applied sciences

MDPI

Article

# Robust Image Inpainting Forensics by Using an Attention-Based Feature Pyramid Network

**Zhuoran Chen, Yujin Zhang *, Yongqi Wang, Jin Tian and Fei Wu**

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; jingyou90719@163.com (Z.C.)
* Correspondence: yjzhang@sues.edu.cn

**Abstract:** Deep learning has injected a new sense of vitality into the field of image inpainting, allowing for the creation of more realistic inpainted images that are difficult to distinguish from the original ones. However, this also means that the malicious use of image inpainting technology to tamper with images could lead to more serious consequences. In this paper, we use an attention-based feature pyramid network (AFPN) to locate the inpainting traces left by deep learning. AFPN employs a feature pyramid to extract low- and high-level features of inpainted images. It further utilizes a multi-scale convolution attention (MSCA) module to optimize the high-level feature maps. The optimized high-level feature map is then fused with the low-level feature map to detect inpainted regions. Additionally, we introduce a fusion loss function to improve the training effectiveness. The experimental results show that AFPN exhibits remarkable precision in deep inpainting forensics and effectively resists JPEG compression and additive noise attacks.

**Keywords:** digital forensics; image inpainting; tampering detection; feature pyramid network; multi-scale convolution attention; deep learning

check for updates

**Citation:** Chen, Z.; Zhang, Y.; Wang, Y.; Tian, J.; Wu, F. Robust Image Inpainting Forensics by Using an Attention-Based Feature Pyramid Network. *Appl. Sci.* **2023**, *13*, 9196. https://doi.org/10.3390/app13169196

Academic Editors: Jiang Zhong, Ying Xie, Weitong Chen and Xue Li

Received: 30 June 2023
Revised: 6 August 2023
Accepted: 9 August 2023
Published: 12 August 2023



## 1. Introduction

With the widespread use of electronic devices and the ubiquitous nature of the internet, accessing information has become increasingly effortless. Digital images, serving as the primary medium for conveying information, are readily available. However, it is important to note that the advancement of image editing software and technology is occurring at an accelerated pace, leading to a decrease in the cost of image tampering. As a result, ensuring the credibility of images has become increasingly challenging [1]. Forged images can appear on social media, in news reports, and even in court, making image forensics a natural focus for researchers. Currently, researchers have developed mature detection methods for common image tampering techniques, such as resampling [2,3], splicing [4], and copy move [5].

Image inpainting is a technique utilized for the purpose of editing images, which can effectively repair damaged or missing regions of images based on the known contents of the original images. It is difficult to distinguish forged images generated by image inpainting. The conventional techniques for image inpainting can mainly be classified into two primary categories, namely, diffusion-based inpainting methods [6,7] and patch-based inpainting methods [8,9]. Constrained by computational expenses, conventional approaches are only suitable for scenarios where the semantic content of the missing region is uncomplicated and small. To achieve a proficient inpainting result in intricate scenarios, an increasing number of researchers are attempting to use numerous deep learning-based image inpainting techniques, such as convolutional neural network (CNN)-based methods [10,11], generative adversarial network (GAN)-based methods [12–15], and transformer-based methods [16]. These methods can obtain more realistic inpainted images and even create new semantic information through large-scale training [17], increasing the challenge of image inpainting forensics.

The conventional inpainting forensics techniques [18,19] primarily depend on the similarity between image blocks to locate inpainting traces. Unfortunately, these methods exhibit significant limitations, such as high computational costs, low accuracy, and limited generalizability. Currently, research on deep learning-based image inpainting forensics remains in the developmental and exploratory phase [20–25], while these deep learning-based techniques exhibit superior performance compared to conventional methods, the majority of them pay attention to enhancing the feature extraction block, with suboptimal optimization of the extracted features, resulting in limited accuracy of inpainting forensics. It is imperative to reiterate that the majority of pixels within the inpainted region are derived from known portions of the image. Therefore, optimizing the utilization of local-to-global contextual information of the inpainting image is crucial in improving the results of inpainting forensics.

Based on the above considerations, we establish an end-to-end image inpainting forensics network, which uses the attention-based feature pyramid network (AFPN) to locate inpainting traces. AFPN was widely used in object detection [26–28] and image segmentation [29,30], but few people use AFPN in the field of image inpainting forensics. By using multi-scale convolution attention [31] (MSCA) after extracting features, the employed AFPN makes the most of low- and high-level features from inpainted images, which are particularly important in inpainting forensics. In the field of image processing, low-level features refer to basic visual characteristics that can be extracted directly from the pixel values of an image, such as color, texture, edges, and corners. The significance of low-level features lies in their ability to capture basic visual cues that are essential for image inpainting detection. High-level features, on the other hand, represent more abstract and semantic aspects of an image, capturing complex visual patterns, structures, and relationships. High-level features can include object shapes, object categories, spatial arrangements, or even more sophisticated attributes, enabling a deeper understanding of the visual content of an image. By combining low-level and high-level features, computer vision systems can achieve a more comprehensive understanding of images.

In order to observe the low-level feature, high-level feature, and the key differences between them more intuitively, we show an example of applying AFPN in Figure 1.

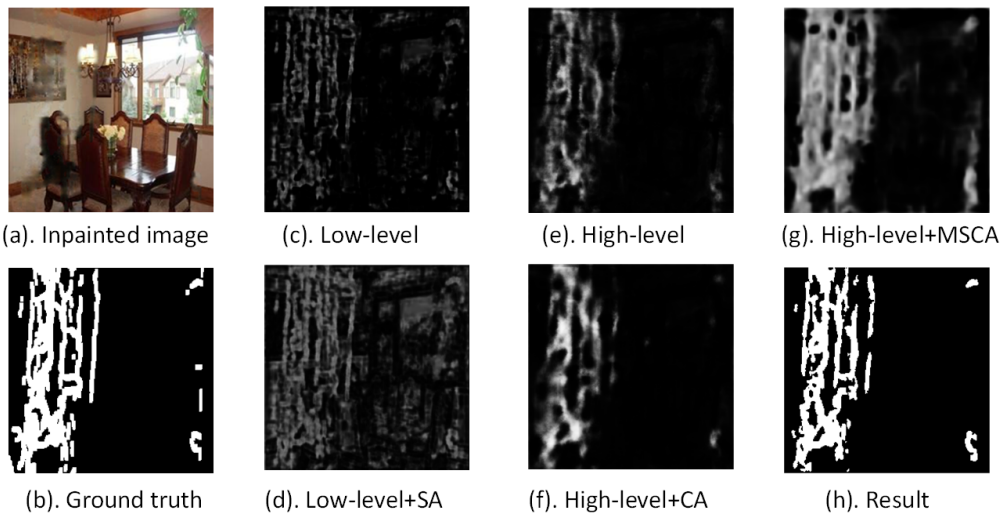


**Figure 1.** An example of applying AFPN.

We obtain Figure 1c by adding the low-level feature map to itself and then add spatial attention (SA) to obtain Figure 1d. Figure 1e–g can be obtained in a similar way. The low-level feature map of the inpainting network exhibits a high degree of similarity to the ground truth, as depicted in Figure 1b,c. Conversely, the high-level feature map in Figure 1e appears to be deficient in crucial inpainting traces, necessitating the application of an effective module to optimize it. Initially, we attempt to optimize the high-level

features through the utilization of channel attention (CA). However, it was observed that despite the increased visibility of the previous inpainting traces, a significant number of inpainting traces remained undetected, as depicted in Figure 1f. Subsequently, considering the peculiarity of the inpainted image, MSCA was employed to extract multi-scale contextual information ranging from local to global information in the feature map. The results in Figure 1g indicate that MSCA is an effective module for enhancing the existing inpainting traces and detecting previously indiscernible inpainting traces. Additionally, we also attempt to employ SA to handle low-level features, yet it is observed that the manipulation adds noise to the low-level features, as depicted in Figure 1d. Following different operations, the high- and low-level features exhibit a complementary relationship, and fusing them by direct addition can result in an inpainting trace-detection map that closely approximates the ground truth, as depicted in Figure 1b,h. In summary, AFPN effectively achieves a high detection accuracy by appropriately processing the low- and high-level features.

