entropy

MDPI

Article

MFAN: Multi-Level Features Attention Network for Fake Certificate Image Detection

Yu Sun [1,2], Rongrong Ni [1,2,*] and Yao Zhao [1,2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China; yu.sun@bjtu.edu.cn (Y.S.); yzhao@bjtu.edu.cn (Y.Z.)

[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China

* Correspondence: rrni@bjtu.edu.cn

Abstract: Up to now, most of the forensics methods have attached more attention to natural content images. To expand the application of image forensics technology, forgery detection for certificate images that can directly represent people's rights and interests is investigated in this paper. Variable tampered region scales and diverse manipulation types are two typical characteristics in fake certificate images. To tackle this task, a novel method called Multi-level Feature Attention Network (MFAN) is proposed. MFAN is built following the encoder–decoder network structure. In order to extract features with rich scale information in the encoder, on the one hand, we employ Atrous Spatial Pyramid Pooling (ASPP) on the final layer of a pre-trained residual network to capture the contextual information at different scales; on the other hand, low-level features are concatenated to ensure the sensibility to small targets. Furthermore, the resulting multi-level features are recalibrated on channels for irrelevant information suppression and enhancing the tampered regions, guiding the MFAN to adapt to diverse manipulation traces. In the decoder module, the attentive feature maps are convoluted and unsampled to effectively generate the prediction mask. Experimental results indicate that the proposed method outperforms some state-of-the-art forensics methods.

Keywords: image forensics; certificate image; multi-level features; feature recalibration

check for updates

Citation: Sun, Y.; Ni, R.; Zhao, Y. MFAN: Multi-Level Features Attention Network for Fake Certificate Image Detection. *Entropy* **2022**, *24*, 118. https://doi.org/10.3390/e24010118

Academic Editors: Luis Javier García Villalba and Vincent A. Cicirello

Received: 13 December 2021
Accepted: 7 January 2022
Published: 13 January 2022

1. Introduction

With the development of computer technology, image editing software is becoming more and more popular, such as Photoshop, CorelDRAW and Fireworks. After simple operations, you can create any images you want. We cannot deny that the popularity of image editing tools has brought convenience to our life, but at the same time, the threshold of tampering has been greatly reduced. Through the Internet and newspaper, a large number of tampered images are used to spread rumors, fabricate fake news and obtain illegal benefits. Therefore, digital image forensics emerges as the times require.

After more than ten years of development, image forensics technology has been widely used in news, justice, criminal investigation and other fields. Techniques used to identify image authenticity can be divided into two groups: active methods [1,2] and passive methods [3,4]. In active methods, the embedded watermark is regarded as an image fingerprint. The authenticity can be confirmed if the retrieved information is consistent with the original one. Passive methods locate the tampered area by analyzing features left by manipulations rather than the extrinsic information of test image. Due to the wider application of passive techniques, they are playing a constructive role in image forensics.

Up to now, many passive forensics techniques have been reported. Among these methods, copy-move [5,6], splicing [7,8], removal [9], enhancement [10], face anti-spoofing [11] and deepfake [12] are hot topics. Copy-move forgery is solved by searching two identical regions in an image. Splicing, removal and enhancement detection depend on distinguishing the abrupt spliced boundary and verifying a unique manipulation that is existed in tampered area. Depth and temporal information play vital roles in face anti-spoofing tasks. Deepfake is a hotspot in recent years. Color cues, fingerprint of GAN and head pose estimation are key properties for deepfake detection.

It is worth noting that during the COVID-19, some people tampered the nucleic acid detection report to illegally obtain the freedom of movement, which brought great hidden danger to public safety. However, the current forensics research is more focused on natural content images. To promote the application of image forensics technology in broader areas, in this paper, we identify the authenticity of certificate images, which can directly represent people's rights and interests. In order to make the dataset closer to those in the real world, the tampered images in our experiments include arbitrary operations, such as splicing, copy-move and object removal. Examples of tampered certificate images are illustrated in Figure 1.

