IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 May 2024, accepted 26 May 2024, date of publication 7 June 2024, date of current version 19 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3410974

RESEARCH ARTICLE

Tri-Path Backbone Network for Image Manipulation Localization

H. SONG[1], BAICHUAN LIN[1], AND D. YE[2]
[1]Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China
[2]Jiangsu Police Institute, Nanjing, Jiangsu 220019, China

Corresponding author: H. Song (15298398611@163.com)

This work was supported in part by the Ministry of Science and Technology of the People's Republic of China under Grant 2023YFC3303702, in part by the Institute of Forensic Science of the Ministry of Public Security under Grant 2022JB024, in part by the Ministry of Public Security of the People's Republic of China under Grant 2022JC16, in part by the Key Laboratory of Forensic Marks of the Ministry of Public Security under Grant 2023FMKFKT05, and in part by Jiangsu Education under Grant 23KJB620002.

ABSTRACT We propose a novel Tri-Path Backbone Network (TPB-Net) and train it end-to-end to effectively detect multiple types of image manipulations. The key challenge for image manipulation localization lies in the difficulty and diversity of extracting forgery features. To address this, we adopt a Triple-path Interconnected Backbone (TIB) scheme as the feature extractor, which enables the strong feature detection capabilities. Furthermore, we design and introduce the Dual-path Compressed Sensing Attention (DCSA) module, that incorporates a dual-path attention mechanism. The DCSA module intelligently compresses channels in the spatial path and spatial information in the channel path. These compression operations lead to improved learning efficiency, enhanced representation effectiveness, and increased model robustness. TPB-Net offers an end-to-end framework comprising trainable modules, facilitating joint optimization and enabling the achievement of optimal performance. Through rigorous experiments conducted on four standard image manipulation datasets, we demonstrate the superior performance of our method compared to previous state-of-the-art approaches.

INDEX TERMS Image forensics, tampering localization, triple-path backbone, dual-path sensing attention.

I. INTRODUCTION

Digit images hold a wealth of valuable information and play a pivotal role in numerous domains, encompassing media, social networks, and criminal investigations. However, the growing accessibility of tampering tools presents a significant challenge to the authenticity of these images, leading to a crisis of credibility. Accordingly, severe social consequences have arisen, including the widespread circulation of seditious rumors, telecom fraud, academic misconduct, and the use of manipulated images to fabricate forensic evidence [1]. Among the assortment of image manipulation techniques, three commonly employed methods are copy-move, splicing, and inpainting, all of which directly modify the semantic information within images [2], [3], [4], [5]. Copy-move entails duplicating a patch from one area of an image and pasting it into another region within the same image. Splicing, on the other hand, involves copying a patch from one image and pasting it onto a different image. Inpainting, also known as removal, consists of replacing a selected region in an image with pixel values predicted from the surrounding background, resulting in the disappearance of specific image content. These manipulation techniques are frequently exploited for malicious purposes, emphasizing the urgent necessity for research and development aimed at combating such practices in image manipulation.

In practical scenarios, the process of manipulating images often leaves behind discernible traces. Researchers have developed specialized algorithms to detect these traces. For example, specific algorithms have been designed to identify resampling [6], [7], median filtering [8], [9], contrast adjustment [10], [11], and double JPEG compression [12], [13]. While these approaches have proven useful, they are time-consuming and their accuracy cannot be

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

guaranteed [14]. Since each algorithm focuses on detecting a specific type of trace, a considerable number of tests must be conducted to evaluate an image. Apart from being inefficient, this complex operation is also prone to errors, as different algorithmic errors can overlap. Furthermore, the detection difficulty is exacerbated by the emergence of novel editing techniques. Therefore, there is an urgent need for the development of new techniques capable of identifying manipulated images and even pinpointing the tampered regions within them.

In recent years, deep learning technology has shown a booming trend. Due to its excellent capabilities in the fields of pattern recognition and computer vision, deep learning has been widely employed in image forensic. Various models are designed and good image manipulation localization performance is achieved [15], [16], [17], [18], [19]. Among them, the localization tasks which marked the tampered regions attract much attention [15], [16], [17]. While appearing similar, the working principles of image manipulation localization tasks (IMLTs) differ from the common image segmentation tasks. The main difference is that IMLTs focus on the region that does not originally belong to the image. So, in IMLTs, the models pay more attention to the features that exhibit discontinuities including edge inconsistencies [15], noise pattern [20], color consistency [21], EXIF consistency [22], etc. However, the common image segmentation tasks rely on the semantic information and edges. Obviously, those features that IMLTs need are more complex and diverse which make them elusive.

Efforts have been made to address the above issue. A common approach is to design specific architectures which can suppress the semantic information [15], [16], [19], [23], [24]. For example, spatial pyramid attention network (SAPN) is proposed to suppress the semantic features in the model [19]. And Mvss-net utilizes the Noise-Sensitive Branch and Edge-Supervised Branch to learn the semantic-agnostic features to the greatest extent possible [24]. Apart from the above methods, some works modify the inputs and feed the concerned features to the model. For instance, P. Zhou utilizes the SRM filter to extract noise patterns as the input to a Faster R-CNN network [23]. And M. Kwon provides the model with the Discrete Cosine Transform (DCT) features [24]. However, this type of methods extract overly simplistic features while the forgery features are characterized by their multifaceted nature [16]. In theory, depending exclusively on a single feature might limit the improvement of performance. For instance, using only noise analysis to detect images, which are manipulated by copy-move, can be challenging. Since no new element was introduced to the image and noise distribution in such images is uniform. Hence many schemes adopt the two- or three-streams structures [23], [24]. They simultaneously feed the image and the extracted features to the model. It's clear these features pre-selection methods are kind of both inconvenient and has the risk of being not sufficiently accurate. Therefore, we come with the inspiration of enhancing the ability to extract features from raw images. In conjunction with suppressing semantic features in subsequent networks, our solution is derived.

Based on the aforementioned opinion, we propose our Tri-Path Backbone Network (TPB-Net) for the image manipulation localization. We adopt a Triple-path Interconnected Backbone (TIB) scheme to provide strong feature detection capability. Then, the features pyramid obtained by this TIB is fed to the subsequent networks for fusion and selection. One critical work in the upcoming network is to effectively reduce semantic information while preserving non-semantic ones. As in the features pyramid, the high-level low-resolution features generally contain more semantic information [24]. We used a Dual-path Compressed Sensing Attention (DCSA) module in the fusion step. This DCSA adopts a dual-path attention mechanism, incorporating both spatial and channel attention. What sets it apart from Dual Attention Network [25] is that DCSA compresses the channels in the spatial path and compresses the spatial information in the channel path. These compression operations not only enhance learning and representation efficiency but also contribute to improving the robustness of the model. The final decision is made based on the above structure. Our TPB-Net offers a comprehensive end-to-end framework comprising trainable modules. It allows for joint optimization, enabling us to achieve the highest level of performance. Our contributions are as follows:

- We introduce an innovative TPB-Net designed specifically for image manipulation localization, showing a robust capability for extracting and selecting prominent features. The structure of TPB-Net is illustrated in Fig. 1
- We developed a robust backbone network that excels in feature mining, complemented by an attention mechanism designed for effective feature filtering.
- Extensive experiment results conducted on public datasets unequivocally demonstrate the remarkable superiority of TPB-Net over the existing state-of-the-art methods.