Our major contributions can be summarized as follows:

1. We use a forensic network to detect traces left by deep learning-based inpainting methods. The network employs a feature pyramid to extract multi-scale inpainting features. To fully utilize multi-scale feature information, we employ MSCA to optimize high-level features and fuse the optimized high- and low-level features for inpainting forensics. The efficacy of the attention module and feature fusion module is verified through ablation analysis.
2. We design a fusion loss function to assess the quality of not only the fused feature maps but also the high-level feature maps. Experimental results demonstrate that the fused loss function can optimize the training process and enhance the performance of our network.
3. To indicate the generalization performance of our network, we employ six state-of-the-art deep learning-based image inpainting methods to set up a diverse inpainting test dataset. Extensive experiments show that the employed AFPN can achieve good detection performance across diverse inpainting test datasets. Furthermore, we assess the robustness of the proposed methods on JPEG compression and additive noise attacks.

The rest of this paper is organized as follows. Section 2 summarizes the related work on inpainting forensics methods and attention mechanisms. Section 3 presents our network. The experimental results are presented in Section 4, while Section 5 concludes the paper.

## 2. Related Works

### 2.1. Inpainting Forensics Methods

Traditional methods for image inpainting forensics usually depend on calculating the similarity between image blocks to identify the location of the inpainted region. Wu et al. [32] proposed a blind detection approach that relies on zero-connectivity and fuzzy membership. Similarly, Lin et al. [33] employed quantization table estimation to evaluate the incongruity among images for detecting forged images. Liang et al. [34] provided empirical evidence, supporting the notion that traditional methods of image inpainting and forensics are fundamentally similar, which presented a proficient algorithm for detecting forgeries by integrating central pixel mapping, maximum zero-connectivity component labeling, and fragment splicing detection. However, these methods that depend on the similarity of image blocks are limited by some challenging issues. Firstly, the search for highly similar image blocks necessitates the computation of nearly every block in the image, resulting in a significant drain on computational resources. Furthermore, the computational cost escalates rapidly with the increasing image size. Secondly, the high degree of similarity among original image blocks, such as those depicting oceans and deserts, results in a high false alarm rate for traditional methods. Finally, the similarity between image blocks is easily affected by common image post-processing operations, reducing the

robustness of traditional methods. Consequently, traditional inpainting forensics methods have poorer performance.

To address the aforementioned limitations and improve the detection performance, researchers have used deep learning-based image inpainting forensics methods. Li et al. [20] designed HP-FCN, which incorporated a high-pass pre-filtering module prior to the residual network to mitigate the interference of image content and facilitate the location of inpainting traces. Wu et al. [21] proposed MT-Net, a more versatile tampering location network that extracted tampering traces from the image and subsequently located anomalous regions by assessing the disparities between local features and their reference features. Wu et al. [22] proposed the IID-Net, which utilized the neural architecture search (NAS) algorithm to automatically design feature extraction blocks. Zhang et al. [23] improved upon the U-Net architecture by integrating it with feature pyramid networks (FPNs), resulting in a method that effectively detected diffusion-based inpainting traces. Zhu et al. [24] built GLFFNet, which incorporated the Swin Transformer and CNN to extract global and local features of inpainted images. Dong et al. [25] built MVSS-Net, which uses multi-view feature learning to jointly exploit tampering boundary artifacts and the noise view of the input. In contrast to the aforementioned deep learning-based methods, our AFPN focuses on the optimization of extracted features to enable the network to effectively acquire and utilize local-to-global contextual information from the inpainted image.

### 2.2. Attention Mechanisms

The utilization of the attention mechanism in neural networks enables the allocation of computing resources toward tasks of greater significance, directing the network's focus toward crucial components, and ultimately improving network performance. Currently, attention mechanisms have been effectively implemented in a diverse range of tasks, including machine translation [35], saliency detection [36], semantic segmentation [31], anomaly detection [37], object recognition [38], and image captioning [39].

Attention mechanism has been shown to significantly improve the efficacy of image inpainting networks. For instance, Yu et al. [40] employed contextual attention to acquire feature information from known image blocks, thereby enabling the generation of a more realistic inpainted image. Similarly, Wu et al. [41] utilized SA to enhance the semantic consistency between the inpainting area and the original area, as well as within the inpainting area. Since most image inpainting methods use information from the original areas to repair damaged areas, there is a strong correlation between the inpainted areas and the original areas. The attention mechanism enables the network to pay attention to this correlation, improving the performance of image inpainting forensics networks.

### 2.3. AFPN

The feature pyramid network (FPN) is a classic network for realizing multi-scale representation. In order to achieve better results, researchers improve FPN by introducing the attention mechanism and establishing one AFPN after another. Liu et al. [26] proposed an AFPN, which not only facilitates better integration between high-level and low-level feature maps but also enhances the accurate semantic information from low-level features. Wu et al. [27] performed two types of attention mechanisms on the output of the feature enhancement module, modeling the semantic interdependencies in both spatial and channel dimensions, respectively. Jiao et al. [28] devised an AFPN by introducing a learnable fusion factor, which controls the feature information conveyed from deep layers to shallow layers. Hu et al. [29] proposed an attention aggregation-based feature pyramid network to improve multi-scale feature learning through attention-guided feature aggregation. Sun et al. [30] proposed a global–local channel spatial attention module, aimed at capturing global contextual information and image segmentation.

The above methods achieved good results in the fields of object detection and image segmentation, but they are of little help to the field of image inpainting forensics. As depicted in Figure 1, the general attention mechanism is not beneficial to detecting in-

painting traces. However, the MSCA used in our AFPN can effectively enhance the useful information in the high-level feature map, efficiently completing the task of detecting the inpainted area. In addition, the structure of MSCA is light, which has little influence on the processing speed.

## 3. Methods

In this paper, we propose a novel inpainting forensics method, containing a context-aware pyramid feature extraction (CPFE) [36] module and an MSCA module to capture context-aware multi-scale multi-receptive-field high-level features to enhance inpainting traces. Additionally, our method contains one fusion loss function to guide the network to learn valid features for inpainting forensics. The overall architecture is illustrated in Figure 2.