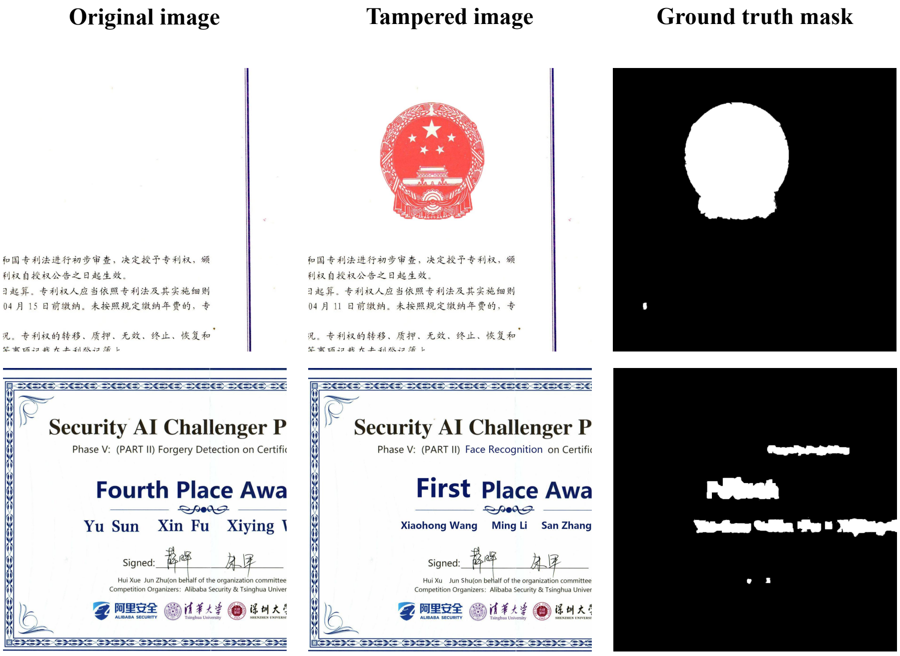

Original image
Tampered image
Ground truth mask
和国专利法进行初步审查，决定授予专利权，颁
利权自授权公告之日起生效。
日起算。专利权人应当依照专利法及其实施细则
04 月 15 日前缴纳。未按照规定缴纳年费的，专
况。专利权的转移、质押、无效、终止、恢复和
04 月 11 日前缴纳。未按照规定缴纳年费的，专
Security AI Challenger P
Phase V: (PART II) Forgery Detection on Certific
Fourth Place Awa
Yu Sun Xin Fu Xiying W
Signed:
Hui Xue Jun Zhu(on behalf of the organization committee
Competition Organizers: Alibaba Security & Tsinghua Univer
Phase V: (PART II) Face Recognition on Certific
First Place Awa
Xiaohong Wang Ming Li San Zhang
Hui Xu Jun Shu(on behalf of the organization committee
阿里安全 ALIBABA SECURITY
清華大学 Tsinghua University
深圳大学 SHENZHEN UNIVERSI

Figure 1. Examples of certificate images. From left to right, the first column is the original image; the second column is the tampered certificate image, and the last column is the ground truth mask.

Obviously, fake certificate images have two remarkable characteristics. The first one is the variable tampered region scales. The tampered region will be very small if only a single word or number is modified, whereas the tampered area of the stamp can be much larger. The second one is the diversity of manipulation types. Each fake certificate image in our experiments contains at least one type of manipulation.

