electronics

MDPI

Article

A Multi-Path Inpainting Forensics Network Based on Frequency Attention and Boundary Guidance

Hongquan Wang [1], Xinshan Zhu [1,*], Hao Sun [1], Tongyu Qian [1] and Ying Chen [2]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; wanghongquan@tju.edu.cn (H.W.); sunhao17@baidu.com (H.S.); tongyuqian0115@hotmail.com (T.Q.)
[2] Beijing SGITG-ACCENTURE Information Technology Co., Ltd., Beijing 100052, China; cheny15@tsinghua.org.cn
* Correspondence: xszhu@tju.edu.cn

Abstract: With the continuous advancement of image-editing technologies, it is particularly important to develop image forensics methods for digital information security. In this study, a deep neural network called multi-path inpainting forensics network (MPIF-Net) was developed to locate the inpainted regions in an image. The interaction of shallow and deep features between different paths was established, which not only preserved detailed information but also allowed for the further mining of deep features. Meanwhile, an improved residual dense block was employed as the deep feature extraction module of each path, which can enhance the feature extraction ability of the model by introducing a frequency domain attention mechanism. In addition, a boundary guidance module was constructed to alleviate the prediction distortion in the boundaries of the inpainted region. Finally, extensive experimental results regarding various deep inpainting datasets demonstrated that the proposed network can accurately locate inpainted regions, exhibit excellent generalization and robustness, and verify the effectiveness of the designed module.

Keywords: image-inpainting forensics; deep learning; attention mechanism; ffrequency domain

Citation: Wang, H.; Zhu, X.; Sun, H.; Qian, T.; Chen, Y. A Multi-Path Inpainting Forensics Network Based on Frequency Attention and Boundary Guidance. *Electronics* **2023**, *12*, 3192. https://doi.org/10.3390/electronics12143192

Academic Editor: Chiman Kwan

Received: 11 June 2023
Revised: 14 July 2023
Accepted: 19 July 2023
Published: 24 July 2023

1. Introduction

With the advancement of deep learning (DL), especially with respect to image generation technology, many novel image-tampering methods, such as inpainting and deepfakes, have rapidly developed. These methods leave weak tampering traces and thus their influence is more difficult to recognize. This poses a serious threat to the security of digital media, which, in turn, negatively impacts fields such as science, politics, and commerce and, ultimately, may even undermine social stability. On the contrary, existing research on image-tampering forensics is still relatively limited, and it is difficult to cope with the emerging image-tampering technologies. Therefore, in the field of information security, the demand for digital image forensics technology continues to increase. It is critical to develop practical forensic methods to counter these rapidly improving digital image-editing technologies [1].

Image inpainting is a fundamental research topic in the field of computer vision and image processing. The aim of this process is to repair damaged image information or remove unwanted content in an imperceptible way [2–4], as shown in the example in Figure 1. In the past two decades, academia has proposed a large number of image-inpainting algorithms. In particular, the rapid development of deep learning technology has greatly accelerated the progress of this technology in recent years [5–8]. Many mature inpainting technologies have not only been widely used but also integrated into some advanced image-editing software. However, image-inpainting technologies have also become convenient tools for maliciously tampering with images, so the development of forensic methods for inpainting has become a major topic in the field of digital image forensics. Due to the diversity and continuous innovation of inpainting technologies,

along with the joint effect produced by their combination with other image manipulations, inpainting forensics is more problematic and more complex than other areas of image manipulation forensics [1].

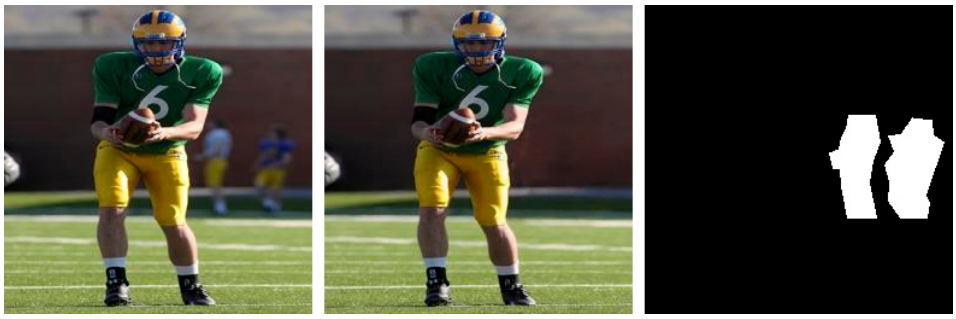

Figure 1. Original image (**left**), inpainted image (**middle**), and ground truth mask (**right**).

The existing forensic methods for image inpainting can be divided into two categories: conventional inpainting forensics methods and deep-learning-based inpainting forensics methods.

1.1. Conventional Inpainting Forensics Methods

Traditional forensic methods rely on manually designed feature extraction. Initially, Wu et al. [9] proposed the zero-connectivity length (ZCL) feature to measure the similarity between image patches; however, a significant drawback of this feature was that it requires the manual selection of suspicious regions in advance. Bacchuwar et al. [10] and Chang et al. [11] proposed similar nearest neighbor image patch search methods that can accelerate the search process, but they can also lead to a decrease in search accuracy. The maximum zero-connectivity component was constructed to label features in [12], which can improve the search speed of suspicious image patches through central pixel mapping, but the inpainted region is easily identified as some isolated and suspicious regions that should be ignored, leading to a decrease in accuracy.

Jin et al. [13] designed a robust inpainting forensic method based on canonical correlation analysis (CCA); however, it could not locate the tampered region. Zhang et al. [14] established a joint probability density matrix (JPDM) to represent the correlation of adjacent discrete cosine transform (DCT) coefficients. This method has good robustness against post-processing operations, but like the method presented in [13], it also fails to achieve the localization of inpainted regions.

The above forensics methods are all proposed for exemplar-based inpainting [2]. For diffusion-based inpainting technology [4], Li et al. [15] proposed constructing feature vectors using local variances within and between color channels, and an ensemble classifier was trained to locate inpainted regions. This method is effective, but its robustness needs to be further improved. On this basis, the method's forensic performance was further improved via weighted least squares filtering [16].

Since image inpainting, especially deep-learning-based inpainting, does not leave apparent traces of manipulation, it is often quite difficult to manually design the corresponding forensics features. In addition, the features of traditional forensics are generally designed based on the principles of inpainting approaches and the observation of a small number of sample images. Therefore, the limitations of manually designing features result in the unsatisfactory effectiveness and robustness of the corresponding forensics methods.

1.2. Deep-Learning-Based Inpainting Forensics Methods

Regarding exemplar-based inpainting [2], Zhu et al. [17] developed a full convolutional network (FCN) based on the encoder–decoder structure and adopted pixel-level labelling and a weighted cross entropy loss function to determine the location of the inpainting region. This method is significantly superior to conventional inpainting forensics in terms

of detection accuracy and robustness and can further improve performance by introducing a skip structure [18]. An inpainting forensics method combining a CNN and a long short-term memory (LSTM) network was proposed in [19]; it helped to improve robustness and reduce the false alarm rate. Wang et al. [20] constructed a forensic network based on Mask R-CNN. This method enables location, recognition, and density prediction for exemplar-based inpainting. Further, considering that the FPN of Mask R-CNN cannot fully utilize all scale feature information, a multi-task deep learning method [21] was designed by combining feature pyramid networks and back connections, allowing for the acquisition of more feature information.