II. RELATED WORK

A. FEATURE EXTRACTOR FOR MANIPULATION DETECTION

Many studies aiming for IMLTs employ the encoder-decoder structure [14], [16], [19], [23], [24], [26], [27]. Wherein the backbone serves as a critical component, enabling the features extracting from the raw images, noise pattern and other relevant elements. Some methods rely on a single backbone [16], [19], while others employ multiple independent backbones to extract features from different sources [23]. The backbones that utilized are generally pre-trained for the ImageNet classification task. For example, in related works, pretrained backbones such as VGG16 [19], ResNet50 [24], and ResNet101 [23] have been employed to detect and analyze forgery features. Given that these backbone networks were originally designed and trained for image classification tasks, directly using them

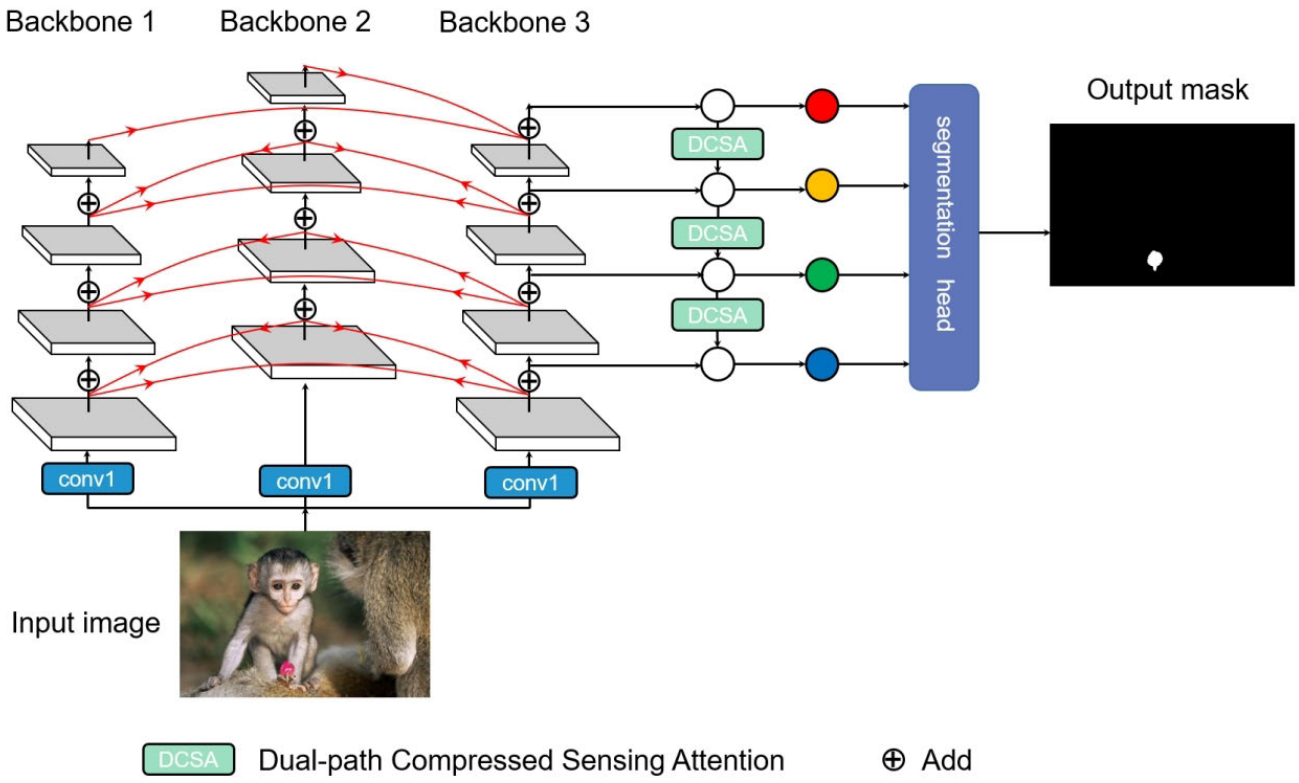


FIGURE 1. Scheme of our Tri-Path Backbone Network. We utilize a tri-path interconnected backbone to extract features from the raw image. Subsequently, we employ a feature pyramid structure in conjunction with a dual-path compressed sensing attention mechanism to refine the features for decision-making.

to extract forgery features for manipulation localization might result in suboptimal performance. Hence, we try to elevate the performance of the backbone network for IMLTs. Since design and pretrain a whole new backbone network requires considerable manpower and computational resources, we get the inspiration from Composite Backbone Network (CB-net) [28] and put forward a more cost-effective and efficient approach to construct a potent backbone for features extraction in IMLTs by integrating specifically existing backbone networks.

B. ATTENTION MECHANISM

Attention mechanism shows to be a promising approach for improving deep Convolutional Neural Networks (CNNs). Over time, research on attention mechanisms has evolved with the aim of achieving two key functions: (1) enhancing feature aggregation and (2) combining channel and spatial attention. CBAM [29] and DAnet [25] independently proposed structures that integrate both channel and spatial attention mechanisms. The main difference between these two approaches is that CBAM computes spatial attention utilizing a 2D convolution of kernel size $k \times k$, then incorporate it with channel attention in a sequential manner. On the other hand, DAnet adopts a parallel structure, where both the channel and spatial attention are calculated simultaneously and then combined through summation or another fusion operation. These approaches balance both channel and spatial information, thereby achieving excellent results. While in our view, the parallel structure of DAnet has the capability to preserve a more comprehensive set of information. Besides, considering features aggregation facilitates the extraction of more abstract and holistic features, we introduce an attention mechanism built upon feature aggregation and parallel structure for feature selection in IMLTs.

III. PROPOSED MODEL

The main concept of our approach is centered around three key steps: Initially, we construct a comprehensive feature extractor designed to capture a diverse set of features from the raw image, including but not limited to aspects such as texture, color distribution, and geometric patterns. This is crucial as it allows for a thorough analysis of various image features. Next, we introduce a module that refines these features by selectively suppressing less relevant semantic information, thereby enhancing the more critical features for our analysis. Finally, the decision-making process is based on these refined, high-quality features.