In order to identify the authenticity of certificate images, variable tampered area scales and diverse manipulations are two important issues to be solved. For the first problem, Atrous Spatial Pyramid Pooling (ASPP) [13] is a common technique for multi-scale features extraction in deep learning based methods. A recent work [14] made use of ASPP on the last attention module and fused features by element-wise product. DOA-GAN [15] applied two ASPP operations with different parameters on a concatenated feature to capture two types of information on different scales. In terms of the second problem, each type of manipulation leaves its own unique traces. Copy-move forgery could be distinguished by finding at least two similar objects in the image [15]. The differences of features between host image and spliced region were exploited in splicing detection [16,17]. Object removal detection depends on the similarity among image patches' features [9,18,19]. However, the exiting forensic methods have two drawbacks: (1) The local information would become very weak with the increase of network layers, making the discriminative features especially in small targets difficult to retain. As a result, small tampered region would be hard to discover. (2) Most methods don't work well when identifying image which contains more than one type of manipulation.

In this paper, we address the above issues and propose a multi-level features attention network for fake certificate image detection. We employ a pre-trained residual network to obtain rich features by three steps. Firstly, ASPP module is applied to capture the contextual information at different scales. Secondly, high-level semantic features are concatenated with low-level features to prevent the properties of a small target from losing after pooling. Finally, we recalibrate the weights of feature map channels to make the network more discriminative for variable manipulation traces. Experimental results verify that the proposed method outperforms the state-of-the-art image forensics methods.

Our contributions are summarized in three-fold. (1) To preserve more information and avoid failures in localizing small tampered objects, low-level convolution layers are made use of to fuse the final feature map. (2) We implement recalibration on feature channels to effectively capture traces of different types of manipulations. (3) We propose a novel network called MFAN for fake certificate images detection, which outperforms some state-of-the-art detection methods.

The rest of the paper is organized as follows. Section 2 presents a brief review of forensics technology and attention mechanism. The proposed method for fake certificate image detection is described in Section 3. Section 4 shows our experimental results, and we conclude this paper in Section 5.

2. Related Work

In this section, related works are briefly introduced. Some state-of-the-art methods for image forgery including copy-move, splicing, removal, enhancement and nonspecific manipulation types are illustrated in Section 2.1. Some recent works about attention mechanism are given in Section 2.2.

2.1. Image Forgery Detection and Localization

Copy-move: A common pipeline for copy-move forgery detection algorithms has three essential stages: feature extraction, matching and post-processing. There are two types of traditional features: block-based and keypoint-based features. Block-based methods such as DCT [20], FT [21], PCA [22] and Zernike moments [23] divided image into several overlapping blocks and extracted features for each block. Keypoint-based methods include SIFT [5,24], SURF [25], LBP [26] and DAISY [27] relied on selecting the extreme value in scale space. Recently, CNN based methods have played an important role in copy-move forgery detection. Wu et al. [28] designed a two-branch CNN-based solution called BusterNet to localize source/target regions separately. Islam et al. [15] proposed a dual-order attention model with two ASPP operations to extract contextual features.

Splicing: The most effective tools which splicing detection approaches are based on can be divided into two categories: (1) statistical characteristics; (2) deep features. Color Filter Array (CFA) demosaicking artifacts, Photo Response Non-Uniformity (PRNU) and noise discrepancy are the most widely used statistical characteristics. Ferrara et al. [16] made advantages of CFA model at a local level to obtain the forgery map. In the paper [29], they could get tampering probability map by calculating the correlations of PRNU on nonoverlapped patches. Liu et al. [30] proposed an adaptive-SVD method to estimated local and global noise. Nowadays, deep features have shown their capability in detecting spliced images. By adopting the consistency of EXIF, Huh et al. [17] introduced a self-supervised method to judge whether the input patches come from the same image. Mayer et al. [31] localized the spliced patches by exploring whether they were captured by the same camera model.

Removal: Object removal is a common way of tampering, which can be simply and effectively achieved by image inpainting. Due to the difficulties in feature extraction, the works on inpainting forensics is limited. Considering the similarity among blocks, Chang et al. [18] proposed a two-stage searching algorithm to detect the tampered images containing uniform areas. Liang et al. [19] exploited the central pixel mapping, greatest zero-connectivity component labeling and fragment splicing detection to improve the performance. With the label matrix and designed loss function, an encoder–decoder network was guided to learn the inpainting features [9].

