/o Multi-scale	53.9	88.3	46.9
Multi-scale + w/o SLW	72.3	93.7	66.4
Ours	<b>74.5</b>	<b>94.6</b>	<b>68.5</b>

**TABLE 5.** Localization results of ICF and IPF ablation study in ICPM on CASIA2.

Method	F1	AUC	IOU
w/o ICF + w/o IPF	66.0	91.8	59.5
Only ICF	69.9	93.3	63.7
Only IPF	69.4	93.2	63.1
IPF + ICF	70.7	94.4	64.6
ICPM5	68.2	93.9	61.7
Ours	<b>74.5</b>	<b>94.6</b>	<b>68.5</b>

improved by 3.4%/1.3%/3.1% respectively. The reasons for this improvement can be explained as follows: On the one hand, by utilizing both color spaces, a more all-around representation of color information can be captured, aiding in distinguishing different regions. On the other hand, when images are manipulated, the HSV color space provides discriminative features that are less sensitive in the RGB color space, such as changes in hue, color saturation, and brightness anomalies.

## 2) ADAPTIVE CONVOLUTION PYRAMID MODULE

In this experiment, we aim to demonstrate the significance of ACPM. For comparison, we replaced the ACPM with two  $3 \times 3$  convolution blocks. The results are displayed in Table 4, the absence of the multi-scale component resulted in a substantial decrease in localization performance, with F1/AUC/IOU score dropping by 18.4%/5.4%/19.5%. This clearly illustrates the critical importance of multi-scale feature extraction in the encoder for image tampering detection. Compared to the multi-scale structure without self-learned weights (SLW), using self-learned weights in ACPM resulted in improvements of 2.2%/0.9%/2.1% in F1/AUC/IOU. This indicates that self-learned weights can more effectively integrate multi-scale features.

## 3) ICF AND IPF MODULES IN ICPM

We designed five experiments to assess the impact of these two modules on the final localization performance. Specifically, we also investigated the reason for using a  $3 \times 3$  convolution instead of the ICPM in the last layer. The results are as shown in Table 5. Compared to the proposed

method, the absence of ICF and IPF led to a decrease of 8.5%/2.8%/9% in F1/AUC/IOU scores, respectively. When using only ICF (IPF), a noticeable drop in localization performance was observed, with F1/AUC/IOU decreasing by 4.6%/1.3%/4.8% and 5.1%/1.6%/5.4%, respectively. We also conducted experiments to determine the order of using ICF and IPF and found that applying ICF first, followed by IPF, resulted in an F1/AUC/IOU improvement of 3.8%/0.2%/3.9%. When ICPM was used in the last layer, F1/AUC/IOU dropped by 6.3%/0.7%/6.8% respectively. The aforementioned results can be explained as follows: On the one hand, combining ICF and IPF enhances feature representation by exploiting the interdependencies between channel and spatial mappings. On the other hand, replacing ICPM with a  $3 \times 3$  convolution in the decoder was done because the initial layers of the encoder contained too much low-level information. Utilizing ICPM for fusion not only degrades the model's performance but also introduces a heavy computational burden.

## D. ROBUSTNESS ANALYSIS

To assess the robustness of our proposed image manipulation localization method, we performed pre-processing operations on the tampered images in the test dataset. These pre-processing steps included applying JPEG compression with different quality factors, Gaussian blurring with varying kernel sizes, and the addition of Gaussian noise and Salt pepper noise with a standard deviation of sigma. Subsequently, we tested the pre-processed images using models trained on the NIST16 [3] and Columbia [2] datasets. In our experiments, we selected RRU-Net [12], ManTra-Net [9], SATFL [35], MTSN [15], and CFL-Net [34] as comparative methods and compared their localization performance on manipulated images subjected to various pre-processing operations. Fig. 8 and 9 display the F1 scores of these methods under different pre-processing operations. The following observations can be made from the figures.

### 1) ROBUSTNESS RESULTS AGAINST JPEG COMPRESSION

In image forgery tasks, the tampered images are often subjected to JPEG compression to conceal the forgery traces. To evaluate the robustness of our proposed DIF-Net under different JPEG compression quality factors (QF), we varied the QF values as  $QF \in \{50, 60, 70, 80, 90, 100\}$ . Subsequently, we employed the DIF-Net model trained on the NIST16 and Columbia datasets to perform forgery localization on the test images, comparing its performance with five other methods. The experimental results on both datasets are illustrated in Fig. 8(a) and 9(a). From the figures, it can be observed that when confronted with JPEG-compressed images, only ManTra-Net exhibits a relatively noticeable performance drop on the Columbia dataset, while the performance of the other five methods remains relatively stable. Among these methods, our DIF-Net achieves the best localization results.