A. OVERVIEW

The schematic of our proposed model is depicted in Fig. 1. In this study, we utilize only the raw image as the input source, other than elements like noise patterns [23] and

DCT [30]. Our rationale is that since noise and DCT components are derived from the raw image, a well-designed, robust feature extractor is capable of uncovering all essential features inherent in the raw image, thereby enabling a comprehensive analysis without the need for additional data sources. Consequently, the raw image is fed into our specially designed Triple-path Interconnected Backbone (TIB). The TPB (Tri-Path Backbone) consists of three DenseNet169 networks. Within the TPB, feature maps are added between the DenseNet169 after each block to acquire rich and comprehensive features. Subsequently, a feature pyramid structure is employed to integrate features from different levels. In this integration process, it is crucial for the feature pyramid to maintain less semantic information across all scales. Acknowledging that high-level feature maps generally contain richer semantic information, we introduce our DCSA attention mechanism in the top-down pathway to facilitate effective feature selection and aggregation. In the end, the final decision is made based on the fused features.

In the following subsections, we will provide detailed explanations of the following aspects: firstly, the functioning of the TIB (Triple-Path Interconnected Backbone) as an improved feature extractor; secondly, the rationale behind adopting a feature pyramid scheme; and thirdly, the architecture of the DCSA (Dual-Path Compressed Sensing Attention).

B. TIB AS A POWERFUL FEATURE EXTRACTOR

TIB consists of 3 identical DenseNet169. The selection of DenseNet169 is based on extensively comparative experiments. As illustrated in Fig. 1, the backbone B_1, B_2 and B_3 constitute the TIB. Each backbone consists of four blocks, with each block containing several convolutional layers that produce feature maps of the same size. The L_{th} block of the backbone performs a transformation denoted as F^L. In a single backbone, the L_{th} block takes the output (denoted as x^{th}) of the previous $((L-1)_{th})$ block as its input. This can be expressed as follows:

$$x^L = F_k^L(x^{L-1}) \tag{1}$$

In contrast to using separate backbone networks, we employ three to obtain more diverse and comprehensive features. The output features from the previous (L_{th}) block of the three backbones (B_1, B_2, B_3) are summed and passed to L_{th} block of each backbone. This process can be formulated as follows:

$$x^L = \sum_{k=1}^{3} F_k^L(x_1^{L-1} + x_2^{L-1} + x_3^{L-1}) \tag{2}$$

As a result, each individual backbone shares and collectively contributes to the feature maps between blocks. The obtained feature maps from different levels are fed into the following network. Given that this composition style utilizes three individual backbones and facilitates internal interconnections among them, we refer to it as the Triple-path interconnected backbone. Apart from the TIB, there are various other combination structures of backbones. Comparative experiments are conducted to demonstrate the superiority of TIB. Several structures used for comparison are introduced in the following.

Fig. 2 illustrates several alternative structures. Fig. 2(a) represents our method TIB, while Fig. 2(b) shows a Series Connection (SC) within the same level block of backbones. In SC, B_1 provides features to B_2, and B_2 supports B_3 in generating the feature maps. The operation of SC can be formulated as follows:

$$x^L = F_3^L(x_1^{L-1} + x_2^{L-1} + x_3^{L-1}) + F_2^l(x_1^{L-1} + x_2^{L-1}) + F_l^L(x_1^{L-1}) \tag{3}$$

In SC, B_1 and B_2 perform the auxiliary roles. The feature maps for the subsequent task are exclusively computed by B_3. Fig. 2(c) demonstrates another transform based on SC. In the scheme depicted in Fig. 2(c), the output of the high-level block of the previous backbone is fed into the succeeding backbone. We refer to this operation as Descending-level Series Connection (DSC) and it can be represented by the following equation:

$$x^L = F_3^L(x_3^{L-1} + x_2^L + x_1^{L+1}) + x_2^{L+1} + x_2^{L+2} \tag{4}$$

Furthermore, another structure which feeds the output of the low-level block of the previous backbone to the succeeding backbone. We name this operation as Ascending-level Series Connection (DSC) and it can be mathematically expressed as follows:

$$x^L = F_3^l(x_3^{L-1} + x_2^{L-2} + x_1^{L-3}) + x_2^{L-1} + x_2^{L-2} \tag{5}$$

Among the SC, DSC and ASC methods, there is a hierarchical relationship between the backbones, with the auxiliary backbones providing supplementary features to the primary backbone network. It is apparent that our method stands out from others in one significant aspect: each backbone works together without any primary or secondary differentiation. The comparison results will be presented in Section IV.

C. DCSA FOR FEATURE SELECTION

With the extraction of features at different stages, a feature pyramid is adopted, as illustrated in Fig. 1. This feature pyramid enhances the model's capability to capture and represent information across various scales, leading to improved performance. However, while leveraging the advantages of it, we also need to address a significant drawback associated with it, namely the presence of semantically strong information it brings along. Since the original Feature Pyramid Networks utilize the top-down pathway and lateral connections and the high-level feature maps are semantically rich [31], there will be a large amount of semantic information flowing into the lower levels. However, semantic information is irrelevant to our task and provides no benefit. To tackle this issue, we incorporated our own