Enhancement: Image enhancement includes many kinds of operations, the general ones are JPEG compression, blurring, contrast enhancement, resampling and additive noise. Choi et al. [32] proposed a CNN-based architecture to learn the statistical information. Wen et al. [33] utilized the significant differences in the image pixel histogram to find traces left by contrast enhancement. Bayar et al. [34] developed a constrained convolutional layer to make the extracted features less affected by the image content. Chen et al. [10] introduced a rotation-invariant CNN for image enhancement forensics. Moreover, they employed isotropic architecture to the network to reduce the required number of parameters.

Nonspecific manipulation types: Most works detect forgeries by the unique characteristic, which is corresponding to a specific manipulation type. This usually leads to mistakes if irrelevant manipulation has been applied to the test image. Some recent papers such as RGB-N [35], ManTra-Net [8] and RRU-Net [36] proposed more general forensics methods that didn't need to specify the type of tampering. RGB-N built a two-stream Faster R-CNN network, which contained RGB stream and noise stream for forensics. RGB stream was employed to capture unnatural tampered boundaries, and the purpose of the noise stream was to find the discriminative features from noise information. The limitation of [35] is that the prediction mask is a rectangular box rather than the pixel level localization. ManTra-Net treated image forensics as a task of local anomaly detection. The backbone architecture of the feature extractor was VGG that was trained by 385 types of manipulations. The local anomaly detection network designed a Z-score feature to measure the difference between a local feature and its references. Paper [36] designed an RRU-Net to solve the gradient degradation problem and differentiate image attributes between the original and tampered regions.

2.2. Attention Mechanism

The basic idea of attention mechanism is making the neural network learn to pay more attention to the important information instead of giving the same weight to all words or regions of the input. The application of the attention mechanism has improved the performance a lot in many tasks, such as machine translation, image caption, image classification and semantic segmentation.

Some recent works employed the attention mechanism into image forgery detection. DOA-GAN [15] introduced a dual-order attention module to localize copy-move forgery. The first-order attention is for location information, and the second-order attention is for patches co-occurrence. Rao et al. [14] proposed a conditional random field-based attention network to generate a series of multi-scale attention maps where the interdependence between pixels are taken into account. To achieve good performance for face forgery detection, PRRNet [37] took advantage of the spatial attention mechanism to learn more competitive features on manipulated regions and the original regions.

3. Proposed Method

3.1. Overview

There are two obvious characteristics in fake certificate images. The first one is the variable tampered region scales. Due to the complex content in the certificate image, the tampered area can be as small as a single letter or as large as a stamp. The second one is that each tampered image contains more than one type of manipulation. Recent studies have shown that applying atrous convolution with different rates can obtain multi-scale information with fewer parameters. In terms of deep learning networks, with the increase of the network layers, the local information becomes weaker, causing the discrimination of features, especially in the small objects, to become weaker and difficult to retain. Taking advantage of low-level features can help to capture extra local information for small objects.

Different types of manipulation leave different tampering traces. Feature map channels calculated by different convolution kernels contain different traces of manipulations, but not all channels contain useful signals. Hence, feature recalibration on channels is beneficial to obtaining more competitive features.

In order to solve the above difficulties, we propose a novel approach following the encoder–decoder network structure called MFAN to detect fake certificate images. The framework of MFAN is illustrated in Figure 2. MFAN is composed of two sub-networks, i.e., the encoder network for feature extraction and the decoder network for localizing forgery regions. In the encoder network, we employ a pre-trained residual network as the backbone architecture. The last layer with ASPP and two low-level layers are resized to the same size and concatenated. Then, the feature recalibration module is for irrelevant information suppression and enhancing the tampered regions, guiding the MFAN to adapt to diverse manipulation traces. At last, the final attentive feature map is fed into decoder network to generate the prediction mask.