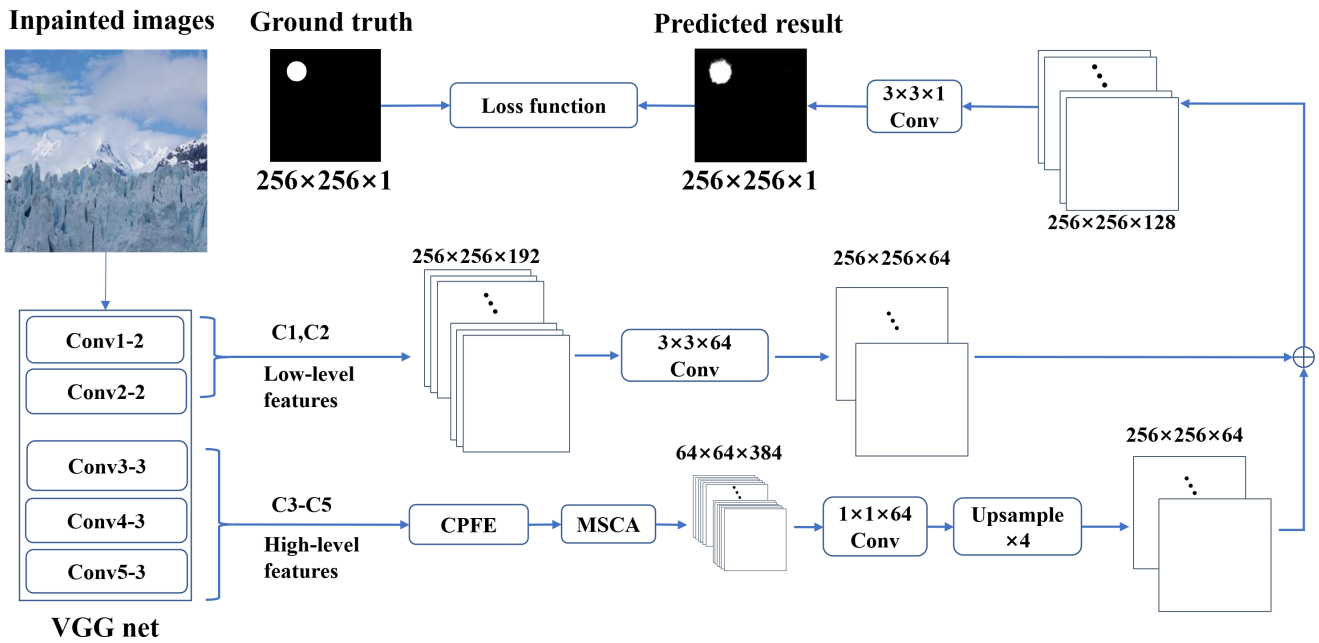


**Figure 2.** The overall architecture of our method.

### 3.1. Multi-Scale Feature Extraction

We take conv1-2, conv2-2, conv3-3, conv4-3, and conv5-3 of VGGNet [42] to extract multi-scale features $\{C_i\}_{i=1}^{5}$ from an input inpainted image. The low-level feature maps are obtained by rolling up $C_1$ and $C_2$, and the high-level feature maps are obtained by rolling up $C_3$, $C_4$, and $C_5$. To extract basic advanced features, we utilize the CPFE module, capturing contextual information at a constant scale, shape, and position. Specifically, the CPFE module employs atrous convolution with dilation rates of 3, 5, and 7 to capture multi-scale contextual information, and then combines the feature maps of different convolution layers with a $1 \times 1$ dimension reduction feature across channels to obtain three different scale feature maps, as illustrated in Figure 3. After that, the CPFE upsamples the two smaller ones to the largest one, making the concatenation possible. Finally, we obtain $64 \times 64 \times 384$ feature maps, which are optimized and restored by CPFE.

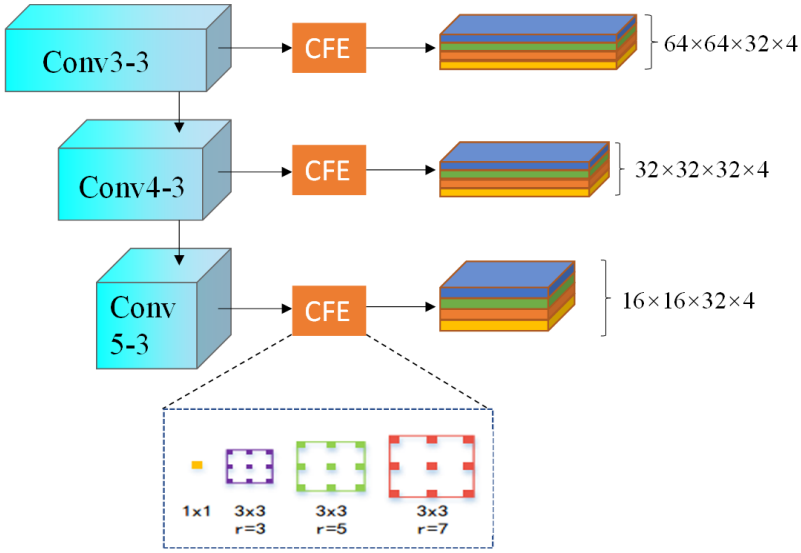


**Figure 3.** Detailed structure of the context-aware pyramid feature extraction (CPFE).

### 3.2. Attention Mechanism

We utilize a powerful attention mechanism called MSCA, which is illustrated in Figure 4.

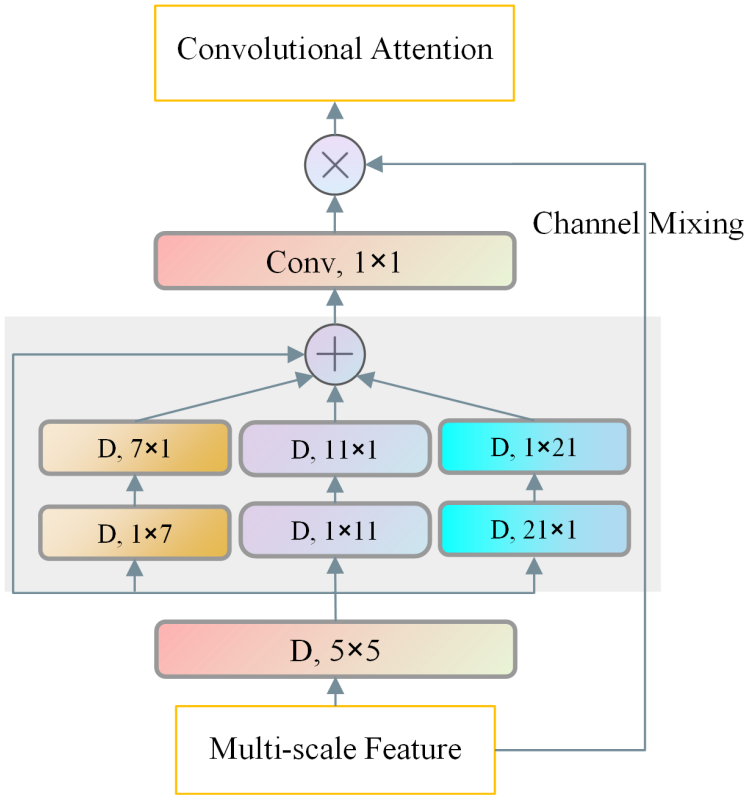


**Figure 4.** Illustration of the multi-scale convolution attention (MSCA).

We utilize the CPFE module to obtain high-level features of multi-scale and multi-receiving fields. However, the high-level feature map in Figure 1e requires further appropriate optimization. The experimental results demonstrate that the MSCA effectively enhances the inpainting traces for high-level feature maps, see Figure 1g. Furthermore, the low-level feature map keeps the original state, as improper handling may result in the magnification of noisy features, including the contour and texture, as depicted in Figure 1d, negatively influencing the inpainting forensics process.