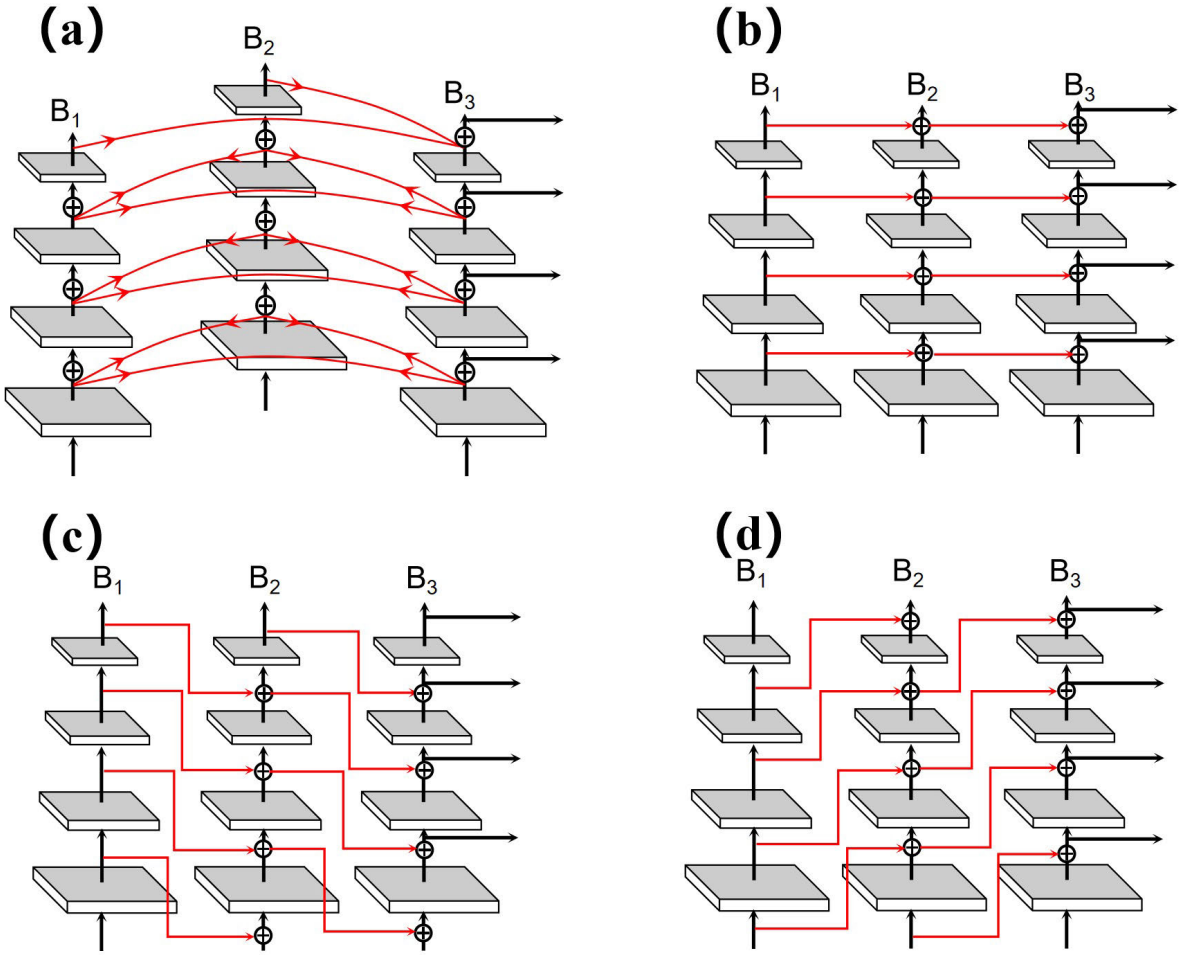


FIGURE 2. Four kinds of combination methods (a) Triple-Path Interconnected Backbone (TIB). (b) Series Connection (SC). (c) Descending-level Series Connection (DSC). (d) Ascending-level Series Connection (ASC).

designed DCSA, specifically tailored to filter out irrelevant information, including semantics.

The structure of DCSA is illustrated in Fig. 3. It consists of two branches similar to DAnet [25]. To minimize information loss as much as possible, we employed a dual-module attention mechanism to model both channel and spatial attention. The channel module adopts the structure of Efficient Channel Attention (ECA) [32]. The features are first aggregated by average pooling. Then, channel weights are generated by applying a 1D convolution with a size of 5 and a sigmoid activation function. The computation process can be described by the following equation:

$$X_c = o(\mathrm{Con}\, v_{5\times1}(\mathrm{MaxPool}(X_{input}))) \times X_{input} \tag{6}$$

where X_{input} is the input and X_C is the output of the channel module. In this module, dimensionality reduction and cross-channel interaction are applied. The space module is demonstrated in the lower part of Fig. 3. The features are first condensed using a 3×3 2D convolution, and spatial weights are generated by applying a 3×3 2D convolution with a softmax activation function. Similarly, the spatial branch also incorporates channel dimensionality reduction with cross-space interaction. The computation process of the space module can be described by:

$$X_s = \mathrm{SoftMax}(\mathrm{Conv}_{3\times3}(\mathrm{Conv}_{3\times3}(X_{input}))) \times X_{input} \tag{7}$$

where X_s is the output of the space module. Hence the output X_{output} of the DCSA equals:

$$X_{out} = X_c + X_s \tag{8}$$

Dimensionality reduction and cross interaction techniques are employed in both modules, resulting in a substantial reduction in model complexity while maintaining performance.

IV. EXPERIMENTS

In this section, we present experiments conducted on four distinct image manipulation datasets to evaluate the effectiveness of our TPB-Net. The obtained results are compared against other state-of-the-art (SOTA) methods.

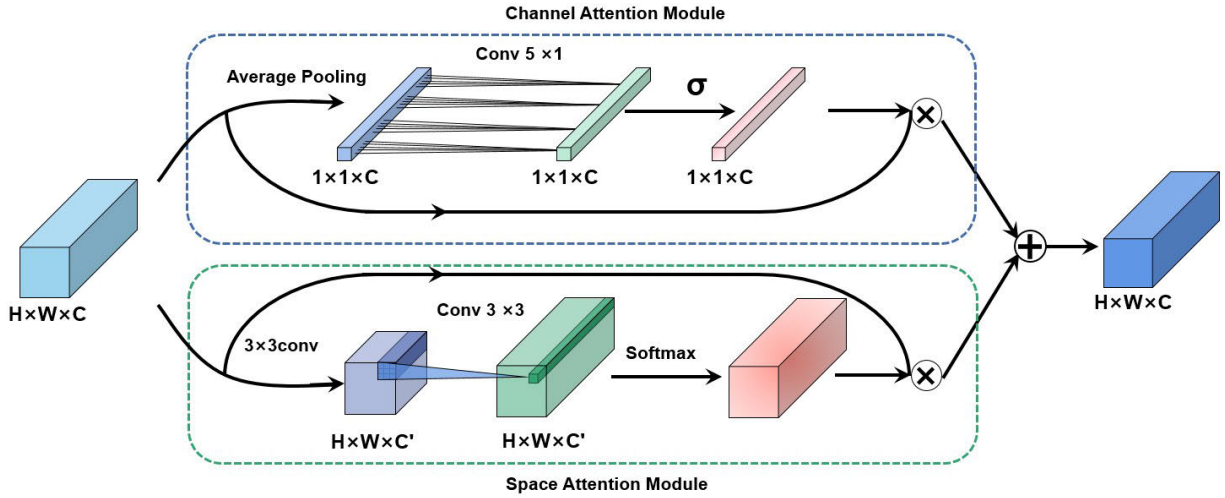


FIGURE 3. The diagram of the Dual-Path Compressed Sensing Attention (DCSA). DCSA adopts a parallel structure, with the upper branch is channel attention and the lower branch is spatial attention. The innovation of DA lies in the adoption of compression cross interaction in both branches.

A. EXPERIMENTAL SETUP

1) DATASETS

Building upon the methodology presented in [2], we utilize the synthetic dataset suggested in [2] for pretraining. And we employ four well-known benchmarks, namely CASIA [33], COVERAGE [34], Nist Nimble 2016 (NIST16),[1] and Columbia [35] datasets, to evaluate our approach against state-of-the-art (SOTA) methods in manipulation detection.