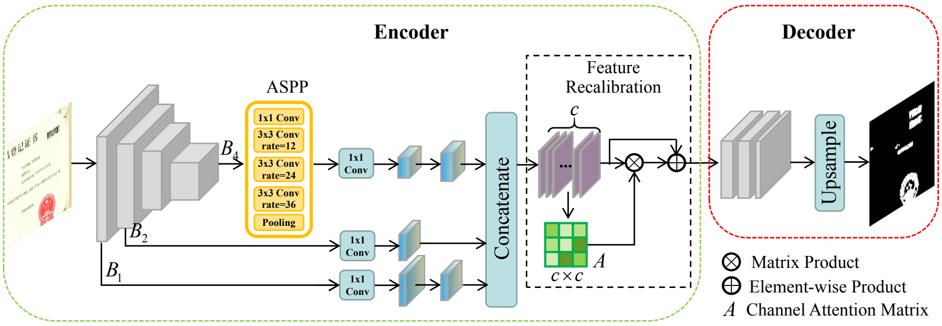


Figure 2. The framework of Multi-level Feature Attention Network (MFAN). There are two sub-network in MFAN: the encoder is used to extract rich features and the decoder is designed to generate binary localization map.

3.2. Encoder

Multi-level features fusion module is designed for forming more powerful features which contain rich information. Given an input image $I \in R^{H\times W\times 3}$, features are extracted by feeding it into a pre-trained ResNet-50. There are four groups of blocks in ResNet-50. B_1, B_2, B_3 and B_4 are feature maps output by four blocks from the first one to the last one. The sizes decrease by ratio 2 for every block from B_1 to B_4. Motivated by [13], in order to capture multi-scale information, ASPP is applied on the final layer B_4 of the backbone. The ASPP consists of one 1 × 1 convolution, three 3 × 3 convolutions with atrous rates 12, 24, and 36, and an average pooling. Then, by passing through a 1 × 1 convolution, feature representation ${B_4}'$ is obtained with less channels.

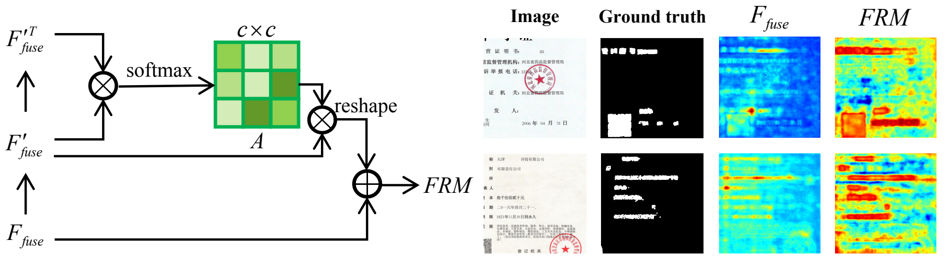
To prevent the information of small objects from losing when the network architecture is deeper, low-level features are utilized. We found through experiments that taking advantage of B_1 and B_2 is helpful to gain richer features. After another 1 × 1 convolution, the corresponding features ${B_1}'$ and ${B_2}'$ are obtained, respectively. Then, we resize ${B_1}'$, ${B_4}'$ to the same size as ${B_2}'$ and concatenate them to form a fused feature $F_{fuse} = \{{B_1}', {B_2}', {B_4}'\}$, where $F_{fuse} \in R^{c\times h\times w}$, c is the sum of the channels of three features. For time efficiency, we set $h = \frac{H}{4}$ and $w = \frac{W}{4}$ in our implement.

Feature recalibration module is motivated by [38], which is the first paper that proposed to improve the quality of feature representations by fusing channel-wise information. Some channels of feature maps can be regarded as responses to one or more manipulations, whereas some are noise. Besides, not all channels contain useful signals. By strengthening the weights of manipulation-related channels, the learning ability of various manipulations would be improved. Therefore, the feature recalibration module is built to capture multi-manipulations traces more accurately by adjusting the weights on channels.