The MSCA module consists of three parts: the depth-wise convolution of aggregating local information, the multi-branch depth-wise strip convolution of capturing the multi-scale context, and the 1 × 1 convolution of simulating the relationship between different channels. The output of the 1 × 1 convolution is directly used as the attention weight to reweigh the input of the MSCA.

The mathematical expression of the MSCA can be written as follows:

$$Att = Conv_{(1\times1)}(\sum_{i=0}^{3} Scale_i(DWConv(F))), \tag{1}$$

$$Out = Att \otimes F. \tag{2}$$

where $F$ represents the optimized feature maps, and $Att$ and $Out$ represent the attention map and output, respectively. The $\otimes$ operation involves matrix multiplication on a pixel-by-pixel basis. $DWConv$ refers to the depth-wise convolution, and $scale_i$, where $i \in \{0, 1, 2, 3\}$ represents the $i$th branch in Figure 4. $scale_0$ represents the identity connection.

In order to simulate the standard depth convolution with a large kernel, two depth-wise strip convolutions are employed in each of the three branches, with kernel sizes of 7, 11, and 21, respectively. The selection of depth-wise strip convolution is motivated by two primary factors. Firstly, strip convolution is characterized by its lightweight nature, requiring only a pair of 7 × 1 and 1 × 7 convolutions to simulate the standard 2D convolution with a kernel size of 7 × 7. Secondly, most of the inpainted areas contain some strip objects, such as people and slogans, increasing the difficulty of the inpainting forensics. As such, the implementation of strip convolution as a complementary technique to grid convolution can improve the detection of inpainting traces. In addition, compared with the attention used by AFPN in [26–30], MSCA is more portable, powerful, and suitable in image inpainting forensics.

### 3.3. Loss Function

The task of inpainting forensics can be characterized as a binary classification problem, wherein the primary objective of the forensics network is to categorize every pixel in the input image as either tamper-free or tampered. In most binary classification tasks, the binary cross entropy (BCE) loss function is widely employed as the preferred loss function. The BCE can be mathematically defined as follows:

$$L_B(G, O) = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(G(i,j)\log O(i,j)(1 - G(i,j))\log(1 - O(i,j))), \tag{3}$$

in this equation, $G(i, j)$ and $O(i, j)$ represent the $(i, j)$th pixel point in the ground truth and the output map, respectively. The resolution of the input image, with dimensions $H \times W$, is set to 256 in this paper.

However, in the case of the majority of inpainted images, the ratio of tampered regions is minimal, resulting in a significant disparity between negative samples (un-inpainted areas) and positive samples (inpainted areas). Consequently, if the sole supervision method for training is binary cross-entropy (BCE), the trained model may possess a robust capability to classify negative samples, but it may encounter challenges in accurately classifying positive samples. This, in turn, poses a difficulty for the model to precisely detect the inpainted area. We propose the utilization of the focus loss function [43] as a solution to address the issue of class imbalance. The focus loss function incorporates a modulation factor into the BCE loss function, thereby decreasing the significance of over-classified negative samples in the overall loss. This approach effectively improves the classification performance of positive samples. The focus loss function is mathematically defined as follows:

$$L_F(G, O) = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(\alpha(1 - O(i,j))^{\gamma}G(i,j)\log O(i,j) + (1-\alpha)(O(i,j))^{\gamma}(1 - G(i,j))\log(1 - O(i,j))), \tag{4}$$

where the focus loss employs a weighting factor, denoted by $\alpha \in \{0{\sim}1\}$ to achieve a balanced representation of the positive and negative sample categories. Specifically, $\alpha$ represents the weight assigned to positive samples, while $(1-\alpha)$ represents the weight assigned to negative samples. Given that the inpainted areas utilized in our paper constitute a range of 5–15%, we set $\alpha = 0.9$. Additionally, the focus loss employs a focusing parameter, denoted by $\gamma \in \{0{\sim}5\}$, to effectively address the imbalance of difficult and easy samples. Empirical evidence suggests that the optimal experimental outcome is achieved when $\gamma = 2$.

As the attention mechanism constitutes the core algorithm of this study and there are essential differences between the processing of high-level feature maps and low-level feature maps, the direct impact of the quality of high-level feature maps on network performance necessitates the adoption of focus loss for high-level feature map $H$, called HF loss. The resulting loss function utilized in this paper is obtained by combining the two focus losses:

$$L = \lambda_1 L_F(G, H) + \lambda_2 L_F(G, O), \tag{5}$$

where the hyperparameters $\lambda_i, i \in \{1, 2\}$ indicate the relative significance of the high-level and output feature maps. For the purposes of this study, both hyperparameters are assigned a value of 1 to underscore the paramount importance of the high-level feature map.

## 4. Experiments

Our AFPN was implemented in PyTorch on a single RTX 3090 GPU and trained with stochastic gradient descent (SGD) [44]. We used the VGG-16 model that was initialized from the pre-trained weights of the ImageNet [45] dataset as the foundational model, and the default parameters of Adam [46] as the optimizer. The initial learning rate was set at $10^{-4}$, with a batch size of 12 and with 50 epochs of training. During the training phase, all images were cropped to a size of 256 × 256. For comparisons, we used publicly available implementations of the state-of-the-art methods, such as [20,21,23,25], the F1 score was utilized as the evaluation criterion. Furthermore, we conducted ablation experiments and introduced novel evaluation criteria, including recall, accuracy, and intersection over union (IoU), to comprehensively assess the efficacy of the primary components of the proposed method. Finally, we evaluated the network's robustness to further test its effectiveness.

### *4.1. Training and Testing Datasets*

We employed a training set comprising 24,000 groups of images, wherein each group consisted of an inpainting image and a corresponding ground truth mask image. Specifically, a random selection of 24,000 images from the Places [47] and Dresden [48] datasets was made, and blank regions with an area of 5–15% were generated within these images. Subsequently, the deep learning inpainting method described in [12] was utilized to inpaint these blank regions, resulting in the creation of 24,000 inpainted images.

To demonstrate the universality of the algorithm presented in this paper, a test set comprising six distinct deep learning inpainting methods was utilized. These methods, namely GC [12], CA [40], SA [41], SN [49], RN [50], and EC [51], each consisted of 1000 groups of images. The inpainted area in each group was manually selected to encompass meaningful objects, with the total area of the inpainted region ranging from 0 to 30% of the entire image.