- **CASIA** dataset comprises two versions, CASIAv1 and CASIAv2, which offer a diverse range of spliced and copy-moved images. The tampered regions within these images are meticulously chosen, and additional post-processing techniques such as filtering and blurring are applied. The dataset includes binary ground-truth masks for the tampered regions. Consistent with prior studies [19], [23], [24], we adopt the common practice of utilizing CASIA v2 for training purposes and CASIAv1 for testing in our experiments.
- **Coverage** is a small collection comprising only 100 images that have been generated using copy-move method. Special care has been taken to meticulously eliminate any visible traces of manipulation within the images. Furthermore, binary ground-truth masks are provided alongside the dataset.
- **NIST16** encompasses all three tampering techniques. The manipulations within this dataset have undergone post-processing to hide the traces, thus posing a significant challenge. Ground-truth tampering masks are provided with the dataset to facilitate evaluation of the detection methods.
- **Columbia** is a Splicing based dataset which focuses on splicing based on uncompressed images. Ground-truth masks are provided.

[1] https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation/

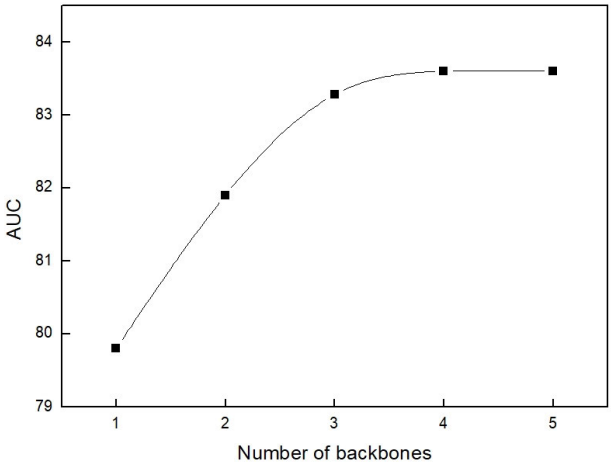


FIGURE 4. Comparison results by using different number of backbones in TPB-Net.

TABLE 1. Training and testing splits, along with the corresponding number of images, for the four standard datasets.

Datasets	CASIA	COVERAGE	NIST16	Columbia
Training	5123	75	404	0
Testing	920	25	160	180

To ensure a fair and comprehensive comparison with the current state-of-the-art (SOTA) methods, we follow the widely adopted training-testing splitting configurations [19], [23] on CASIA, NIST16, and Coverage datasets. Regarding the Columbia dataset, we utilize the entire dataset for validation purposes after training our models on the synthetic dataset. For more detailed information about training settings, please refer to Table 1.

2) EVALUATION METRIC

In our task, we are focused on binary segmentation, where each pixel in the input image is assigned a label as either

TABLE 2. Comparison of our method with four SOTA methods on CASIA, NIST16, Columbia, and Coverage. We evaluated the AUC and F1 (%) metrics to assess the effectiveness of our approach.

Datasets	CASIA		Coverage		NIST16		Columbia	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
ManTra-Net	75.2	40.6	77.5	40.9	91.2	60.7	73.5	52.1
SPAN	77.2	42.4	76.6	41.2	91.1	58.8	81.1	48.8
CAT-Net	72.4	31.2	70.2	32.6	87.5	42.8	73.3	49.7
MVSS-Net	76.8	42.8	78.9	42.1	89.1	61.2	76.4	53.4
Ours	**83.5**	**65.0**	**89.6**	**65.1**	**95.1**	**91.1**	**89.4**	**70.2**

tampered (white) or authentic (black). This results in the generation of 2-dimensional binary arrays that have the same dimensions as the input image. To evaluate the performance and make comparisons, we utilize two evaluation metrics: pixel-level F1 score and Area Under the receiver operating Characteristic Curve (AUC). It is worth noting that while many previous studies optimize the decision threshold for F1 score based on the test set, we adopt a different approach. Given that determining the optimal threshold for tampered images in real-life scenarios is challenging, we employ a fixed threshold of 0.5. This choice allows for a more objective reflection of performance across different methods and avoids potential bias introduced by threshold optimization.

3) IMPLEMENTATION

The proposed network is trained end-to-end, with the input images resized to 512 × 512 pixels. The three DeseNet169 networks that constitute TIB are ImageNet pretrained. For the remaining normal convolutional layers, the kernel weights are initialized using He initialization [36], and the biases are initialized to zero. We employ the Adam optimizer with a fixed learning rate of 5×10^{-5} [37]. Throughout the training process, we monitor the validation loss at each epoch. If the validation loss does not decrease for 10 consecutive epochs, we halve the learning rate until it reaches a minimum value of $\times 10^{-7}$. The model is trained for 100 epochs using a batch size of 12. To address the data imbalance commonly observed in forgery datasets [38], we utilize dice loss as the loss function [39]. This loss function effectively handles the imbalanced nature of the data. Our model is implemented in PyTorch and trained on 2 NVIDIA RTX3090 GPUs. To ensure a fair and impartial performance evaluation of our model, we deliberately refrain from employing any data augmentation techniques during the training process. This decision ensures that our model's performance is assessed solely based on its inherent capabilities and avoids any potential bias introduced by data augmentation.

B. RESULTS

1) COMPARISON WITH THE STATE-OF-THE-ART

In order to evaluate the effectiveness of our approach, we conduct a comparative analysis of its performance against baseline models including namely ManTra-Net [16], SPAN [19], CAT-Net [30], and MVSS-Net [24]. The models

TABLE 3. Comprehensive Comparison of Various Backbone Structures Utilizing TPB-Net and DenseNet169 as Reference Architectures. Initially, each model is pretrained on synthetic dataset, followed by fine-tuning on the CASIA dataset, following the procedure described earlier. The backbone structures under evaluation include: 'SD' (Single DenseNet169 Connection), 'SC' (Series Connection), 'DSC' (Descending-level Series Connection), and 'ASC' (Ascending-level Series Connection). Key metrics for comparison encompass Area Under the Curve (AUC), computational load (measured in Giga Multiply-Accumulate Operations, GMAC), the number of parameters (Num parameters), and computational time (Time).