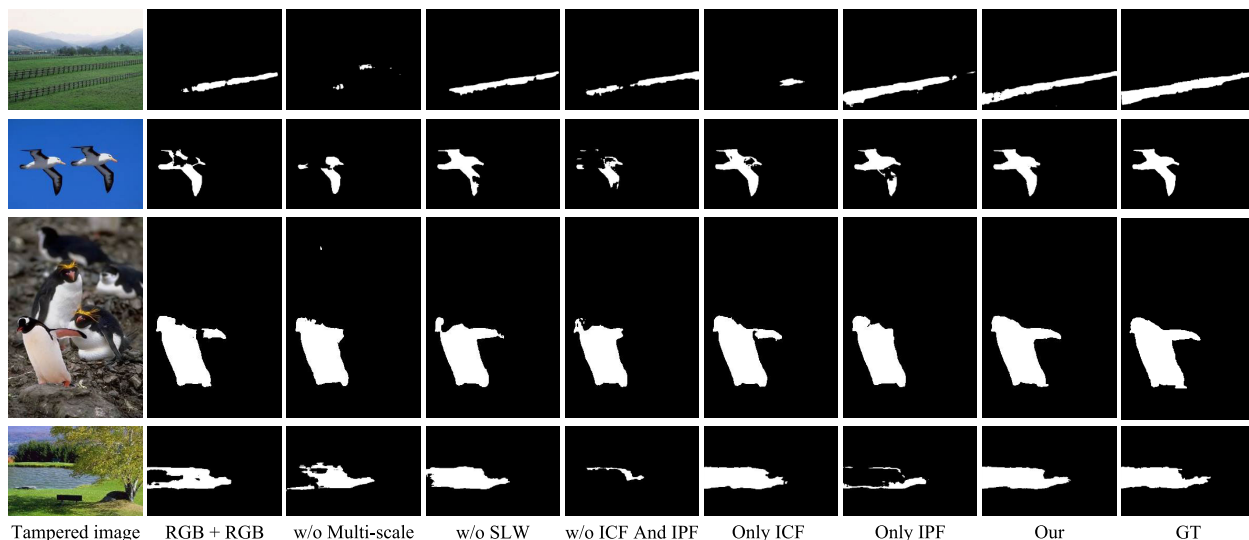


FIGURE 7. Visualization results of partial ablation experiments.