Li et al. [22] proposed the first forensic network for deep inpainting approaches. The network uses residual blocks to construct the backbone for feature extraction, and high-pass filters were designed as a preprocessing module to enhance inpainting traces and improve localization performance. For diffusion-based inpainting, Zhang et al. [23] designed a U-Net-based forensics network that utilizes a feature pyramid network to enhance multi-scale feature representation and constructs a stagewise weighted cross entropy loss function to improve localization accuracy. Liu et al. [24] proposed a progressive spatial channel correlation network (PSCC-Net) based on the backbone network of a multi-stream structure [25] that can locate the tampered regions in images that are spliced, copy-moved, and inpainted. Recently, Wu et al. [26] proposed an end-to-end inpainting detection network that was dubbed IID-Net. IID-Net includes a feature enhancement module, a feature extraction module, and a decision module, and uses a neural structure search algorithm [27] to automatically design the structure of the feature extraction module, thereby significantly improving forensic accuracy and robustness. Finally, Wu et al. [28] constructed a general forensic network, known as ManTra-Net, which is composed of two parts: a manipulation-trace feature extractor and a local anomaly detection network. This network can detect multiple types of image manipulation, such as splicing, copy-moving, inpainting, etc.

Through end-to-end learning, DCNN can directly learn features and optimize final decisions from a set of data. Therefore, the DL-based inpainting forensics method can circumvent the difficult process of manually extracting features and achieve significantly better performance than conventional forensics methods. However, in response to the continuous development of inpainting technology, there are still some issues that need to be addressed with respect to the current deep forensic methods, such as the lack of discriminative inpainting features, significant false alarms and missed detections, loss of detail information, distorted boundary localization, robustness that needs further improvement, etc.

In order to enhance inpainting forensics, the employed network should be able to effectively capture traces of inpainting procedures and accurately locate the boundaries of the inpainted region. Thus, in accordance with the above consideration, this paper proposes a multi-path inpainting forensics network (MPIF-Net) based on frequency attention and boundary guidance. The main contributions of this work are as follows.

First, a network structure with multiple parallel paths is proposed that facilitates the capture of more abundant detail information, and connections between shallow and deep layers on different paths are constructed to fuse features and promote the reuse of deep features.

Second, by introducing an improved attention mechanism, a deep feature extraction module is designed to further enhance the feature extraction ability in relation to inpainting traces. Then, a boundary guidance module is designed to make the network more attentive to the boundaries of the inpainted region.

Finally, the proposed inpainting forensics method is compared with state-of-the-art methods with respect to their performance when applied to established deep inpainting datasets, and the robustness and generalization of the model are verified.

The remainder of this paper is organized as follows. In Section 2, the design concept and details of MPIF-Net are carefully described. Next, Sections 3 and 4 introduce two spe-

cially designed modules in the network, respectively, while the loss function for training the network is proposed in Section 5. Then, a series of experiments performed to evaluate the proposed MPIF-Net is presented in Section 6. Finally, Section 7 concludes this paper.

2. The Multi-Path Inpainting Forensics Network

In this section, the design concept of the MPIF-Net is described in detail, followed by its structure, and the decoder and encoder designs of the MPIF-Net are introduced.

2.1. Overview of MPIF-Net

In this paper, a network structure with a multi-path framework designed in parallel is proposed (as shown in Figure 2). The network structure can combine information from different paths to reduce the loss of detail information, thereby improving the ability of the model to capture detail information. Considering the great success of the encoder–decoder structure [29] in image segmentation and generation and other visual tasks, this structure is employed as the basic structure of each path in this paper.

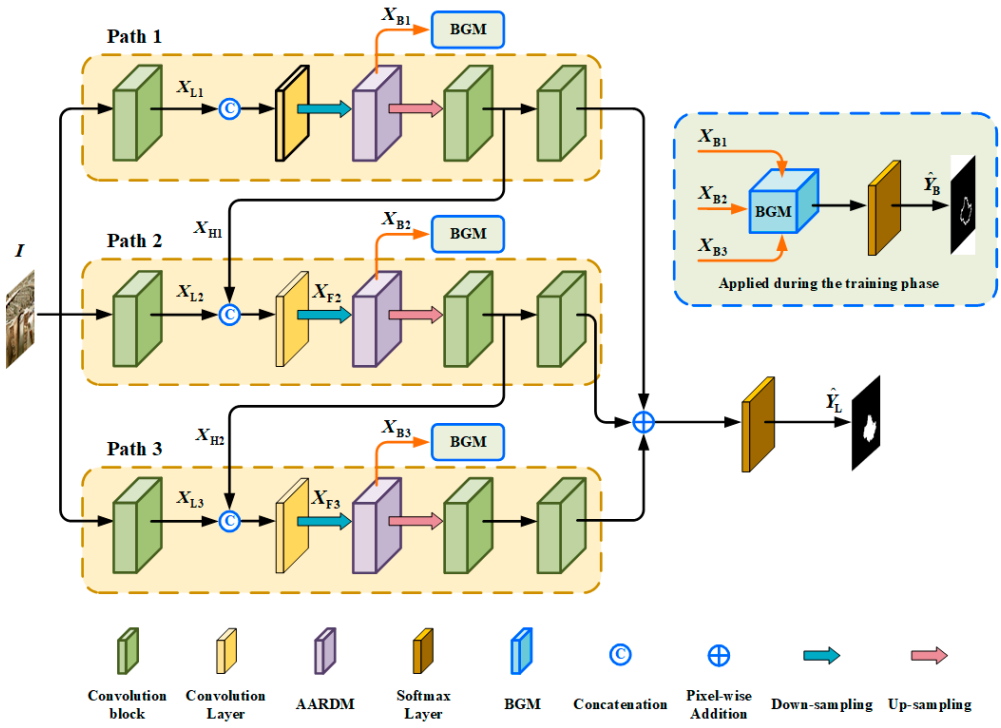

Figure 2. The architecture of MPIF-Net.

At the same time, an attention-aware residual dense block (AARDB) with an attention mechanism is designed as the deep feature extraction module of the network. This module combines the attention mechanism with a Residual Dense Block (RDB) [30] to effectively enhance the feature extraction ability of the network.

In addition, due to the importance of boundary information in inpainting forensics tasks, a boundary guide module (BGM) is proposed to further improve the forensics performance of the proposed model. The BGM is utilized to encourage the model to pay more attention to the boundaries of the inpainted region.

In the conventional encoder–decoder network, progressive down-sampling operations are usually used to capture high-level semantic features, but this inevitably causes the loss of detail information. This is unfavorable for image inpainting forensic tasks. To address the issue and improve feature expression, a network for inpainting forensics was established in a parallel multi-path style, and the connections between shallow and deep layers on different paths were constructed for information exchange and the reuse of deep features. Shallow features that do not lose resolution retain abundant detail information,

while deep features contain higher-level semantic information, allowing for the acquisition of more discriminative features through their fusion. Finally, the efficient feature extraction module AARDB was employed to further mine deep features related to inpainting traces from fused features while providing them to the next path for reuse.

2.2. Design of Encoder

The three paths of the encoder were developed using the same structure, that is, a convolution block, an independent convolution layer, and an attention-aware residual dense module (AARDM).