Structures	AUC	GMAC	Num parameters(Million)	Time(s)
SD	79.8	27.82	14.32	1.446
ASC	76.2	65.23	15.21	1.541
DSC	81.1	65.17	15.20	1.533
SC	79.5	63.81	15.10	1.491
TIB	83.5	63.78	15.11	1.484

undergo a two-step training process, starting with pre-training on the synthetic dataset, followed by fine-tuning on CASIA, COVERAGE, and NIST16, with the exception of COLUMBIA, which was reserved exclusively for validation purpose. The pixel-level localization performance of the models is presented in Table 2. Our method demonstrates obvious superiority over the other models and achieves remarkable scores. The AUC scores of our work on all four datasets demonstrate a notable improvement of approximately 4-10, all exceeding 83%. Particularly impressive are the results for Coverage, NIST16, and Columbia, where the AUC scores exceed 89%. Our TPB-Net exhibits strong feature extraction capabilities by leveraging the TIB, enabling it to search for more diverse and richer features compared to single backbone architecture. As a result, we achieve enhanced localization accuracy. However, the AUC score for CASIA, at 83.5%, is relatively lower compared to the other datasets. This discrepancy may arise from the cross-dataset training and validation between CASIAv1 and CASIAv2. The F1 score shows a remarkable improvement of nearly 20, surpassing the AUC increase. This enhancement can be attributed to our method's ability to generate prediction masks with pixel values predominantly clustered around 0 or 255. Consequently, our method produces less ambiguous decisions, as illustrated in Fig. 6. Unlike other approaches [24], our TPB-Net does not heavily rely on an optimal threshold to achieve a high F1 score. We attribute this to the effectiveness of the DCSA module incorporated within our model, which helps in excluding interference information and providing precise features that aid in making accurate decisions.

2) COMPARISON WITH ALTERNATIVE BACKBONE STRUCTURES

We conducted experiments to compare our proposed TIB architecture with alternative structures, namely SC, DSC, and ASC, as depicted in Fig. 2. For comparative analysis, a single DenseNet169 structure was also evaluated. These experiments utilized the TPB-Net architecture, with variations only in the backbone component. The pretraining on a synthetic dataset and subsequent fine-tuning on the CASIA

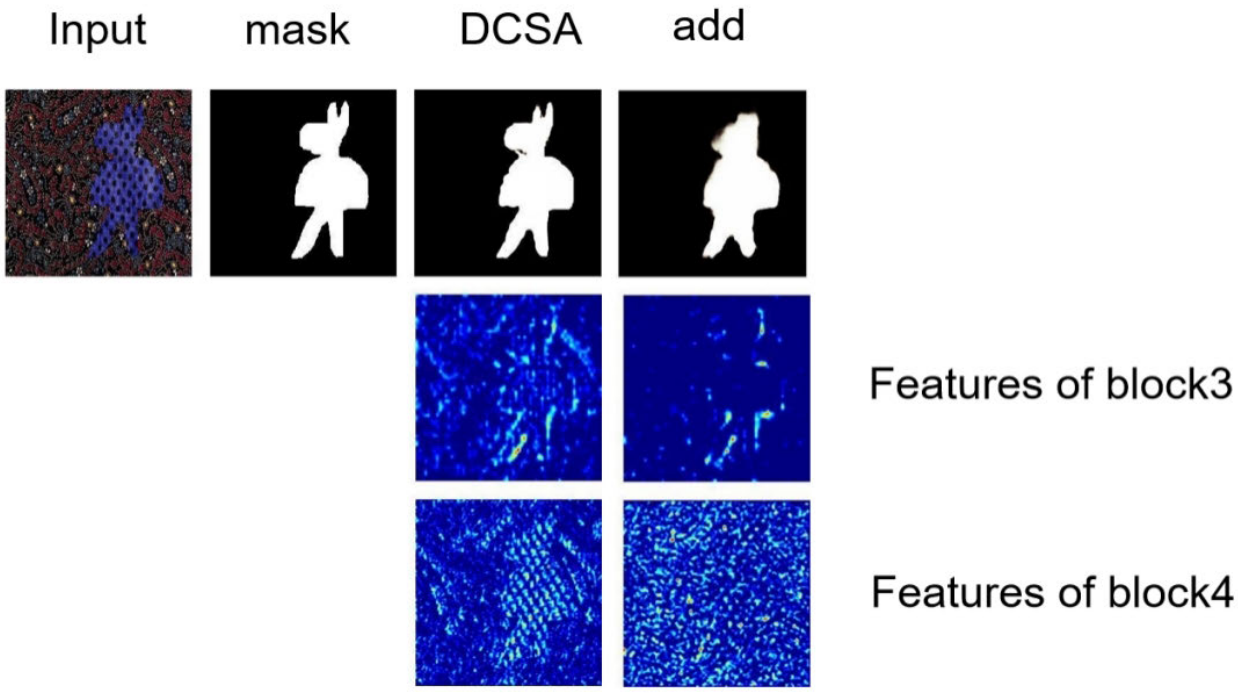


FIGURE 5. Visualization comparison of the features extracted by backbone when DSCA or add operation is employed. we visualize the feature maps of 3rd and 4th block of the backbone. Notably, the feature maps exhibit more representational qualities when DSCA is employed, as evidenced by their stronger activation values and clearer boundaries.

dataset were conducted following our previously described methodology, with results presented in Table 3.

Regarding the Area Under the Curve (AUC) performance, the SD configuration achieved 79.8%, while ASC recorded the lowest at 76.2%. Our analysis suggests that lower-level features, being closer to raw data, capture finer details, whereas deeper features are more abstract. Consequently, integrating lower-level features from a preceding backbone directly into the higher-level of a subsequent backbone can adversely affect the abstraction capacity of deep features. DSC, on the other hand, scored an AUC of 81.1%, indicating an improvement over ASC. In DSC, the deeper features from one backbone are merged with the shallow features of the next, enhancing the abstract information within deep features. However, there remains a noticeable performance gap between SC and our TIB approach. We infer that the unidirectional feature transmission in SC does not effectively foster inter-backbone collaboration. Our TIB scheme, with an AUC of 83.5%, demonstrates superior performance, which we attribute to its interconnected structure that efficiently reduces parameter redundancy and fosters collaborative functionality among backbones.

In terms of computational load, as measured by GMAC, the SD configuration, with a value of 27.82, demonstrates significantly lower complexity compared to the other structures, all of which incorporate three backbones. This lower GMAC value for SD reflects its simpler, less computationally intensive nature. In contrast, the GMAC values for ASC, DSC, SC, and TIB are relatively similar, owing to their utilization of three backbones. However, ASC and DSC exhibit slightly higher values than SC and TIB. This is attributed to their more complex interconnections between backbones.

When examining the number of parameters, it's noted that all structures (SD, ASC, DSC, SC, TIB) maintain a comparable level, with only marginal differences. This similarity in parameter count indicates a balanced design approach where the increase in model complexity is counterbalanced by efficient architectural choices. The SD configuration, with its single backbone, naturally presents a lower parameter count, aligning with its lower computational load.

Regarding computational time, the efficiency of each structure was monitored. All multi-backbone models, despite their complexity, achieved computational times that were in close range with each other, suggesting optimized processing paths in their design. The SD model, with its streamlined architecture, demonstrated a slightly quicker computational time, reflecting its less complex characteristic. These observations highlight the trade-offs between model complexity and efficiency, suggesting that while more complex models like ASC, DSC, and TIB offer advanced feature processing, they do so with an acceptable increase in computational demands.