### *4.2. Quantitative Comparisons*

The advantages of AFPN can be effectively demonstrated through comparative experiments. This study employs three state-of-the-art inpainting forensics techniques, namely HP-FCN [20], MT-Net [21], U-FPN [23], and MVSS-Net [25], to detect the inpainted areas generated by GC inpainting methods. HP-FCN is a full convolution network with high precision that is utilized to identify the forged region generated by deep inpainting. The reason why we choose HP-FCN is that it is the first one to use the deep learning method to detect the deep inpainting image. MT-Net leverages the robust learning capability of neural networks to classify anomalous features in input images and exhibits a strong

generalization performance across various conventional operation types, including inpainting operations. MVSS-Net uses multi-view feature learning to jointly exploit tampering boundary artifacts and the noise view of the input. Both MT-Net and MVSS-Net study the variety of possible attacks on the content, devising a generic method. The U-FPN model extends the feature pyramid network approach, leveraging the benefits of network feature extraction to effectively identify and inpainting traces. U-FPN is the first one to use FPN for image inpainting forensics. To ensure impartiality, this study evaluates the performance of three models, including those provided by the networks, and retrains them using the proposed training set. The results of this comparison are presented in Table 1, where a higher F1 score indicates superior performance.

In order to provide a more precise explanation of the network's generalization ability, the gray value in the table was excluded from the average calculation. The data presented in the table indicate that the AFPN, as employed in this paper, outperforms the other three methods across all test sets. The performance of the U-FPN, which utilizes multi-scale feature information akin to AFPN, is deemed barely satisfactory at 76.45%. Conversely, HP-FCN's poor performance of 8.57% on the EC dataset and low average of 52.51% suggest limited efficacy and universality. Notably, the retrained MT-Net model exhibits a lower efficacy at 15.12% compared to the original model's 46.41%, yet it yields excellent results at 92.10% on the GC test set, which suggest that MT-Net's performance is acceptable. MVSS-Net is very similar to MT-Net. The performance of MVSS-Net is gratifying on SN (94.08%) and EC (83.52%), but is very bad on GC(1.86%). Their universality is notably lacking.

**Table 1.** Quantitative comparisons by using the F1 score as an evaluation criterion.

| Models | Retrain | Test Dataset | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|
| | | GC | CA | SN | EC | SA | RN | |
| MT-Net | - [1] | 14.17 | 28.80 | 72.63 | 67.55 | 60.14 | 35.22 | 46.41 |
| MT-Net | GC | 92.10 [2] | 19.02 | 32.78 | 10.62 | 2.38 | 10.80 | 15.12 |
| HP-FCN | - | 0.04 | 0.22 | 0.38 | 0.42 | 0.05 | 1.98 | 0.52 |
| HP-FCN | GC | 76.93 | 35.75 | 81.43 | 8.57 | 55.78 | 56.58 | 52.51 |
| U-FPN | - | 31.12 | 28.60 | 19.26 | 10.41 | 20.74 | 23.55 | 22.28 |
| U-FPN | GC | 80.14 | 70.40 | 70.26 | 72.18 | 87.28 | 82.22 | 76.45 |
| MVSS-Net | - | 1.86 | 76.22 | 94.08 | 83.52 | 67.63 | 77.07 | 66.73 |
| Proposed | GC | 98.91 | **87.03** [3] | **94.69** | **84.21** | **94.55** | **85.53** | **89.20** |

[1] The "-" in the "Retrain" column indicates that the models are officially released without retraining. [2] The gray value means that the inpainting methods used in the test dataset are used in the training, not testing generalization. [3] The highest value is highlighted in black.

### 4.3. Qualitative Comparisons

To facilitate a more intuitive evaluation of the performances of the four image forensics methods, this study opted to visually present the selected images from each test set. Notably, the retraining effect of the MT-Net is comparatively inferior and, thus, the original model parameters were utilized, while the remaining networks employed the model parameters post-retraining. The visualizations of these images are presented in Figure 5.

The visualized content depicted in Figure 5 exhibits a fundamental consistency with the data presented in Table 1. Notably, the MT-Net demonstrates a sub-optimal performance on the CA and GC test sets, with an accuracy rate of 28.80 and 14.17%, respectively. Consequently, the MT-Net fails to accurately obtain the majority of inpainting information on the CA and GC test sets. MVSS-Net is more special. Its performance on GC is a mess. Conversely, the U-FPN exhibits a commendable performance across all test sets, albeit with some writing defects. The simple network architecture of the HP-FCN renders it challenging to achieve optimal results in more complex tasks.

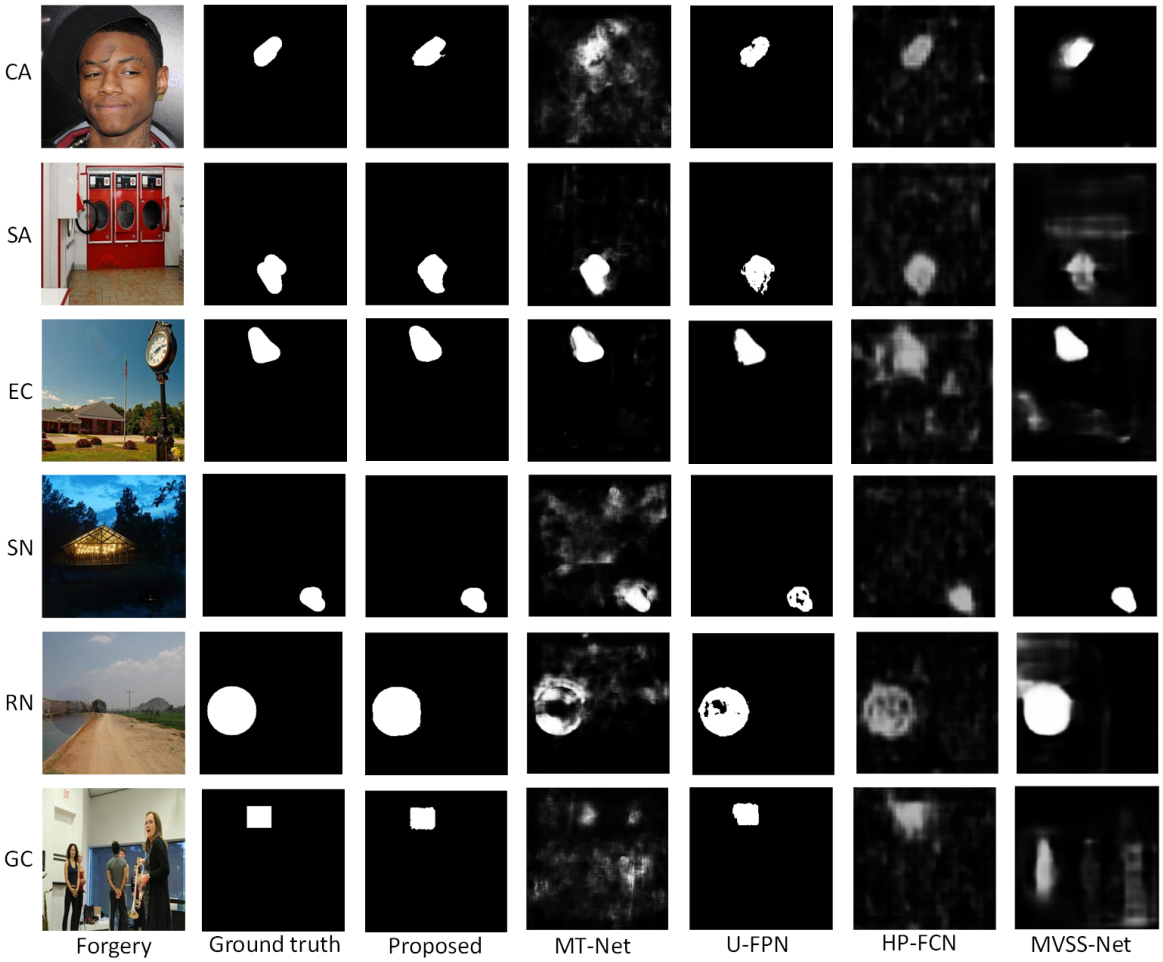


**Figure 5.** Qualitative comparisons for the detection of inpainting forgeries.

### *4.4. Ablation Studies*

This paper conducted three types of ablation experiments to examine the impacts of three innovations (i.e., feature fusion mode, attention module, and loss function) on the final inpainting trace detection outcomes. The results are presented in Table 2.