Firstly, given an image $\boldsymbol{I} \in \mathbb{R}^{W \times H \times 3}$ that is considered forensic, low-level features with 16 channels are extracted through a convolutional block for each path, which consists of a convolution layer with a kernel size of 3×3, a batch normalization layer, and a ReLU activation layer. The process of generating low-level features $\boldsymbol{X}_{\mathrm{L}i}$ on the i-th path can be expressed as

$$\boldsymbol{X}_{\mathrm{L}i} = \mathcal{F}_{\mathrm{C}i}(\boldsymbol{I}) i \in \{1, 2, 3\} \tag{1}$$

where $\mathcal{F}_{\mathrm{C}i}(\cdot)$ represents the composite operation performed by the above convolution block on the i-th path. Then, excluding the first path, the currently obtained shallow features in path 2 and path 3 are concatenated and fused with the deep features input from the previous path, which is implemented through an independent 3×3 convolutional layer. Although the first path does not introduce high-level features, the first path is also set with the same convolutional layer at the same location to ensure a balance of structure and performance. Let $\mathcal{F}_{\mathrm{F}i}(\cdot)$ represent the feature fusion performed by the independent convolutional layer on the i-th path; then, the process of obtaining fused features $\boldsymbol{X}_{\mathrm{F}i}$ described above can be written as follows:

$$\boldsymbol{X}_{\mathrm{F}i} = \mathcal{F}_{\mathrm{F}i}([\boldsymbol{X}_{\mathrm{L}i}, \boldsymbol{X}_{\mathrm{H}i-1}]) i \in \{2, 3\} \tag{2}$$

where $\boldsymbol{X}_{\mathrm{L}i}$ and $\boldsymbol{X}_{\mathrm{H}i-1}$ represent low-level features generated in the shallow layer of the i-th path and high-level features generated in the deep layer of the previous path (i.e., the $i-1$-th path), respectively, and $[\cdot]$ denotes the concatenation of features along the channel. Next, feature down-sampling is performed through convolution with a kernel size of 3×3 and a stride of 2, reducing the resolution of the feature maps by $1/2$ and expanding the channel number of feature maps to 32. Finally, the feature maps are fed into an efficient feature extraction module, namely, AARDM, to further extract deep features. The extracted deep features are fed into the decoders in their respective paths to generate the final localization prediction results and are also provided to a boundary guidance module for additional supervision of the inpainted region boundary.

2.3. Design of Decoder

The decoder for each path is also designed using the same structure. First, the feature maps are bilinearly interpolated using the up-sampling layer to restore their spatial resolution. Then, two consecutive convolutional blocks are utilized to gradually convert the feature maps into 16-channel and 2-channel feature maps in sequence, while further refining the feature representations regarding the localization results.

In addition, it should be noted that the feature maps output by the first convolutional block of the decoders in path 1 and path 2 are input to the next paths of the network (namely, path 2 and path 3, respectively), which are concatenated with the output features of the first convolution block of the encoder in the next path to carry out feature fusion and feature learning. This design can further expose the inpainting traces by reusing deep features from previous paths, and it can reduce the loss of detail information through shallow and deep feature fusion.

Finally, the two-channel feature maps obtained from the three paths are fused through pixel-wise addition, and the fused results are fed to a Softmax layer, yielding the final probability map $\hat{\boldsymbol{Y}}_{\mathrm{L}}$ for localization.

3. Attention-Aware Residual Dense Module

3.1. Detailed Design of Module Structure

Residual dense blocks [30] constitute a feature extraction module with excellent feature-learning ability. They are designed through combining dense connections [31] with residual learning [32]. This study designed an improved attention module based on frequency information, namely, the frequency convolutional block attention module (FCBAM), which is introduced into the RDBs to enhance the representation of output features. We refer to these RDBs introduced into the FCBAM as attention-aware residual dense modules (AARDMs) and place them at the end of the decoder in each path to extract more discriminative deep features.

As shown in Figure 3, an AARDM consists of four densely connected 3×3 convolution blocks with a ReLU layer, one convolution layer that reduces the channel number of features, one FCBAM for improving feature representation, and one residual connection.

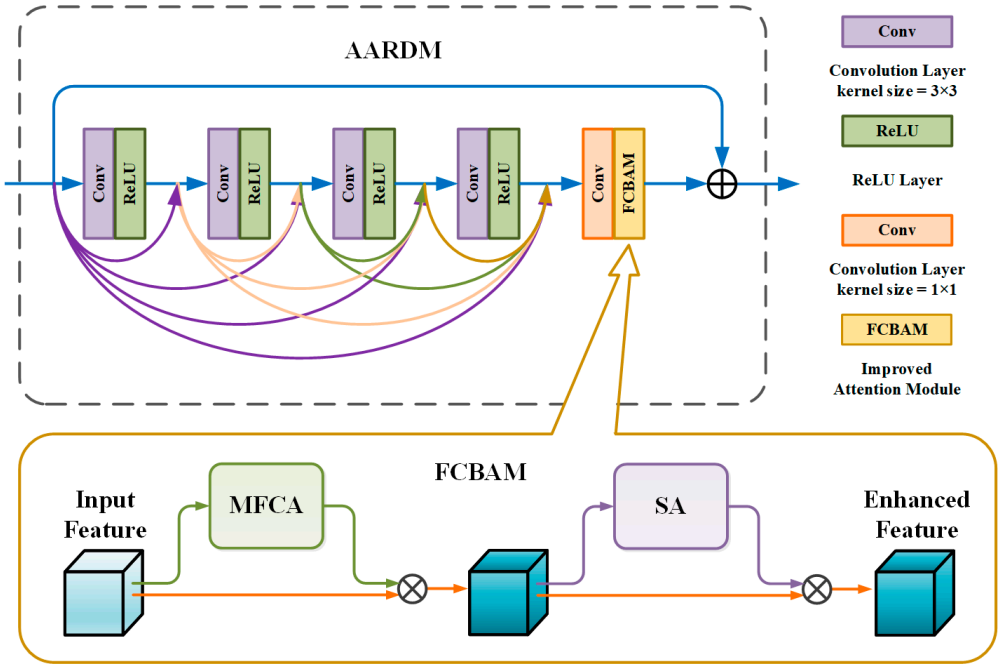

Figure 3. The structure of the residual dense block with FCBAM.

Firstly, each convolutional block can utilize the input features $\boldsymbol{X}_0$ of AARDM and the ones extracted by all previous convolution blocks to facilitate feature learning and yield 16 feature maps (i.e., a growth rate of 16). Taking the fourth convolution block as an example, the above process can be expressed as follows:

$$\boldsymbol{X}_4^i = \mathcal{F}_{\mathrm{C}}([\boldsymbol{X}_0^i, \boldsymbol{X}_1^i, \boldsymbol{X}_2^i, \boldsymbol{X}_3^i]) \tag{3}$$

where $\boldsymbol{X}_n^i, n \in \{1, 2, 3, 4\}$ denotes the output of the n-th convolutional block of AARDM in the i-th path, and $\boldsymbol{X}_0^i$ denotes the input features of this AARDM. $\mathcal{F}_{\mathrm{C}}(\cdot)$ and $[\cdot]$ represent the composite operation performed by the convolution block and the feature concatenation along the channel dimension, respectively.

Then, after all generated features and the input features are concatenated, and a 1×1 convolutional layer is used to fuse the connected features and convert the channel number of connected features into the same values as the input features. Subsequently, in order to improve the representation of fused features, FCABM is employed to recalibrate them in order to obtain enhanced features $\boldsymbol{X}_{\mathrm{R}}^i$:

$$\boldsymbol{X}_{\mathrm{R}}^i = \mathcal{F}_{\mathrm{FA}}(\mathcal{F}_{\mathrm{C1}}([\boldsymbol{X}_0^i, \boldsymbol{X}_1^i, \ldots, \boldsymbol{X}_4^i])) \tag{4}$$

where $\mathcal{F}_{C1}(\cdot)$ and $\mathcal{F}_{FA}(\cdot)$ indicate the operation performed by the convolution layer with the kernel size of 1×1 and the FCBAM, respectively. The specific design of the FCBAM will be comprehensively described later.