The detailed structure of feature recalibration module is illustrated in Figure 3a. The final output FRM of the feature recalibration module is defined as:

$$\begin{aligned} FRM &= \left[AF'_{fuse}\right]_{reshape} + F_{fuse} \\ &= \left[softmax\left(F'_{fuse}F'^{T}_{fuse}\right)F'_{fuse}\right]_{reshape} + F_{fuse} \end{aligned} \tag{1}$$

where A is the attention matrix calculated by a series of operations on channel, $F'_{fuse} \in R^{c\times(hw)}$ is a flattened matrix of F_{fuse}, F'^{T}_{fuse} is the transpose of F'_{fuse} and $[\cdot]_{reshape}$ means reshaping the dimensions of tensors. We first perform a matrix product between F'_{fuse} and F'^{T}_{fuse}. After a softmax layer, a channel attention matrix A with dimension $c \times c$ is generated, where $A_{i,j}$ represents the effect on the ith channel from the jth channel. Then, we reshape the dimensions of the result of matrix multiplication between A and F'_{fuse} to the same as that of F_{fuse}. Finally, the recalibrated feature FRM is obtained after a element-wise sum. Figure 3b shows the visualization of F_{fuse} and FRM on two fake images. As illustrated in Figure 3b, the FRM generated by adjusting the impacts of channels highlights the tampered regions more prominently than that of F_{fuse}.



(a) Structure of feature recalibration module.

(b) Visualization of F_{fuse} and FRM on two fake images.

Figure 3. The structure of feature recalibration module and visualization of F_{fuse} and FRM.

3.3. Decoder and Loss Functions

The decoder network shown in Figure 4 is designed to predict the forgery map of the input image. The encoder features are computed with 4 output stride. It recovers the encoder features FRM to the original size by three convolution layers and an upsample layer. Three convolution layers consist of two 3×3 layers with 512 filters and one 1×1 layers with 2 filters. Each convolution layer is followed by ReLU activation. The stride of the convolution layers is 1. As a result, the spatial dimension of the feature map can stay at $\frac{H}{4} \times \frac{W}{4}$ even if passing through three convolution layers. In the last upsample layer, the feature map is bilinearly upsampled by a factor of 4. Therefore, after the decoder network, the spatial dimension of prediction mask can be the same as the input's.

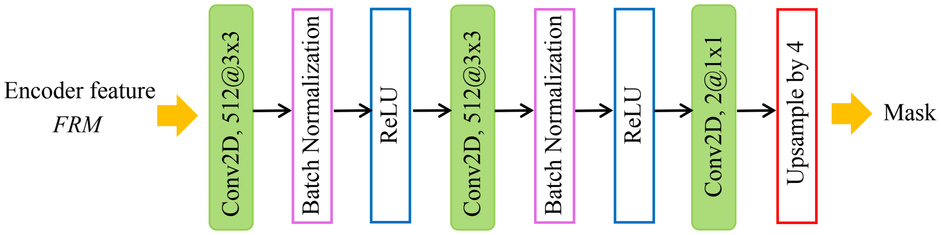


Figure 4. The structure of decoder network.

The loss function L_{total} in MFAN is composed of localization loss and auxiliary loss. The localization loss L_{loc} is the binary cross-entropy loss calculated between the prediction mask and the ground truth label:

$$L_{loc} = y_{gt} log(y_p) + (1 - y_{gt}) log(1 - log(y_p)) \tag{2}$$

where $y_{gt} = 1$ if the pixel is tampered, otherwise $y_{gt} = 0$, and y_p is the prediction mask. The auxiliary loss L_{aux} is added after the third block of ResNet-50, which helps to optimize the learning process. It is also a binary cross-entropy loss defined as follows:

$$L_{aux} = y_{gt} log(y_b) + (1 - y_{gt}) log(1 - log(y_b)) \tag{3}$$

where y_b is the result of feature map B_3 after several convolution layers and resizing.