Regarding the feature fusion approach, three distinct methods for feature fusion were proposed, based on the network structure. These methods include utilizing solely low-level features as the output, solely high-level features as the output, and utilizing low- and high-level fusion features as the output. When solely low-level features are utilized, the accuracy is notably high (99.11%); however, the recall rate is relatively low (91.76%), indicating that part of inpainting traces remain undetected. The utilization of solely advanced features results in an increased recall rate of 94.74%; however, this value is significantly lower than the outcome obtained through feature fusion, which is 98.21%. This indicates that the inpainting traces of images accurately identified by low- and high-level features are restricted, and optimal recall rates can be attained by amalgamating them. Thus, the feature fusion technique employed in this study is highly efficacious, enabling the network to acquire valuable information from both low- and high-level features.

Three distinct approaches for selecting an attention mechanism exist, namely, utilizing only MSCA, exclusively employing CA, or abstaining from an attention mechanism altogether. The recall rate for the latter option is the least favorable at 93.79% when compared to the other two. Upon implementation of CA, the recall rate is increased to 95.94%, albeit at the cost of a decrease in accuracy from 98.95 to 98.05%. Thus, it can be inferred that using the CA in the context of inpainting forensics is limited. The results of the ablation experiments demonstrate that using the MSCA module, as presented in this paper, can

significantly improve the efficacy of inpainting forensics networks by effectively leveraging the contextual information from advanced features at both local and global levels.

**Table 2.** Localizationresults (%) obtained by different variants of the proposed method.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Feature fusion method | Low-Level | | ✓ [1] | | | | | |
| | Low-High | ✓ | | | ✓ | ✓ | ✓ | ✓ |
| | High-Level | | | ✓ | | | | |
| Attention | w/o Att | | | | | ✓ | | |
| | CA | | | | ✓ | | | |
| | MSCA | ✓ | ✓ | ✓ | | | ✓ | ✓ |
| Loss | HF Loss | | | | | | | ✓ |
| | Focal Loss | | | | | | ✓ | |
| | Focal-HF | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Recall | | **98.21** [2] | 91.76 | 94.74 | 95.94 | 93.79 | 97.10 | 96.01 |
| Precision | | **99.62** | 99.11 | 98.42 | 98.05 | 98.95 | 98.25 | 99.04 |
| IoU | | **97.85** | 91.01 | 93.32 | 94.15 | 92.87 | 95.45 | 95.12 |
| F1 | | **98.91** | 95.29 | 96.54 | 96.98 | 96.30 | 97.67 | 97.50 |

[1] ✓ means to adopt this method. [2] The highest value is highlighted in black.

To ascertain the efficacy of the loss function posited in this study, a final ablation experiment was conducted. The outcome of training the network with the loss function $L_F(G, O)$ in Equation (4) (yielding an F1 score of 97.67%) is found to be nearly identical to that of $L_F(G, H)$ (yielding an F1 score of 97.50%). But they are all lower than the results of fusion loss function $L$ (98.91%). This finding serves to reaffirm the criticality of the high-level feature integrated into the network architecture employed in this research. It is noteworthy that altering the loss function has minimal impact on the network performance. Empirical findings indicate that utilizing the fusion loss function, denoted as $L$ in Equation (5), can significantly enhance the network performance.

In order to verify the hyperparameters we used in Equation (5), we perform a concrete study about the setting of the hyperparameter. The results are presented in Table 3.

**Table 3.** Localization results (%) obtained by different loss function hyperparameter settings.

| Hyperparameter Setting | Recall | Precision | IoU | F1 |
|---|---|---|---|---|
| $\lambda_1 = \lambda_2$ | 98.21 | 99.62 | 97.85 | 98.91 |
| $\lambda_1 = 2\lambda_2$ | 97.93 | 96.61 | 95.85 | 97.87 |
| $2\lambda_1 = \lambda_2$ | 94.13 | 99.67 | 96.18 | 98.12 |

From Table 3, we observe that the best results are obtained when the $L_F, (G, O)$, and $L_F(G, H)$ account for the same proportion in the loss function.

In addition, we also study the influence of the pre-training model on the detection results. We use different pre-trained models, i.e., VGGNet [42], ResNet-50 [52], and Swin-T [53] to train AFPN. The results are presented in Table 4.

**Table 4.** Localization results (%) obtained by different pre-trained models.

| Pre-Trained Model | Recall | Precision | IoU | F1 |
|---|---|---|---|---|
| VGG-16 | 98.21 | 99.62 | 97.85 | 98.91 |
| ResNet-50 | 99.20 | 99.56 | 97.83 | 98.84 |
| Swin-T | 99.30 | 99.74 | 98.21 | 99.16 |

From Table 4, we observe that using the Swin-T pre-trained model can further stimulate the potential of our method, but this improvement is limited. Usually, CNN has an

advantage over the transformer in the processing speed because of their different calculation methods. So, we choose VGG16 as our pre-training model.

### 4.5. Robustness Evaluations

The evaluation of the employed AFPN's robustness is conducted, whereby the impact of common image post-processing operations, including noise addition and JPEG compression, on the trace of inpainting is examined, thereby posing challenges for inpainting forensics. The inadequacy of robustness remains a significant drawback of conventional inpainting forensics approaches. The results are presented in Tables 5 and 6.

**Table 5.** Localization results (%) under different JPEG compression quality factors.

| QF | Recall | Precision | IoU | F1 |
|---|---|---|---|---|
| 85 | 94.46 | 98.24 | 92.89 | 96.31 |
| 75 | 92.52 | 98.35 | 91.11 | 95.35 |
| 65 | 80.60 | 99.61 | 80.35 | 89.10 |

**Table 6.** Localization results (%) under Gaussian noise with different standard deviations.

| Std | Recall | Precision | IoU | F1 |
|---|---|---|---|---|
| 0.1 | 92.37 | 98.29 | 90.90 | 95.24 |
| 0.2 | 90.20 | 98.92 | 89.32 | 94.36 |
| 0.3 | 83.70 | 99.32 | 83.22 | 90.84 |

Consequently, this study employed various post-processing techniques of different magnitudes on the test datasets, presenting statistical detection outcomes in Tables 5 and 6. The findings indicate that the overall performance is satisfactory when disturbance intensity is low. The performance remains relatively stable at a JPEG compression quality factor of 85. Conversely, a significant decline in performance is observed as the disturbance intensity increases to 65. This assertion holds when Gaussian noise is introduced, as the stability collapses at a standard deviation of 0.3. Whether the quality factor is 65 or the standard deviation is 0.3, images are significantly degraded, leading to the loss of the original purpose of inpainting forensics.