Finally, the input features of the AARDM are introduced through residual connections and added to the enhanced features, which can ensure stable network training and further promote network convergence; this process can be expressed as follows:

$$\boldsymbol{X}_{\text{out}}^{i} = \boldsymbol{X}_{\text{R}}^{i} \oplus \boldsymbol{X}_{0}^{i} \tag{5}$$

where $\oplus$ denotes pixel-wise addition, and $\boldsymbol{X}_{\text{out}}^{i}$ denotes the final output of the AARDM in the i-th path. In MPIF-Net, the resolution of the output and input feature maps of AARDM is consistent in each path, and their channel number is also set to 32 in order to perform residual learning through direct addition.

Unlike RDBs, we employ the improved attention mechanism to optimize the feature fusion, which can be beneficial for enhancing the representation of deep features.

3.2. Improved Attention Mechanism

As shown in Figure 3, all features in the AARDM are fused into a feature response containing abundant information through a 1×1 convolutional layer in the AARDM. In order to further enhance the features closely related to forensic tasks and suppress redundant information, an improved attention module FCBAM is introduced at the end of the AARDM to improve the representation of fused features.

An FCBAM includes two parts, namely, frequency channel attention (FCA) and spatial attention (SA), which recalibrate the input features on the channel and spatial dimensions, as shown in the orange box of Figure 3. An FCBAM has a similar structure to a CBAM [33], except that it utilizes channel attention based on frequency to enhance features.

Qin et al. [34] found that only the lowest frequency component is preserved through global average pooling (GAP) in the channel attention, while other frequency information is completely discarded. Inspired by this, we adopt abundant frequency information in the frequency channel attention of the FCBAM to generate channel attention maps. Specifically, the implementation process of FCA is as follows:

(1) Firstly, the input $\boldsymbol{X}$ is evenly divided into n parts along the channel dimension, which are denoted by the set $\{\boldsymbol{X}^0, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^{n-1}\}$, $\boldsymbol{X}^i \in \mathbb{R}^{W \times H \times C'}$, and $C' = C/n$. Each part is assigned a corresponding 2D DCT frequency component, which is used to calculate the 2D DCT results of that part.

(2) Then, the 2D DCT results of all the parts are concatenated as the multi-spectral channel descriptors $\boldsymbol{F}_{\text{C}}$ of the input $\boldsymbol{X}$, which contains more information, and the process can be expressed as

$$F_{\text{C}} = [F_{\text{C}}^0, F_{\text{C}}^1, \cdots, F_{\text{C}}^{n-1}] \tag{6}$$

where $F_{\text{C}}^i \in \mathbb{R}^{C'}$ is a C'-dimensional vector that denotes the 2D DCT results of the i-th parts $\boldsymbol{X}^i$, and $[\cdot]$ represents the concatenation of features along the channel dimension.

(3) Finally, the channel attention maps $\boldsymbol{M}_{\text{MSC}}$ based on a multi-spectral descriptor $\boldsymbol{F}_{\text{C}}$ are obtained as follows:

$$\boldsymbol{M}_{\text{MSC}} = \sigma(\mathcal{F}_{\text{MLP}}(\boldsymbol{F}_{\text{C}})) \tag{7}$$

where $\sigma(\cdot)$ is a Sigmoid function, and $\mathcal{F}_{\text{MLP}}(\cdot)$ denotes a multi-layer perceptron (MLP) with a hidden layer and an output layer.

Based on the above design, the FCBAM module in the AARDM processes the fused features $\boldsymbol{X}$, and this procedure can be expressed as follows

$$\boldsymbol{X}' = \boldsymbol{M}_{\text{S}} \otimes (\boldsymbol{M}_{\text{MSC}} \otimes \boldsymbol{X}) \tag{8}$$

where $\otimes$ denotes element-wise multiplication, and X' represents the enhanced feature output by the FCBAM. An ablation experiment demonstrated the effectiveness of the FCBAM in the AARDM.

4. Boundary Guidance Module

The target of image forensics is very similar to the target of semantic segmentation, and accurately fitting the boundary of the mask is one of its main challenges. Therefore, we have designed a boundary guidance module to enhance the localization of the boundary of the inpainted region in order to further improve forensic performance.

As shown in Figure 4, the final feature responses generated by the decoders of the three paths are fused via the boundary guidance module (BGM); then, the fused features are utilized to predict the boundary of the inpainted region. By monitoring the prediction results, the network can pay more attention to the boundary of the inpainted region; thus, the prediction distortion of the boundary can be effectively alleviated.

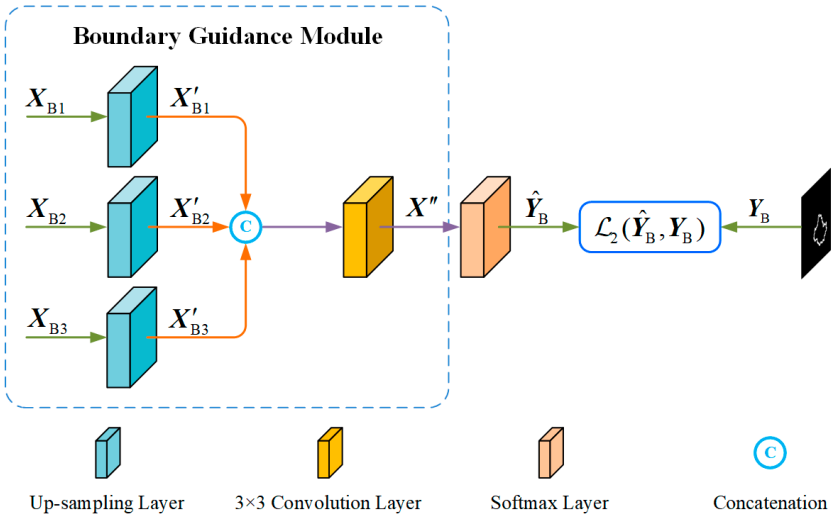

Figure 4. The structure of the boundary guidance module.

Specifically, through the proposed network structure, shallow and deep features on different paths are fused and fed to the specially designed AARDB to further improve feature expression. Let X_{B1}, X_{B2}, and X_{B3} represent the deep features extracted by the AARDM module of each path. In this process, the shallow features on the current path can compensate for the detail information, while the deep features introduced from the previous path can be further utilized. The resolution of the above features is first restored to the size of the input image through the up-sampling layers as follows:

$$X'_{Bi} = \mathcal{F}_{US}(X_{Bi}) i \in \{1, 2, 3\} \tag{9}$$

where $\mathcal{F}_{US}(\cdot)$ denotes the up-sampling operation, and X'_{Bi} represents the result of up-sampling the corresponding feature X_{Bi}. Afterward, all the restored resolution features are concatenated along the channel dimension and then fused through a convolutional layer with a kernel size of 3×3, as follows:

$$X'' = \mathrm{Conv}([X'_{B1}, X'_{B2}, X'_{B3}]) \tag{10}$$

where $[\cdot]$ represents the features' concatenation along the channel dimension, $\mathrm{Conv}(\cdot)$ denotes the convolutional operation, and the dimensions of the fused features X'' are consistent with those of X'_{Bi}. Therefore, the fused features X'' contain abundant spatial details and high-level semantic information obtained through different paths.

It is worth noting that the BGM is only used during the training phase of the model, so the application of this module will not increase the inference time during the testing phase. An ablation study verified the effectiveness of BGM.