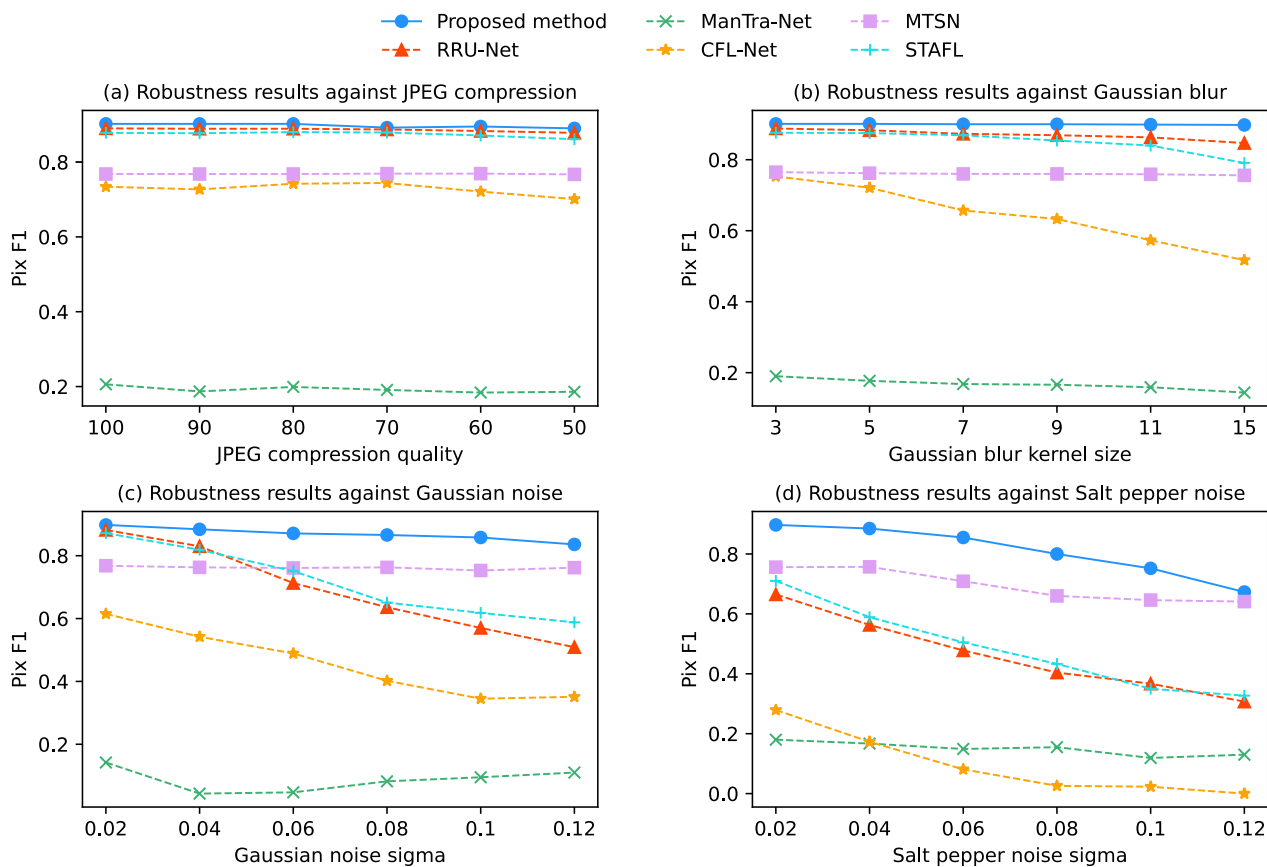
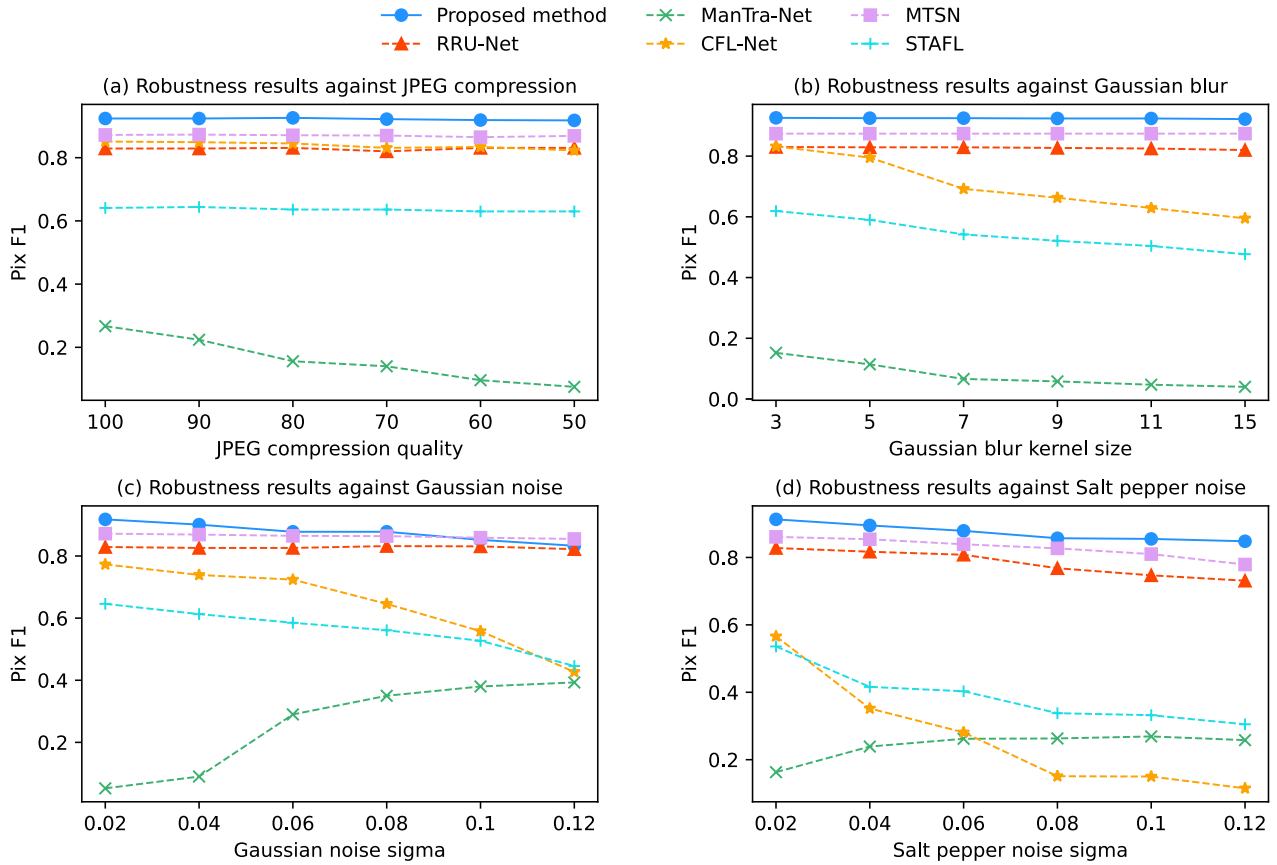