In order to show the advantages of our network in robustness, we made comparisons. The visualizations of these comparisons are presented in Figures 6 and 7.

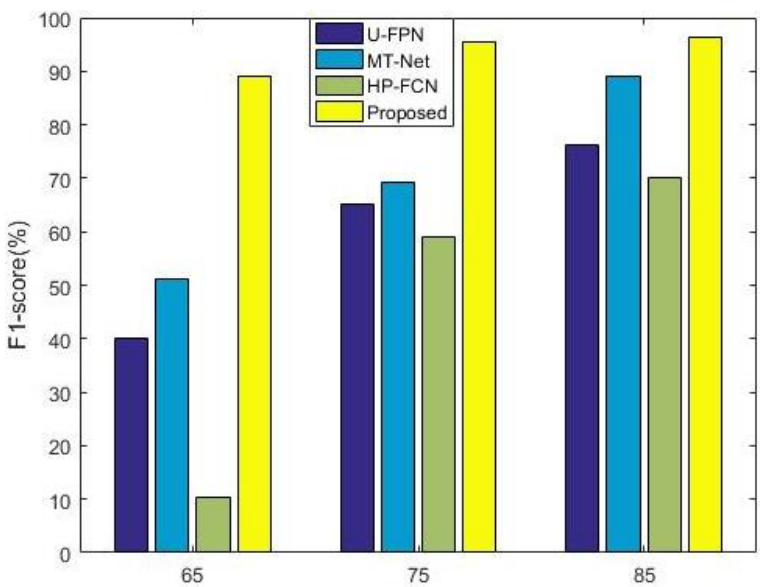


**Figure 6.** Comparisons in the robustness of JPEG compression by using the F1 score as an evaluation criterion.

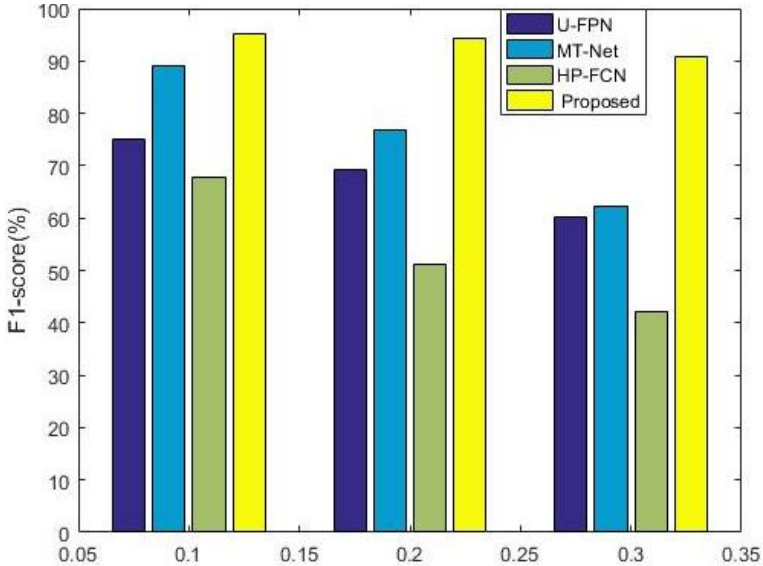


**Figure 7.** Comparisons in the robustness of Gaussian noise by using the F1 score as an evaluation criterion.

In conclusion, the robustness of AFPN is demonstrated.

### *4.6. Limitations*

Our method also has some limitations, as shown in Figure 8.

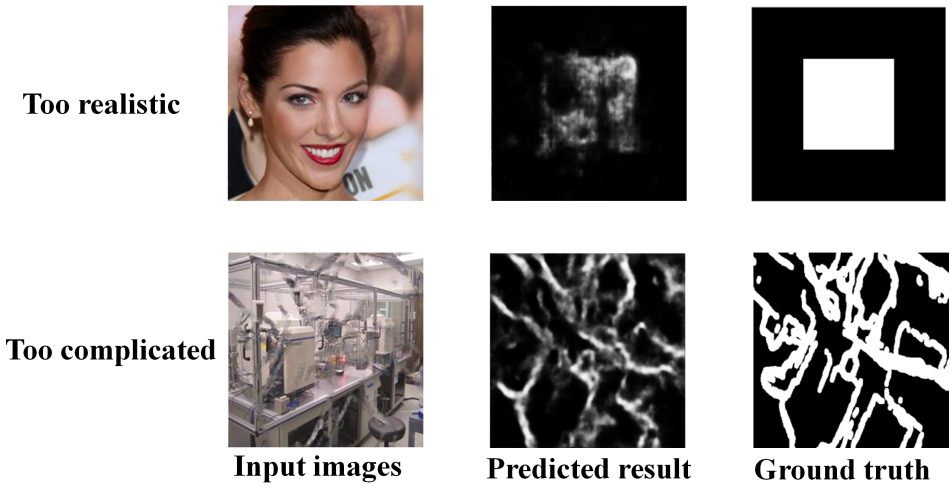


**Figure 8.** Limitation of our method.

The detection effect of the method proposed in this paper is not good for graphs whose repair marks are too complicated and extremely difficult to identify.

## 5. Conclusions

In this paper, we use a deep learning-based inpainting forensics approach called AFPN. AFPN utilizes the feature pyramid network to predict pixel-wise class labels for inpainting manipulation and optimizes high-level feature maps by the MSCA model. For training AFPN, we introduce the fusion loss function, which takes the effect of high-level feature maps into account. By adopting a data-driven approach, AFPN avoids the challenges associated with designing hand-crafted features.

We extensively test AFPN on various images and compare its performance with state-of-the-art inpainting forensics methods. The experimental results demonstrate that AFPN effectively learns manipulation features for deep image inpainting and accurately locates

inpainted regions. In terms of location accuracy, AFPN outperforms representative forensics methods. Additionally, AFPN exhibits superior robustness against typical post-processing operations, such as JPEG compression and additive noise attacks.