5. Loss Function

In order to obtain an optimal forensics model, it is necessary to design a loss function to measure the consistency between the predicted results yielded by the model and the true values and provide a basis upon which the optimizer can update the network parameters simultaneously. An appropriate loss function can promote convergence and improve the forensic performance of a network. As a loss function commonly used in various dense prediction tasks, including inpainting forensics [17,24], this paper employs cross-entropy loss as the main component in order to propose a loss function consisting of three parts. Firstly, loss L_1 indicates the difference between the final localization prediction results of the inpainting region $\hat{\boldsymbol{Y}}_{\mathrm{L}}$ obtained via the pixel-wise addition of the outputs of the three decoders and the ground truth mask $\boldsymbol{Y}_{\mathrm{L}}$:

$$L_1 = \mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{Y}}_{\mathrm{L}}, \boldsymbol{Y}_{\mathrm{L}}) \tag{11}$$

where $\mathcal{L}_{\mathrm{CE}}(\cdot, \cdot)$ denotes the cross-entropy loss (CE) function. Then, the cross-entropy loss L_2 between the output $\hat{\boldsymbol{Y}}_{\mathrm{B}}$ of the boundary guidance module and the boundary label $\boldsymbol{Y}_{\mathrm{B}}$ is introduced to enhance the ability to learn information about the pixels near the boundary of the inpainting region during training, which can be obtained using the following expression:

$$L_2 = \mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{Y}}_{\mathrm{B}}, \boldsymbol{Y}_{\mathrm{B}}) \tag{12}$$

Further, after adding the prediction results $\hat{\boldsymbol{Y}}_{\mathrm{L}}$ and $\boldsymbol{X}_{\mathrm{B}}$, the cross-entropy loss L_3 can be used to calculate the difference between the sum result and the ground truth mask, and L_3 can be expressed as

$$L_3 = \mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{Y}}_{\mathrm{L}} + \hat{\boldsymbol{Y}}_{\mathrm{B}}, \boldsymbol{Y}_{\mathrm{L}}) \tag{13}$$

Finally, the overall forensics loss function L of MPIF-Net is as follows

$$L = \alpha * L_1 + \beta * L_2 + \gamma * L_3 \tag{14}$$

where α, β, and γ are the hyperparameters indicating the weights of each loss.

6. Experimental Results and Analysis

In order to validate the forensic performance of the proposed MPIF-Net, we established an inpainting forensics dataset on which extensive experiments were carried out. Intersection over Union (IoU), the F1 score, the true positive rate (TPR), and the false positive rate (FPR) were employed as performance metrics. Finally, an ablation study was conducted to verify the main components of our proposed network.

6.1. Dataset

In this paper, 19,350 color images with dimensions of 256 × 256 were randomly selected from the Places2 [35] dataset. A random region was removed from each image, and these regions were tampered with using image-inpainting approaches. In addition, there were three different shapes of the removed regions, namely, circular, rectangular, and irregular regions, with the sizes equal to 1%, 5%, and 10% of the original image. To fully verify the forensic performance of MPIF-Net, several trained inpainting models [5–8] based on deep learning were utilized to inpaint each image, resulting in a total of 19,350 × 4 inpainted images. The final obtained dataset was randomly divided into a training set with 18,000 × 4 images, a validation set with 450 × 4 images, and a test set with 900 × 4 images. Several sample images with random masks (marked in green) are shown in Figure 5.

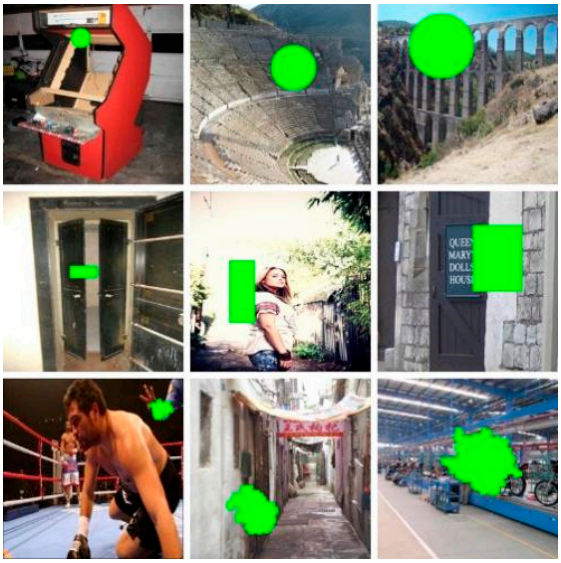

Figure 5. Sample images with masks (marked in green) in datasets.

6.2. Training Details

The proposed MPIF-Net with an input with dimensions of 256 × 256 was implemented using PyTorch. It was trained and tested on an NVIDIA GeForce GTX 3080Ti GPU with 12 GB of memory using the Adam optimizer. The mini-batch was set to 24. The momentum decay parameters β_1 and β_2 of ADAM were set to 0.9 and 0.999, respectively. Training was executed for 100 epochs, and the learning rate of the model was initialized to 3×10^{-3}, decreasing by 50% every 10 epochs.

For comparison, several comparative methods were selected, including those produced by Zhu et al. [17], Chen et al. [36], Li et al. [22], and Wu et al. [26]. The methods presented in [26,36] were trained as described in their papers, while the methods presented in [17,22] were trained using the same training method as that of MPIF-Net.

6.3. Forensic Performance Evaluation for Inpainted Images without Post-Processing Operations

The proposed network MPIF-Net was first subjected to qualitative and quantitative evaluations using inpainted images without post-processing operations. Figure 6 shows the visualized forensic results for all the methods with respect to the deep inpainting dataset produced using the inpainting approach [6]. The forensic results of the comparison methods are shown in Figure 6c–g, and the real inpainted regions are shown in Figure 6b. It can be clearly observed that all the forensic methods were capable of localizing the inpainted regions to some extent. When the inpainted region was relatively large, regardless of its shape, all methods were able to achieve a rough localization of the inpainted region to some extent. However, as the inpainted region became smaller, the forensic difficulty gradually increased (e.g., rows 1, 4, and 7 in Figure 6). Taking the first row as an example, the first row of Figure 6c shows the results of the method presented in [17], which only locates a part of the inpainted region, and there are also obvious false alarm regions. The method presented in [36] did not capture the correct inpainting features, resulting in the recognition of the inherent pixels in the image as inpainted ones, as shown in the first row of Figure 6d. Then, the method presented in [22] located fewer inpainted pixels compared to the other methods, but there were basically no false detections. As for the method presented in [26], although most of the inpainted pixels were captured, a large number of false alarm pixels appeared in the forensic results. Surprisingly, MPIF-Net (the last row of Figure 6f) was able to locate almost every real inpainted region with relatively few false alarms. Overall, the forensic results of the proposed MPIF-Net not only better fit the real inpainted regions in terms of size and shape but also present fewer false alarms, while the results regarding

the other compared methods all exhibit varying degrees of distortion, false alarms, and even omissions.