3) NUMBER OF BACKBONES IN INTERCONNECTED BACKBONES

We conducted experiments to explore the relationship between the performance and the number of backbones in

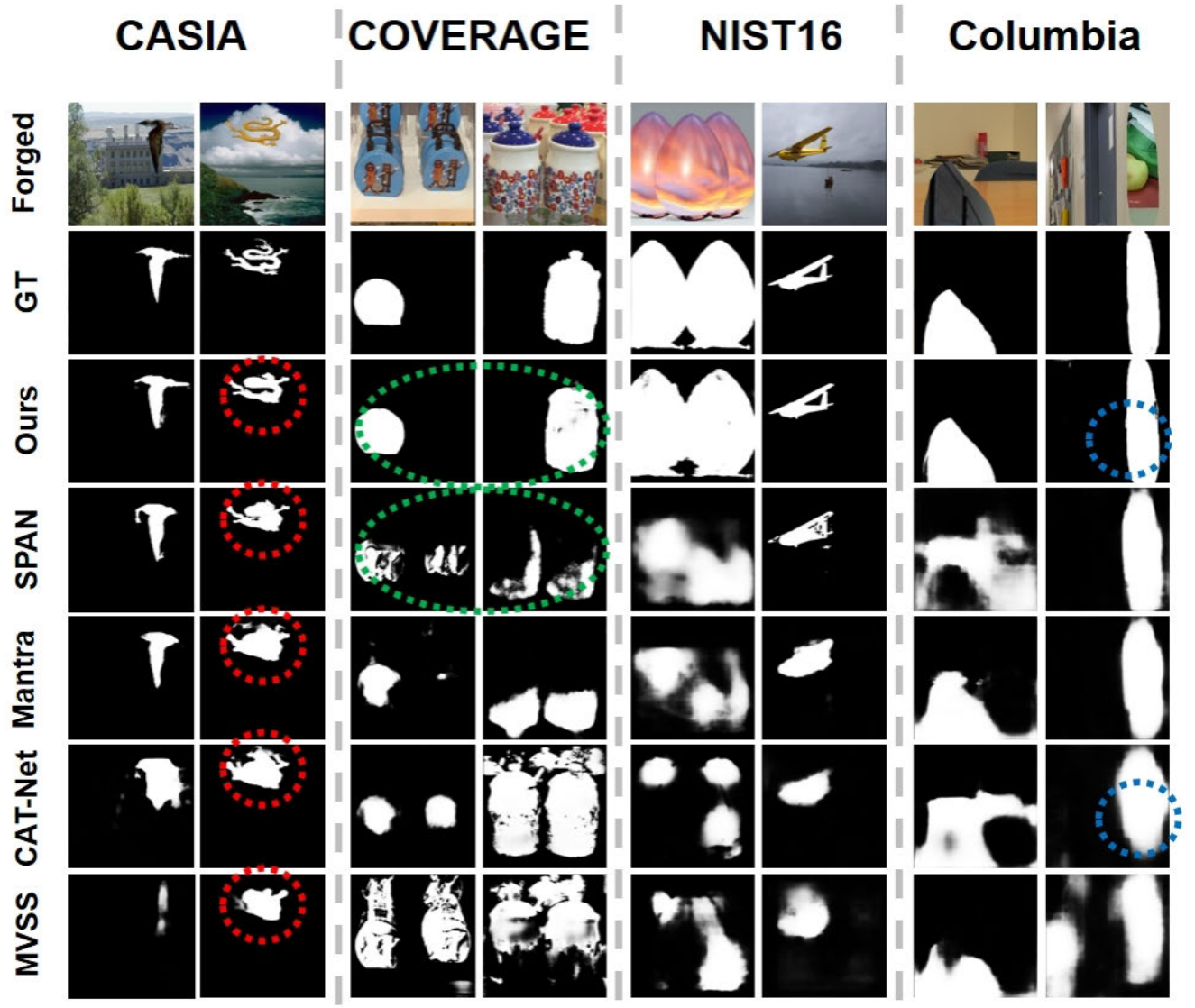


FIGURE 6. Comparison of prediction results on CASIA, Coverage, NIST16, and Columbia datasets. From top to bottom: Manipulated Image, Ground-truth mask, SFEN prediction, SPAN prediction, Mantra-Net prediction, CAT-Net prediction, MVSS prediction. The figure illustrates the visual comparison of the prediction results obtained from the different approaches on the mentioned datasets.

the TIB structure. The dataset used and the training process remainsconsistent with the previous experiments, and the results are presented in Fig. 4.

We observed that the AUC score exhibits a steady increase as the number of backbones grows, eventually reaching a plateau when using three backbones. Based on these findings, we adopted to utilize the Triple Backbone architecture as it demonstrates the optimal performance in terms of AUC score.

4) THE INFLUENCE OF DCSA

To investigate to what extent DCSA improves the final performance, we conducted the ablation experiments about with or without DCSA in our scheme. Element-wise add is used for comparison. When using add, the AUC score dropped by 2.4 compared to using DCSA.

For a more comprehensive understanding, we visually analyze and compare the feature maps extracted by the backbone when employing DCSA and the add operation. Fig. 5 demonstrates cases with input images that contain a single tampered region. It is evident from the visualization that the feature maps exhibit stronger activation values within the tampered region when DCSA is utilized, in contrast to when the add operation is employed. This visualization example highlights the beneficial role of DCSA in providing more representative features for the given task.

5) ROBUSTNESS EVALUATION

In order to assess the robustness of TPB-Net, we conducted a series of experiments. We tested various manipulation methods as listed in Table 4. These methods included resizing,

TABLE 4. Robustness analysis of TPB-Net on NIST16. ↓ indicates the decrease in AUC compared to the case when no manipulation is applied.

Manipulation method	SPAN	PSCCNet	Ours
None	83.95	85.47	99.50
Resize (0.78×)	↓0.71	↓0.18	↓0.31
Resize (0.25×)	↓3.63	↓0.46	↓1.62
Gaussian Blur (kernel size=3)	↓0.85	↓0.09	↓0.20
Gaussian Blur (kernel size=15)	↓4.80	↓5.54	↓3.77
Gaussian Noise (sigma=3)	↓8.78	↓7.05	↓6.33
Gaussian Noise (sigma=15)	↓16.67	↓8.82	↓9.21
JPEG Compress (quality=100)	↓0.36	↓0.07	↓0.09
JPEG Compress (quality=50)	↓3.27	↓0.1	↓1.26

Gaussian blur, Gaussian noise, and JPEG compression, which are applied to the NIST16 dataset to generate samples. The implementation of these manipulation techniques was carried out using OpenCV, a Python-based library for computer vision.