**Author Contributions:** Conceptualization, Z.C.; Methodology, Z.C.; Software, Z.C.; Writing—original draft, Z.C.; Writing —review & editing, Y.Z. and Y.W.; Supervision, Y.Z.; Project administration, J.T. and F.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** Industry-University-Research Innovation Fund of the Chinese Ministry of Education, Item Number: 2021ZYB01003; Shanghai Natural Science Foundation Project, Item Number: 17ZR1411900.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Tyagi, S.; Yadav, D. A detailed analysis of image and video forgery detection techniques. *Vis. Comput.* **2023**, *39*, 813–833. [CrossRef]
2. Liang, Y.; Fang, Y.; Luo, S.; Chen, B. Image resampling detection based on convolutional neural network. In Proceedings of the 2019 15th International Conference on Computational Intelligence and Security (CIS), Macao, China, 13–16 December 2019; pp. 257–261.
3. Lamba, M.; Mitra, K. Multi-patch aggregation models for resampling detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2967–2971.
4. Ding, H.; Chen, L.; Tao, Q.; Fu, Z.; Dong, L.; Cui, X. DCU-Net: A dual-channel U-shaped network for image splicing forgery detection. *Neural Comput. Appl.* **2023**, *35*, 5015–5031. [CrossRef] [PubMed]
5. Babu, S.T.; Rao, C.S. Efficient detection of copy-move forgery using polar complex exponential transform and gradient direction pattern. *Multimed. Tools Appl.* **2023**, *82*, 10061–10075. [CrossRef]
6. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th annual conference on Computer Graphics and Interactive Techniques, New York, NY, USA, 23–28 July 2000; pp. 417–424.
7. Esedoglu, S.; Shen, J. Digital inpainting based on the Mumford–Shah–Euler image model. *Eur. J. Appl. Math.* **2002**, *13*, 353–370. [CrossRef]
8. Hays, J.; Efros, A.A. Scene completion using millions of photographs. *ACM Trans. Graph.* **2007**, *26*, 4-es. [CrossRef]
9. Chen, Y.; Zhang, H.; Liu, L.; Tao, J.; Zhang, Q.; Yang, K.; Xia, R.; Xie, J. Research on image inpainting algorithm of improved total variation minimization method. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *14*, 5555–5564. [CrossRef]
10. Chen, Y.; Liu, L.; Phonevilay, V.; Gu, K.; Xia, R.; Xie, J.; Zhang, Q.; Yang, K. Image super-resolution reconstruction based on feature map attention mechanism. *App. Intell.* **2021**, *51*, 4367–4380. [CrossRef]
11. Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; Lempitsky, V. Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 2149–2159.
12. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4471–4480.
13. Zheng, C.; Cham, T.-J.; Cai, J. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1438–1447.
14. Chen, Y.; Liu, L.; Tao, J.; Xia, R.; Zhang, Q.; Yang, K.; Xiong, J.; Chen, X. The improved image inpainting algorithm via encoder and similarity constraint. *Vis. Comput.* **2021**, *37*, 1691–1705. [CrossRef]
15. Zhao, S.; Cui, J.; Sheng, Y.; Dong, Y.; Liang, X.; Chang, E.I.; Xu, Y. Large scale image completion via co-modulated generative adversarial networks. *arXiv* **2021**, arXiv:2103.10428.
16. Wan, Z.; Zhang, J.; Chen, D.; Liao, J. High-fidelity pluralistic image completion with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 4692–4701.
17. Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D.J.; Soricut, R. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18359–18369.
18. 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]
19. Li, H.; Luo, W.; Huang, J. Localization of diffusion-based inpainting in digital images. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 3050–3064. [CrossRef]
20. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8301–8310.
21. 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, Long Beach, CA, USA, 16–20 June 2019; pp. 9543–9552.
22. Wu, H.; Zhou, J. IID-Net: Image inpainting detection network via neural architecture search and attention. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1172–1185. [CrossRef]

23. Zhang, Y.; Ding, F.; Kwong, S.; Zhu, G. Feature pyramid network for diffusion-based image inpainting detection. *Inf. Sci.* **2021**, *572*, 29–42. [CrossRef]
24. Zhu, X.; Lu, J.; Ren, H.; Wang, H.; Sun, B. A transformer–CNN for deep image inpainting forensics. *Image Vis. Comput.* **2022**, 1–15. [CrossRef]
25. Dong, C.; Chen, X.; Hu, R.; Cao, J.; Li, X. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3539–3553. [CrossRef] [PubMed]
26. Liu, Z.; Gong, P.; Wang, J. Attention-Based feature pyramid network for object detection. In Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, Beijing, China, 23–25 October 2019; pp. 117–121.
27. Wu, H.; Dong, B.; Ding, L.; Dong, Y. Attention feature pyramid network for scene text detection. In Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications, Chengdu, China, 9–12 December 2022; pp. 1726–1731.
28. Jiao, L.; Kang, C.; Dong, S.; Chen, P.; Li, G.; Wang, R. An attention-based feature pyramid network for single-stage small object detection. *Multimed. Tools Appl.* **2023**, *82*, 18529–18544. [CrossRef]
29. Hu, M.; Li, Y.; Fang, L.; Wang, S. A2-FPN: Attention aggregation based feature pyramid network for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15343–15352.
30. Sun, Y.; Dai, D.; Zhang, Q.; Wang, Y.; Xu, S.; Lian, C. MSCA-Net: Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recognit.* **2023**, *139*, 109524. [CrossRef]
31. Guo, M.; Lu, C.; Hou, Q.; Liu, Z.; Cheng, M.; Hu, S. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv* **2022**, arXiv:2209.08575.
32. Wu, Q.; Sun, S.-J.; Zhu, W.; Li, G.-H.; Tu, D. Detection of digital doctoring in exemplar-based inpainted images. In Proceedings of the 2008 International Conference on Machine Learning and Cybernetics, Kunming, China, 12–15 July 2008; pp. 1222–1226.
33. Lin, G.; Chang, M.; Chen, Y. A passive-blind forgery detection scheme based on content-adaptive quantization table estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 421–434. [CrossRef]
34. 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]
35. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the International Conference on Machine Learning, Amsterdam, The Netherlands, 7–10 July 2017; pp. 1243–1252.
36. Zhao, T.; Wu, X. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3085–3094.
37. Wang, L.; Song, Z.; Zhang, X.; Wang, C.; Zhang, G.; Zhu, L.; Li, J.; Liu, H. SAT-GCN: Self-attention graph convolutional network-based 3D object detection for autonomous driving. *Knowl. Based Syst.* **2023**, *259*, 110080. [CrossRef]
38. Lei, X.; Xia, Y.; Wang, A.; Jian, X.; Zhong, H.; Sun, L. Mutual information based anomaly detection of monitoring data with attention mechanism and residual learning. *Mech. Syst. Signal Process.* **2023**, *182*, 109607. [CrossRef]
39. Dubey, S.; Olimov, F.; Rafique, M.A.; Kim, J.; Jeon, M. Label-attention transformer with geometrically coherent objects for image captioning. *Inf. Sci.* **2023**, *623*, 812–831. [CrossRef]
40. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514.
41. Wu, H.; Zhou, J.; Li, Y. Deep generative model for image inpainting with local binary pattern learning and spatial attention. *IEEE Trans. Multimed.* **2021**, *24*, 4016–4027. [CrossRef]
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
44. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the COMPSTAT'2010, Paris, France, 22–27 August 2010; pp. 177–186.
45. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
46. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
47. 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]
48. Gloe, T.; Böhme, R. The Dresden Image Database for Benchmarking Digital Image Forensics. *J. Digit. Forensic Pract.* **2010**, *3*, 150–159. [CrossRef]
49. Yan, Z.; Li, X.; Li, M.; Zuo, W.; Shan, S. Shift-net: Image inpainting via deep feature rearrangement. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 1–17.
50. Yu, T.; Guo, Z.; Jin, X.; Wu, S.; Chen, Z.; Li, W.; Zhang, Z.; Liu, S. Region normalization for image inpainting. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12733–12740.
51. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.

52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.