FIGURE 8. Robustness results of different methods on the NIST16 dataset. (a) JPEG compression, (b) Gaussian blur, (c) Gaussian noise, (d) Salt pepper noise.

2) ROBUSTNESS RESULTS AGAINST GAUSSIAN BLUR

Gaussian blur is a common post-processing operation that can make the edges of tampered images smoother, thereby reducing the visibility of forgery traces. In the pre-processing stage, we applied Gaussian blur with different

kernel sizes, i.e.,  $k \in \{3, 5, 7, 9, 11, 15\}$ , to simulate various levels of blurring effects. The experimental results in Fig. 8(b) and 9(b) allow us to observe the impact on the performance of the six methods. Specifically, for tampered images processed with Gaussian blur of different



**FIGURE 9. Robustness results of different methods on the Columbia dataset. (a) JPEG compression, (b) Gaussian blur, (c) Gaussian noise, (d) Salt pepper noise.**

sizes, except for CFL-Net and ManTra-Net, which were significantly affected. The localization results of the other four methods remained basically unchanged, with DIF-Net achieving the best localization performance.

### 3) ROBUSTNESS RESULTS AGAINST GAUSSIAN NOISE

Adding Gaussian noise to forged images is also a commonly used method for robustness testing. In this experiment, we compared the localization performance of the six methods under the Gaussian noise perturbation attack. The different standard deviations, set as  $\delta \in \{0.02, 0.04, 0.06, 0.08, 0.1, 0.12\}$ , to simulate varying degrees of Gaussian noise interference. Based on the experimental results presented in Fig. 8(c) and 9(c), it can be observed that as the standard deviation of Gaussian noise increases, the localization results of MTSN remain relatively stable. In contrast, ManTra-Net’s localization performance starts to improve, while the localization performance of the other four methods begins to decline. This indicates that, in the face of increasingly intense Gaussian noise interference, the localization ability of most methods is affected to some extent. For the stable performance of MTSN, we attribute this to the role played by edge guidance and multi-task loss. Edge guidance makes the tampered edges more prominent, reducing the impact of noise on the localization results. As for the unexpected results of ManTra-Net, we speculate that this may be because

ManTra-Net used up to 385 types of tampered images from various image operation types during its training process. Such a training strategy may have made ManTra-Net more adaptable to different levels of Gaussian noise because it learned a wider range of manipulation operation types. Although DIF-Net’s performance is slightly weaker than MTSN on the Columbia dataset, overall, when faced with Gaussian noise attacks, DIF-Net still achieves the best localization results.