We compared the performance of our model with the SPAN [19] and PSCCNet [40] methods on the manipulated dataset. The data used for these two models were obtained from their original papers. Our experimental results demonstrate that SFEN exhibits strong robustness across the tested manipulation methods. In particular, our method outperformed SPAN and PSCCNet in handling Gaussian blur and noise. However, it was observed that SFEN is more sensitive to resizing and JPEG compression.

6) QUALITATIVE RESULT

In Fig. 6, we present predicted results from different methods. These results highlight three key advantages of our method over others: 1)Stronger ability to accurately segment complex tampered regions (indicated by red circles); 2)Capability to differentiate forged regions within copy-move scenarios, effectively avoiding interference from the original area (indicated by green circles); 3)Provision of more precise and clearer boundaries around tampered regions (indicated by blue circles).

These advantages demonstrate the superior performance and effectiveness of our method in tackling various types of tampering scenarios.

V. CONCLUSION

In summary, we propose the Tri-Path Backbone Network (TPB-Net) as an effective solution for image manipulation localization. TPB-Net utilizes the Triple-path Interconnected Backbone (TIB) scheme, enhancing feature detection capabilities. Furthermore, we leverage a features pyramid obtained from TIB, which is subsequently utilized for fusion and selection in subsequent networks. A key aspect of our network is the effective reduction of semantic information while preserving non-semantic information. This is accomplished through the incorporation of the Dual-path Compressed Sensing Attention (DCSA) module, which integrates a dual-path attention mechanism. The DCSA module compresses channels in the spatial path and compresses spatial information in the channel path. These compression operations enhance learning efficiency, improve representation effectiveness, and enhance overall model robustness. The final decision is made based on this proposed structure. TPB-Net provides an end-to-end framework comprising trainable modules, allowing for joint optimization and achieving optimal performance. The method demonstrates both accuracy and robustness in general manipulation localization tasks.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

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] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4980–4989.

[3] 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.

[4] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. Dugelay, and M. Pic, "DEFACTO: Image and face manipulation dataset," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[5] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[6] 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.

[7] M. Kirchner, "Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue," in *Proc. 10th ACM Workshop Multimedia Secur.*, A. D. Ker, J. Dittmann, and J. J. Fridrich, Eds. Oxford, U.K., Sep. 2008, pp. 11–20.

[8] M. Kirchner and J. J. Fridrich, "On detection of median filtering in digital images," *Proc. SPIE*, vol. 7541, Jan. 2010, Art. no. 754110.

[9] X. Kang, M. C. Stamm, A. Peng, and K. J. R. Liu, "Robust median filtering forensics using an autoregressive model," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 9, pp. 1456–1468, Sep. 2013.

[10] M. C. Stamm and K. J. R. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 492–506, Sep. 2010.

[11] H. Yao, S. Wang, and X. Zhang, "Detect piecewise linear contrast enhancement and estimate parameters using spectral analysis of image histogram," in *Proc. IET Int. Commun. Conf. Wireless Mobile Comput. (CCWMC)*. Shanghai, China: IET, 2009, pp. 94–97.

[12] T. Bianchi and A. Piva, "Detection of non-aligned double JPEG compression with estimation of primary compression parameters," in *Proc. 18th IEEE Int. Conf. Image Process.*, B. Macq and P. Schelkens, Eds., Brussels, Belgium, Sep. 2011, pp. 1929–1932.

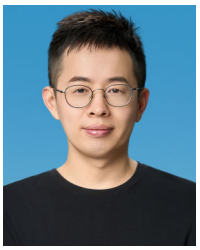
[13] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.

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

[15] 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.

[16] 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)*, Long Beach, CA, USA, Jun. 2019, pp. 9535–9544.

[17] P. Zhou, B.-C. Chen, X. Han, M. Najibi, A. Shrivastava, S.-N. Lim, and L. Davis, "Generate, segment, and refine: Towards generic manipulation segmentation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 13058–13065.
[18] C. Yang, H. Li, F. Lin, B. Jiang, and H. Zhao, "Constrained R-CNN: A general image manipulation detection model," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, London, U.K., Jul. 2020, pp. 1–6.
[19] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Proc. 16th Eur. Conf. Comput. Vis.*, vol. 12366, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Glasgow, U.K.: Springer, 2020, pp. 312–328.
[20] 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.
[21] Y. Fan, P. Carré, and C. Fernandez-Maloigne, "Image splicing detection with local illumination estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 2940–2944.
[22] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proc. 15th Eur. Conf. Comput. Vis.*, vol. 11215, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Munich, Germa: Springer, 2018, pp. 106–124.
[23] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1053–1061.
[24] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "MVSS-net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3539–3553, Mar. 2023.
[25] 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.*, Jun. 2019, pp. 3146–3154.
[26] 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.
[27] H. Zhu, G. Cao, and M. Zhao, "Effective image tampering localization with multi-scale ConvNeXt feature fusion," 2022, *arXiv:2208.13739*.
[28] Y. Liu, Y. Wang, S. Wang, T. Liang, Q. Zhao, Z. Tang, and H. Ling, "CBNet: A novel composite backbone network architecture for object detection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 11653–11660.
[29] 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.
[30] 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.
[31] 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.
[32] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[33] 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.
[34] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "COVERAGE—A novel database for copy-move forgery detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 161–165.
[35] T.-T. Ng, J. Hsu, and S.-F. Chang, "Columbia image splicing detection evaluation dataset," DVMM lab. Columbia Univ. CalPhotos Digit. Lib., Tech. Rep., 2009.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[38] F. Z. El Biach, I. Iala, H. Laanaya, and K. Minaoui, "Encoder-decoder based convolutional neural networks for image forgery detection," *Multimedia Tools Appl.*, vol. 81, no. 16, pp. 22611–22628, Jul. 2022.
[39] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.
[40] 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.



H. SONG received the B.S. and Ph.D. degrees in aerospace engineering in optics engineering from Nanjing University of Science and Technology, Nanjing, China, in 2013 and 2019, respectively.

From 2019 to 2021, he was a Research Assistant with China Academy of Engineering Physics. Since 2021, he has been a Research Assistant with the Institute of Forensic Science, Ministry of Public Security. He is the author more than 20 articles and more than ten inventions. His research interests include computer vision, deep learning, and pattern recognition.

BAICHUAN LIN received the bachelor's and master's degrees in optical engineering from Tianjin University, in 2018 and 2021, respectively. He is currently with the Institute of Forensic Science, Ministry of Public Security. His current research interests include pattern recognition, machine learning, and image processing.

D. YE received the B.S. and Ph.D. degrees in optics engineering from Nanjing University of Science and Technology, Nanjing, China, in 2013 and 2018, respectively. He is currently an Assistant Professor with the Department of Forensic Science and Technology, Jiangsu Police Institute, China. His current research interests include image processing and machine learning.