L_{loc} takes the major responsibility, while L_{aux} is used to assist the network with training. In order to balance the importance between them, the weight α is added to L_{aux}:

$$L_{total} = L_{loc} + \alpha L_{aux} \tag{4}$$

3.4. Implementation Details

The MFAN is implemented by PyTorch on a computer with NVIDIA GeForce RTX 2080 Ti GPU. The ResNet-50 in the encoder is pretrained on the ImageNet dataset. The dilation rate in ASPP is 1, 12, 24, and 36. The learning rates in encoder and decoder module are both 0.01. α is set to 0.2 in Equation (4). Limited by the computer condition, we resize the input images to 512 × 512, and our proposed model is trained with a batch size of 8 by minimizing the loss function L_{total}.

4. Experimental Results

In this section, extensive experiments are conducted to evaluate the performance of the proposed method for fake certificate image detection. We first introduce the dataset and evaluation metrics. Afterward, the demonstration of effectiveness on multi-features fusion and feature recalibration is described in ablation study. Moreover, MFAN is compared with some state-of-the-art forgery detection methods under different cases. Finally, we also carry out experiments on image splicing benchmark datasets to evaluate the universality of MFAN on natural content images.

4.1. Dataset and Evaluation Metrics

In 2020, a security AI challenger program called **Forgery Detection on Certificate Image** was co-sponsored by Alibaba Security and Tsinghua University [39]. This competition provided a TIANCHI dataset that contains tampered certificate images and their corresponding ground truth masks, where part of the data comes from real business scenarios. The image sizes range from 513 × 513 to 1536 × 1536, and we make use of 1000 images from TIANCHI for the experiments. There are seven types of certificate images in TIANCHI dataset: copyright declaration, contract, business license, trademark registration, book cover, honorary certificate and work registration. Tampering manipulations are composed of splicing, copy-move, object removal and text insertion. The challenging thing is that each fake certificate image contains more than one type of manipulation.

The details of experimental data are listed in Table 1. We randomly select 800 images as training data, 100 images as validation data and 100 images for plain testing. What is more, in order to evaluate the robustness of the proposed method, four attacks including JPEG compression, Gaussian noise, resize and median blur are considered on plain testing set. The quality factor of JPEG compression ranges from 60 to 100 with step 10; the Gaussian noise varies in standard deviation from 0.02 to 0.1 with step 0.02; the resize scale ranges from 0.8 to 1.2 with step 0.1 and blur kernel sizes are 3 × 3, 5 × 5, 7 × 7. Thus, 1800 images are generated for robustness evaluation.

Table 1. The details of of training, validation and testing data on TIANCHI dataset.

Sets		Attacks	Parameter Range	Step	Criterion	Number
Training		N/a	-	-	-	800
Validation		N/a	-	-	-	100
Testing	Plain	N/a	-	-	-	100
	Robust	JPEG compression	60∼100	10	Quality factor	500
		Gaussian noise	0.02∼0.1	0.02	Standard deviation	500
		Resize	0.8∼1.2	0.1	Rate	500
		Median blur	3∼7	2	Kernel size	300

To comprehensively evaluate the performance of MFAN, we also conduct experiments on CASIA [40], Columbia [41] and NC2016 [42] to show the universality of our method on natural content images. CASIA is an image forgery benchmark dataset that has two versions, i.e., CASIA v1.0 and CASIA v2.0. In CASIA v1.0, there are 921 compressed images with size 384 × 256. CASIA v2.0 contains 5123 images. Forgeries from both versions are manipulated by splicing or copy-move operations. Columbia consists 180 uncompressed spliced images. In NC2016, images in an average resolution 3561 × 2516 are with finely and fine-grained detailed editing.

To evaluate the performance of our method for certificate image forensics, IoU and F_1 score are applied as evaluation metrics in measuring the accuracy of tampered regions localization.