### 4) ROBUSTNESS RESULTS AGAINST SALT PEPPER NOISE

We also compared the localization performance of the six methods under salt-and-pepper noise attacks. In the pre-processing stage, we set different standard deviations, denoted as  $\delta \in \{0.02, 0.04, 0.06, 0.08, 0.10, 0.12\}$ , to introduce varying degrees of salt-and-pepper noise interference. Salt-and-pepper noise is a common type of noise that randomly adds black and white pixels to an image, simulating image corruption. As shown in Fig. 8(d) and 9(d), we observed that under salt-and-pepper noise attacks, CFL-Net is the most sensitive, and its localization results are significantly affected by the presence of salt-and-pepper noise. This may be attributed to CFL-Net using contrastive loss during training to learn feature distinctions between tampered and untampered regions. However, for salt-and-pepper noise, which is highly random noise, accurately



modeling feature differences becomes challenging, leading to a decline in localization performance. ManTra-Net's performance on the Columbia dataset exhibits the same upward trend as it does when facing Gaussian noise attacks. The other four methods showed some degree of degradation in localization performance under salt-and-pepper noise attacks. Nevertheless, DIF-Net still achieved the best localization results under these noise conditions.

In summary, the performance of almost all models tends to degrade when different attacks are applied to test images. Particularly, the performance degradation is most pronounced when Gaussian noise and salt-and-pepper noise are added. However, DIF-Net shows more stable performance compared to other methods under these attacks, showcasing its strong robustness.

## V. CONCLUSION

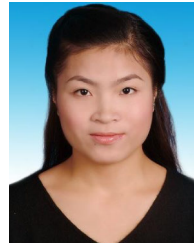
In this study, we propose a novel DIF-Net for the task of image forgery localization. Firstly, we introduce a color space combination strategy during the feature extraction phase, which leverages both RGB and HSV color spaces to capture richer feature information. This strategy allows our model to better discriminate different regions, especially when significant color variations exist between tampered and untampered areas. Secondly, we introduce the Adaptive Convolution Pyramid Module (ACPM) in the encoder of the feature extraction process, using multi-branch depth-wise convolutions with learnable weights to handle features at different scales. This enhances our model's capability to handle tampered regions of various sizes, resulting in improved localization accuracy. Finally, we propose a novel Intermediate Channel Fusion Module (ICPM) to establish coherence between channels and spatial information, enhancing feature representation. The ICPM module conducts feature fusion in the decoder, enabling more effective encoding and expression of features for the forgery localization task through the mutual dependence between channel and spatial mappings.

To evaluate our method's performance on different datasets, we conducted extensive experiments. Compared to previous methods, DIF-Net improved the F1 by 3.3%, 1.1%, and 5% on the CASIA2, NIST16, and Columbia datasets, respectively. On most datasets, the proposed method achieves the best performance in digital image tampering localization, significantly outperforming other hand-crafted traditional methods and deep learning-based approaches. Furthermore, ablation experiments prove the effectiveness of individual modules and parameters in the model's performance. While our DIF-Net exhibits remarkable robustness and maintains stable performance against various types of pre-processing attacks, it still has limitations in terms of generalization and stability. We acknowledge these issues and will strive to address them in future research, developing more adaptable and versatile approaches. Finally, we hope that DIF-Net can provide valuable assistance in scenarios that require ensuring the authenticity of digital images and the integrity of visual information, such as digital forensics, media, and journalism.

## REFERENCES

- [1] 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.
- [2] T.-T. Ng, J. Hsu, and S.-F. Chang, "Columbia image splicing detection evaluation dataset," DVMM lab. Columbia Univ. CalPhotos Digit Libr, Columbia Univ., New York, NY, USA, Tech. Rep. #203-2004-3, 2009. [Online]. Available: <https://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm>
- [3] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhan, J. Smith, and J. Fiscus, "MFC datasets: Large-scale benchmark datasets for media forensics challenge evaluation," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 63–72.
- [4] A. Novozámský, B. Mahdian, and S. Saic, "IMD2020: A large-scale annotated dataset tailored for detecting manipulated images," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Mar. 2020, pp. 71–80.
- [5] W. Wang, J. Dong, and T. Tan, "Exploring DCT coefficient quantization effects for local tampering detection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1653–1666, Oct. 2014.
- [6] P. Ferrara, T. Bianchi, A. De 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.
- [7] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.
- [8] P. Zhuang, H. Li, S. Tan, B. Li, and J. Huang, "Image tampering localization using a dense fully convolutional network," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2986–2999, 2021.
- [9] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9535–9544.
- [10] R. Salloum, Y. Ren, and C.-C. Jay Kuo, "Image splicing localization using a multi-task fully convolutional network (MFCN)," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 201–209, Feb. 2018.
- [11] X. Liu, Y. Liu, J. Chen, and X. Liu, "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7505–7517, Nov. 2022.
- [12] 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.
- [13] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3539–3553, Mar. 2023.
- [14] 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. (ECCV)*, Sep. 2020, pp. 312–328.
- [15] Y. Zhang, G. Zhu, L. Wu, S. Kwong, H. Zhang, and Y. Zhou, "Multi-task SE-network for image splicing localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4828–4840, Jul. 2022.
- [16] 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.
- [17] S. Das, Md. S. Islam, and Md. R. Amin, "GCA-Net: Utilizing gated context attention for improving image forgery localization and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 81–90.
- [18] J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, and Y.-G. Jiang, "ObjectFormer for image manipulation detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2354–2363.
- [19] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022, pp. 1140–1156.
- [20] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to  $31 \times 31$ : Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11953–11965.