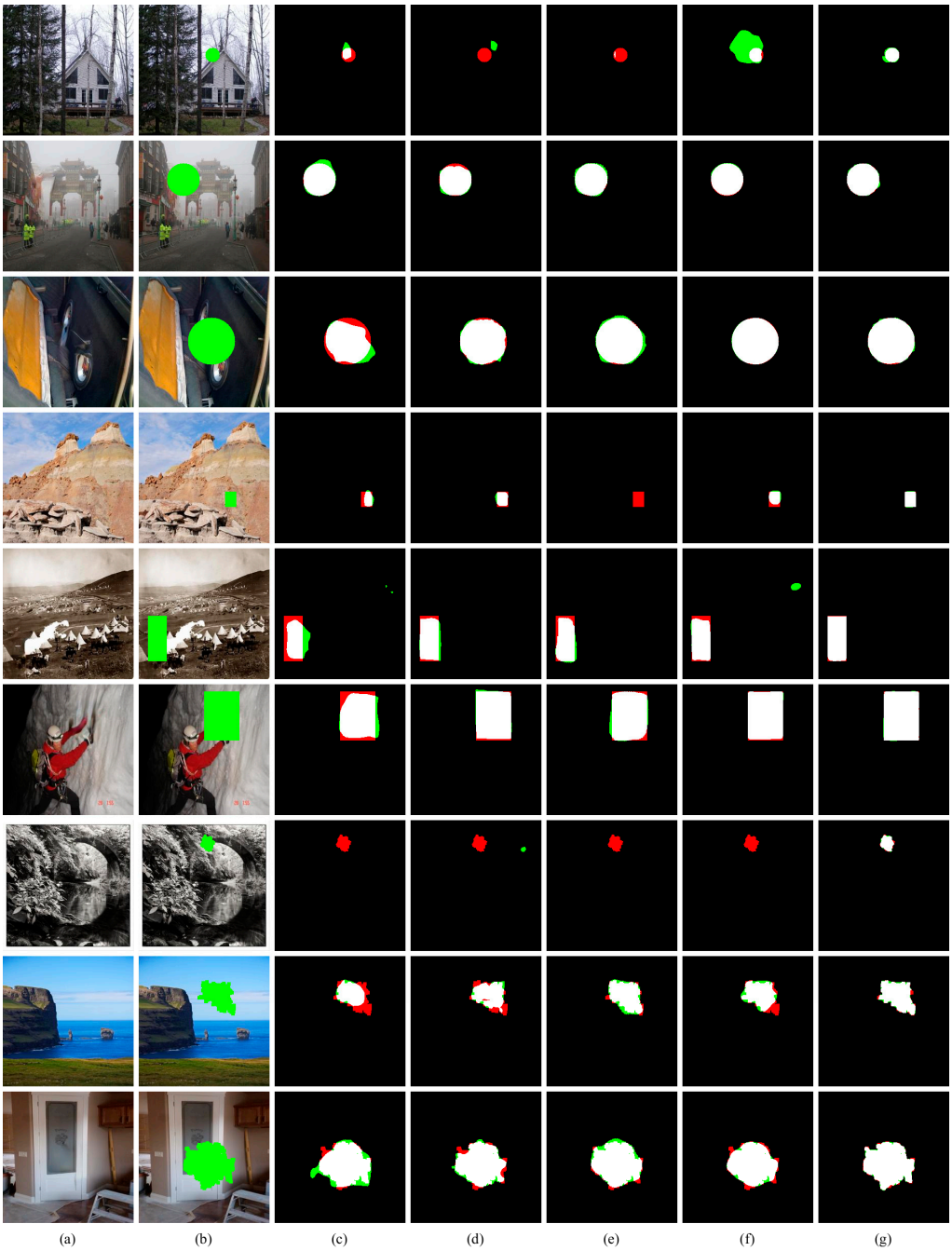

Figure 6. Qualitative comparison results of different forensic methods regarding deep inpainting approach [6]. The original uninpainted images, mask images, forensic results obtained by the methods in Ref. [17], Ref. [36], Ref. [22], Ref. [26] and proposed MPIF-Net are respectively shown in (**a**–**g**), where the pixels in white, black, green, and red indicate true positive, true negative, false positive, and false negative, respectively.

The quantitative forensic results for all the forensic methods applied to the deep inpainting dataset produced using the inpainting approach [6] are presented in Tables 1 and 2, and the best results are indicated in bold.

Table 1. Average IoU (%) and F1 (%) score of different forensic methods with respect to deep inpainting approach [6].

Method	Metric	Circular			Rectangle			Irregular			Mean
		1%	5%	10%	1%	5%	10%	1%	5%	10%	
Ref. [17]	$F1^+$	43.14	82.62	90.33	44.93	78.94	88.52	50.41	78.43	86.37	71.52
	IoU^+	34.71	73.39	83.45	35.26	68.78	80.60	40.00	66.93	77.63	62.30
Ref. [36]	$F1^+$	76.71	94.24	97.59	73.47	93.58	94.72	72.79	87.45	89.46	87.02
	IoU^+	66.76	90.58	95.37	62.91	88.68	90.79	61.83	78.84	81.71	80.27
Ref. [22]	$F1^+$	55.30	91.15	94.46	56.08	90.02	92.78	61.09	86.09	90.37	79.71
	IoU^+	46.16	84.26	89.64	46.32	82.74	87.05	50.18	76.85	83.20	71.83
Ref. [26]	$F1^+$	78.48	95.50	97.12	75.39	94.18	96.42	76.42	87.59	91.35	88.05
	IoU^+	68.84	91.54	94.66	64.55	89.19	93.16	64.55	78.72	84.79	81.11
MPIF-Net	$F1^+$	**93.39**	**97.10**	**98.49**	**93.96**	**98.13**	**98.34**	**90.12**	**95.30**	**96.68**	**95.73**
	IoU^+	**88.08**	**94.78**	**97.03**	**88.95**	**96.36**	**96.83**	**82.71**	**91.07**	**93.59**	**92.17**

Table 2. Average TPR (%) and FPR (%) scores of different forensic methods with respect to deep inpainting approach [6].

Method	Metric	Circular			Rectangle			Irregular			Mean
		1%	5%	10%	1%	5%	10%	1%	5%	10%	
Ref. [17]	TPR^+	39.85	80.71	89.45	44.56	76.85	87.06	47.97	75.70	86.29	69.82
	FPR^-	0.10	0.48	0.81	0.21	0.60	0.89	0.16	0.66	1.21	0.57
Ref. [36]	TPR^+	77.58	93.99	97.08	71.86	93.63	94.64	71.41	87.93	89.73	86.74
	FPR^-	0.16	0.20	0.20	0.17	0.31	0.32	0.16	0.65	0.78	0.35
Ref. [22]	TPR^+	50.51	93.66	97.56	52.53	92.40	94.78	57.08	87.89	93.39	79.98
	FPR^-	0.07	0.62	1.02	0.14	0.63	0.99	0.12	0.75	1.37	0.63
Ref. [26]	TPR^+	77.12	94.56	96.14	70.80	92.51	95.62	71.53	84.92	89.40	85.85
	FPR^-	0.19	0.18	0.18	0.12	0.21	0.30	0.12	0.41	0.60	0.26
MPIF-Net	TPR^+	**93.25**	**97.74**	**98.43**	**93.70**	**98.05**	**98.10**	**88.61**	**95.20**	**96.65**	**95.53**
	FPR^-	**0.05**	**0.12**	**0.16**	**0.06**	**0.09**	**0.15**	**0.07**	**0.24**	**0.36**	**0.16**

By referring to the results regarding the test images with circular masks, it can be found that the proposed MPIF-Net has achieved the best performance among all the other forensic methods. For example, when the size of the inpainted region was 10% of the image, MPIF-Net presented the highest F1 score of 98.49% and an IoU of 97.03%. Simultaneously, its TPR also corresponded to the highest value, namely, 98.43%, and its FPR was the lowest, namely, only 0.16%. Although the various metrics of all the compared methods show obvious degradation as the size of the inpainted region decreases, MPIF-Net still has the best performance among all the forensic methods. Similarly, for masks of other shapes, the proposed MPIF-Net also exhibits a significantly superior IoU, F1 score, and TPR and maintains the lowest FPR compared to the other comparison methods. Specifically, for inpainted images with a tamper rate of only 1%, the network appears to have more outstanding performance advantages. Taking the F1 score as an example, MPIF-Net's scores are about 14.9%, 18.6%, and 13.7% higher than the suboptimal method presented in [26] for circular, rectangular, and irregular masks, respectively. It can be seen that even for smaller inpainted regions that are more difficult to detect, MPIF-Net can obtain relatively accurate localization results, indicating that the proposed network can mine abundant inpainted features and retain more details.