- [21] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Dec. 2023.
- [22] Y. LIU, N. Liu, X. Yao, and J. Han, "Intermediate prototype mining transformer for few-shot semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022, pp. 38020–38031.
- [23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
- [24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [25] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1497–1503, Sep. 2009.
- [26] D. Cozzolino, D. Gragnaniello, and L. Verdoliva, "Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5302–5306.
- [27] X. Pan, X. Zhang, and S. Lyu, "Exposing image splicing with inconsistent local noise variances," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, Apr. 2012, pp. 1–10.
- [28] S. Lyu, X. Pan, and X. Zhang, "Exposing region splicing forgeries with blind local noise estimation," *Int. J. Comput. Vis.*, vol. 110, no. 2, pp. 202–221, Nov. 2014.
- [29] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, "Aligned and non-aligned double JPEG detection using convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 49, pp. 153–163, Nov. 2017.
- [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] N. Krawetz and H. F. Solutions, "A picture's worth," *Hacker Factor Solutions*, vol. 6, no. 2, p. 2, 2007.
- [32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.
- [33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [34] F. F. Niloy, K. Kumar Bhaumik, and S. S. Woo, "CFL-Net: Image forgery localization using contrastive learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4631–4640.
- [35] L. Zhuo, S. Tan, B. Li, and J. Huang, "Self-adversarial training incorporating forgery attention for image forgery localization," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 819–834, 2022.
- [36] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
- [37] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [38] Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, and T. S. Huang, "CCNet: Criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6896–6908, Jun. 2023.
- [39] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.
- [40] B. Xu, X. Wang, X. Zhou, J. Xi, and S. Wang, "Source camera identification from image texture features," *Neurocomputing*, vol. 207, pp. 131–140, Sep. 2016.
- [41] W. Chen, Y. Q. Shi, and G. Xuan, "Identifying computer graphics using HSV color model and statistical moments of characteristic functions," in *Proc. IEEE Multimedia Expo. Int. Conf.*, Jul. 2007, pp. 1123–1126.
- [42] V. Tuba, R. Jovanovic, and M. Tuba, "Digital image forgery detection based on shadow HSV inconsistency," in *Proc. 5th Int. Symp. Digit. Forensic Secur. (ISDFS)*, Apr. 2017, pp. 1–6.
- [43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
- [45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
- [46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [47] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Int. Workshop Deep Learn. Med. Image Anal., 7th Int. Workshop Multimodal Learn. Clin. Decis. Support*, Québec City, QC, Canada. Cham, Switzerland: Springer, Sep. 2017, pp. 240–248.



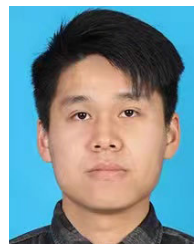
**CAIPING YAN** (Member, IEEE) received the Ph.D. degree in computer science from the University of Macau, China, in 2017. She is currently with Hangzhou Normal University, Hangzhou, China. Her current research interests include digital forensics and image processing.



**RENHAI LIU** received the B.S. degree from Chizhou University, Anhui, China, in 2021. He is currently pursuing the M.S. degree in software engineering with Hangzhou Normal University, Hangzhou, China. His research interests include digital forensics and computer vision.



**HONG LI** 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.



**JINGHUI WU** received the M.S. degree from Hangzhou Normal University, Hangzhou, China, in 2023. His research interests include digital forensics and computer vision.



**HAOJIE PAN** received the B.S. degree from Henan University of Science and Technology, Henan, China, in 2020. He is currently pursuing the M.S. degree in software engineering with Hangzhou Normal University, Hangzhou, China. His research interests include digital forensics and computer vision.

...