6.4. Robustness against Post-Processing Operations

In order to prevent inpainting traces from being easily detected, malicious tamperers generally perform post-processing operations, such as JPEG compression and additive white Gaussian noise (AWGN) procedures, on inpainted images. Therefore, the robustness

of MPIF-Net against JPEG compression and AWGN was evaluated in relation to the dataset produced through the inpainting approach [6].

Firstly, the testing images were JPEG-compressed by quality factors (QF) of 95 and 75, and all the analyzed forensics methods were performed on the compressed images with an inpainted ratio greater than 1%. The average values of the IoU and F1 scores are listed in Table 3. Apparently, due to the influence of JPEG compression, the forensic performances of all methods were significantly weakened with the decrease in the QF. For instance, MPIF-Net obtained an IoU of 81.72% and an F1 score of 89.81% for a QF of 95. When the QF was further decreased to 75, the IoU and F1 of MPIF-Net decreased to 62.13% and 74.32%, respectively. It can be seen that some high-frequency information containing inpainting traces was lost due to JPEG compression, which led to a decrease in forensic performance. However, MPIF-Net exhibited good robustness against JPEG compression and still significantly outperformed the other methods.

Table 3. Average IoU (%) and F1 (%) scores of different forensic methods under the influence of JPEG compression and additive white Gaussian noise.

Types of Distortions	Ref. [17]		Ref. [36]		Ref. [22]		Ref. [26]		MPIF-Net	
	IoU^+	$F1^+$	IoU^+	$F1^+$	IoU^+	$F1^+$	IoU^+	$F1^+$	IoU^+	$F1^+$
w/o Dis.	75.13	84.20	87.66	92.84	83.92	90.81	88.68	93.69	94.94	97.34
JPEG95	52.78	63.94	77.02	86.09	76.29	85.84	71.23	81.13	81.72	89.81
JPEG75	45.89	58.71	54.85	65.47	53.32	65.49	55.98	66.55	62.13	74.32
50 dB	46.09	58.86	55.15	65.69	52.77	64.68	56.38	67.07	62.15	74.27
40 dB	46.22	59.05	55.21	65.78	53.06	65.05	56.30	66.95	62.24	74.29
30 dB	44.54	58.34	53.11	63.70	51.81	64.60	53.51	64.13	57.36	69.74

Then, we further tested the robustness of the forensics methods against AWGN under a QF = 75. Specifically, the forensic results of all the compared methods for inpainted images with signal-to-noise ratios (SNR) of 50 dB, 40 dB, and 30 dB when the QF was equal to 75 are shown in Table 3. Robustness was further tested under the influence of AWGN with signal-to-noise ratios (SNRs) of 50 dB, 40 dB, and 30 dB. It can be seen that when the SNR is 50 dB or 40 dB, the forensic performance of each method is basically the same as that without AWGN. Even if the SNR decreases to 30 dB, the IoU and F1 scores of different forensic methods only appear to experience slight attenuation. In addition, the proposed method not only consistently maintained the best performance compared to all the methods in all the above cases but it was also almost unaffected by AWGN.

6.5. Generalization Performance Evaluation of Networks

We further compared the forensic performance of all forensic methods with respect to the inpainting dataset produced using other deep inpainting approaches [5,7,8] to verify the generalization of the proposed MPIF-Net. The experimental results of all the methods when applied to the three datasets are presented in Tables 4–6 and the best results are indicated in bold.

Table 4. The forensic performance of different methods in relation to the deep inpainting approach [5].

Method	TPR^+	FPR^-	$F1^+$	IoU^+
Ref. [17]	88.95	0.23	90.93	84.53
Ref. [36]	94.88	0.17	94.83	90.73
Ref. [22]	94.93	0.08	95.52	92.42
Ref. [26]	95.18	0.11	95.99	92.53
MPIF-Net	**97.58**	**0.04**	**98.22**	**96.59**

Table 5. The forensic performance of different methods in relation to the deep inpainting approach [7].

Method	TPR^+	FPR^-	$F1^+$	IoU^+
Ref. [17]	86.53	0.69	83.78	74.64
Ref. [36]	82.69	0.34	84.13	76.45
Ref. [22]	71.52	0.29	75.12	66.61
Ref. [26]	86.70	**0.27**	88.70	81.46
MPIF-Net	**90.08**	0.34	**89.89**	**82.80**

Table 6. The forensic performance of different methods in relation to the deep inpainting approach [8].

Method	TPR^+	FPR^-	$F1^+$	IoU^+
Ref. [17]	81.53	0.36	83.12	75.40
Ref. [36]	89.98	0.23	90.26	85.39
Ref. [22]	96.23	0.13	94.70	90.77
Ref. [26]	90.32	0.20	91.60	86.28
MPIF-Net	**98.18**	**0.03**	**98.57**	**97.28**

Regarding the inpainting approach [5], the forensics method presented in [26] achieved the best performance from among the four methods, presenting values of 95.99%, 92.53%, 95.18%, and 0.11% for the F1 score, IoU, TPR, and FPR, respectively. The proposed MPIF-Net's performance was approximately 2.2%, 4.1%, and 2.4% higher than that of the first three metrics of the method presented in [26], and MPIF-Net also maintained the lowest false alarm rate, i.e., FPR. Additionally, compared to the remaining two inpainting approaches, i.e., the methods presented in [7,8], the proposed MPIF-Net still presents almost optimal forensics performance, except for its slightly higher FPR compared to the method presented in [26] when conducting forensics for the inpainting approach [7]. It is evident that MPIF-Net offers a preferable degree of generalization and can be applied to detect multiple inpainting approaches.

6.6. Ablation Studies

To investigate the effects of the FCBAM and BGM on MPIF-Net, an ablation study was conducted on a dataset produced via deep inpainting [6]. Firstly, we designed the following five variants of MPIF-Net to test the influence of FCBAM, as follows:

(1) MPN: A network with three parallel paths was set as the basic model (a multi-path network (MPN)), where the MPN does not include a BGM, and the FCBAM in the AARDM is removed;
(2) MPN-C: This network was established by introducing a CBAM based on an MPN to replace the FCBAM in the original AARDM;
(3) MPN-F4: In this case, the MPN extracts deep features through an AARDM, which employs the four highest DCT components in frequency domain channel attention to fuse feature maps;
(4) MPN-F16: This network has settings similar to those of MPN-F8, except it uses the 16 highest DCT components;
(5) MPN-F32: This network utilizes more DCT components (32) in frequency domain channel attention to improve the representation of deep features extracted via AARDM.

The five network models given above apply the same loss function and training method as MPIF-Net during training.

The results of all the variants are listed in Table 7. It can be seen that compared to the MPN, MPN-C achieved 0.28% and 0.13% improvements in IoU and F1 scores, respectively, through the application of a CBAM. Obviously, by setting the attention mechanisms in the AARDM, the representation of features can be enhanced, thereby improving network performance, but the degree of improvement is limited. If an FCBAM is adopted and the highest DCT component is gradually increased, the forensic performance begins to significantly improve. When the highest DCT component reaches 16, the variant achieves the best

forensic performance, with IoU and F1 scores increased by 1.31% and 0.8%, respectively, compared to the MPN. It is worth noting that this performance cannot be further improved by continuously increasing the highest DCT component, as shown in the results regarding MPN-F32. This may be because the effect of higher-frequency information on improving attention mechanisms is not obvious or possibly even redundant. In summary, an FCBAM is beneficial in terms of improving model performance and can produce the best results among all the variants when using 16 highest DCT components.

Table 7. Average IoU (%) and F1 (%) scores of different variants related to FCBAM of MPIF-Net regarding the deep inpainting approach [6].

Metric	MPN	MPN-C	MPN-F4	MPN-F16	MPN-F32
IoU^+	89.92	90.20	90.78	91.23	90.83
$F1^+$	94.45	94.58	94.95	95.25	94.97

For the boundary guidance module (BGM), several variants of MPIF-Net were designed to verify the effect of this module; these variants are as follows:

(1) MPN-B: This variant was derived by introducing a BGM into the base model MPN;
(2) MPN-F16-B (full model): This network was implemented by applying a BGM to variant MPN-F16, which is the proposed full model.

The results for all the variants are listed in Table 8. Taking the IoU metric as an example, when the BGM module was added to the base network MPN, the prediction result of the MPN-B model was 1.24% higher than that of the MPN. Similarly, when we added a BGM to MPNF-16, i.e., MPIF-Net (MPNF-16-B), the full model exhibited increase that was approximately 1% higher than that of MPNF-16 and was even higher than that of the MPN by over 2%. Therefore, it can be concluded that the BGM is effective for the current task.

Table 8. Average IoU (%) and F1 (%) scores of different variants related to BGM of MPIF-Net regarding the deep inpainting approach [6].

Metric	MPN	MPNB	MPNF-16	MPNF-16-B
IoU^+	89.92	91.16	91.23	92.17
$F1^+$	94.45	94.97	95.25	95.73

7. Conclusions and Future Work

In this paper, we propose a novel deep network for inpainting forensics called MPIF-Net. The proposed MPIF-Net consists of three feature learning paths with the same structure in parallel and promotes feature reuse through fusion between shallow and deep features of different paths. After a description of the network was provided, an attention-aware residual dense module was designed using an improved attention module to enhance the representation ability of deep features; it was employed to efficiently extract deep features in the encoder of each path. Finally, we developed a boundary guidance module, which encourages the model to pay close attention to the boundaries of inpainted regions and reduces boundary distortion of localization results.

MPIF-Net has been extensively tested with regard to multiple inpainting datasets and compared with many state-of-the-art methods. Both the qualitative and quantitative results demonstrate that the proposed network can accurately locate inpainted regions with various sizes and shapes. In addition, the proposed method exhibits excellent generalization for different inpainting approaches and is robust against common post-processing operations, such as JPEG compression and AWGN.

At present, the robustness and generalization of the proposed forensics network still need to be further improved. In future work, we expect that the forensics network

could achieve superior robustness and acquire the ability to detect unknown inpainting technologies by learning directly from inpainted images without post-processing.

Author Contributions: Conceptualization, H.W. and X.Z.; methodology, H.W. and X.Z.; software, H.W. and H.S.; validation, H.W. and T.Q.; formal analysis, H.W. and H.S.; investigation, H.W. and H.S.; resources, H.S. and T.Q.; data curation, H.W. and H.S.; writing—original draft preparation, H.W. and H.S.; writing—review and editing, X.Z.; supervision, X.Z. and Y.C.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China under Grant 61972282 and by the Opening Project of State Key Laboratory of Digital Publishing Technology under Grant Cndplab-2019-Z001.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–931. [CrossRef]
2. Criminisi, A.; Perez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef]
3. Li, Z.; He, H.; Tai, H.-M.; Yin, Z.; Chen, F. Color-direction patch-sparsity-based image inpainting using multidirection features. *IEEE Trans. Image Process.* **2015**, *24*, 1138–1152. [CrossRef]
4. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Internationl Conference on Computer Graphics and Interactive Techniques Conference, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
5. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 107. [CrossRef]
6. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514.
7. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1486–1494.
8. Liu, H.; Jiang, B.; Song, Y.; Huang, W.; Yang, C. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 725–741.
9. Wu, Q.; Sun, S.; Zhu, W.; Li, G.-H.; Tu, D. Detection of digital doctoring in exemplar-based inpainted images. In Proceedings of the IEEE International Conference Machine Learning and Cybernetics (ICMLC), Kunming, China, 12–15 July 2008; Volume 3, pp. 1222–1226.
10. Bacchuwar, K.S.; Ramakrishnan, K.R. A jump patch-block match algorithm for multiple forgery detection. In Proceedings of the IEEE International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), Kottayam, India, 22–23 March 2013; pp. 723–728.
11. Chang, I.; Yu, J.; Chang, C. A forgery detection algorithm for exemplar-based inpainting images using multi-region relation. *Image Vis. Comput.* **2013**, *31*, 57–71. [CrossRef]
12. Liang, Z.; Yang, G.; Ding, X.; Li, L. An Efficient Forgery Detection Algorithm for Object Removal by Exemplar-Based Image Inpainting. *J. Vis. Commun. Image Represent.* **2015**, *30*, 75–85. [CrossRef]
13. Jin, X.; Su, Y.; Zou, L.; Wang, Y.; Jing, P.; Wang, Z. Sparsity-Based Image Inpainting Detection via Canonical Correlation Analysis with Low-Rank Constraints. *IEEE Access* **2018**, *6*, 49967–49978. [CrossRef]
14. Zhang, D.; Liang, Z.; Yang, G.; Li, Q.; Li, L.; Sun, X. A robust forgery detection algorithm for object removal by exemplar-based image inpainting. *Multimed. Tools Appl.* **2018**, *77*, 11823–11842. [CrossRef]
15. Li, H.; Luo, W.; Huang, J. Localization of diffusion-based inpainting in digital images. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 3050–3064. [CrossRef]
16. Zhang, Y.; Liu, T.; Cattani, C.; Cui, Q.; Liu, S. Diffusion-based image inpainting forensics via weighted least squares filtering enhancement. *Multimed. Tools Appl.* **2021**, *80*, 30725–30739. [CrossRef]
17. Zhu, X.; Qian, Y.; Zhao, X.; Sun, B.; Sun, Y. A deep learning approach to patch-based image inpainting forensics. *Signal Process. Image Commun.* **2018**, *67*, 90–99. [CrossRef]
18. Zhu, X.; Qian, Y.; Sun, B.; Ren, C.; Sun, Y.; Yao, S. Image inpainting forensics algorithm based on deep neural network. *Acta Opt. Sin.* **2018**, *38*, 1110005-1–1110005-9.

19. Lu, M.; Liu, S. A detection approach using LSTM-CNN for object removal caused by exemplar-based image inpainting. *Electronics* **2020**, *9*, 858. [CrossRef]
20. Wang, X.; Wang, H.; Niu, S. An intelligent forensics approach for detecting patch-based image inpainting. *Math. Probl. Eng.* **2020**, *2020*, 8892989. [CrossRef]
21. Wang, X.; Niu, S.; Wang, H. Image inpainting detection based on multi-task deep learning network. *IETE Tech. Rev.* **2021**, *38*, 149–157. [CrossRef]
22. Li, H.; Huang, J. Localization of deep inpainting using high-pass fully convolutional network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8301–8310.
23. Zhang, Y.; Ding, F.; Kwong, S. Feature pyramid network for diffusion-based image inpainting detection. *Inf. Sci.* **2021**, *572*, 29–42. [CrossRef]
24. Liu, X.; Liu, Y.; Chen, J.; Liu, X. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Trans. Circuits Syst.* **2021**, *32*, 7505–7517. [CrossRef]
25. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
26. Wu, H.; Zhou, J. IID-Net: Image inpainting detection network via neural architecture search and attention. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1172–1185. [CrossRef]
27. Bender, G.; Kindermans, P.-J.; Zoph, B.; Vasudevan, V.; Le, Q. Understanding and simplifying one-shot architecture search. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 550–559.
28. Wu, Y.; AbdAlmageed, W.; Natarajan, P. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9543–9552.
29. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
30. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
31. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 630–645.
33. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
34. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency channel attention networks. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 783–792.
35. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1452–1464. [CrossRef] [PubMed]
36. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.