4.2. Ablation Study

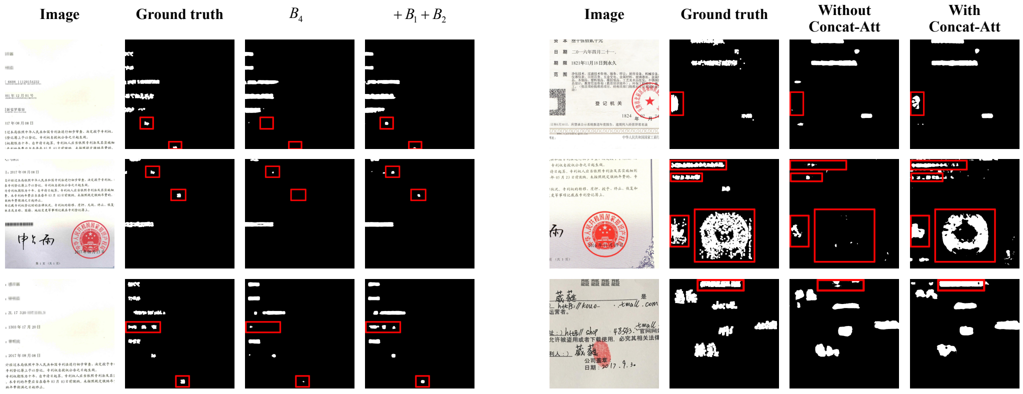
In this subsection, a series of experiments are conducted to explore the effectiveness of some essential components. Table 2 summarizes the localization results by employing multi-level feature fusion, feature recalibration and auxiliary loss, respectively. For multi-level feature fusion, we create a baseline named "B_4" by only feeding the feature map of the fourth block of ResNet-50 into decoder network. For simplicity, "$+B_1 + B_2$" and "$+B_1 + B_2 + B_3$" indicate that low-level features B_1, B_2 and B_3 are concatenated to "B_4" in different ways of combination, respectively. We can see that "$+B_1 + B_2$" achieves 0.6274 IoU and 0.7710 F_1 score, which is the best way for multi-level feature fusion. The visualization results of "B_4" and "$+B_1 + B_2$" are shown in Figure 5a. We can see that "$+B_1 + B_2$" performs better than "B_4" in locating small targets. Furthermore, a quantitative analysis of small targets localization is given in Table 3. We calculate the localization accuracy of "B_4" and "$+B_1 + B_2$" for different scales of tamper regions. Obviously, "$+B_1 + B_2$" overcomes "B_4" from "$\leqslant 20 \times 20$" to "$\leqslant 60 \times 60$".

Table 2. The localization performance comparisons for ablation study.

Methods		IoU	F_1
Multi-level feature fusion	B_4	0.6176	0.7636
	$+B_1 + B_2$	0.6274	0.7710
	$+B_1 + B_2 + B_3$	0.6255	0.7696
Feature recalibration	$+B_1 + B_2$ + ASPP-Att	0.6202	0.7656
	$+B_1 + B_2$ + Concat-Att	0.6346	0.7765
Auxiliary loss	$+B_1 + B_2$ + Concat-Att + aux-0.1	0.6357	0.7773
	$+B_1 + B_2$ + Concat-Att + aux-0.2	**0.6360**	**0.7775**
	$+B_1 + B_2$ + Concat-Att + aux-0.3	0.6352	0.7767

Table 3. The localization performance of small targets.

Methods	Scales of Tampered Regions (Pixel)				
	$\leqslant 20\times 20$	$\leqslant 30\times 30$	$\leqslant 40\times 40$	$\leqslant 50\times 50$	$\leqslant 60\times 60$
B_4	0.1903	0.2387	0.3106	0.3793	0.4661
$+B_1+B_2$	**0.2053**	**0.2423**	**0.3312**	**0.3986**	**0.4948**



(a) Multi-level feature fusion.

(b) Feature recalibration.

Figure 5. Visualization results of (**a**) multi-level feature fusion and (**b**) feature recalibration.

One of the remarkable characteristics of fake certificate images is a mixture of various manipulations. Channels on the feature maps have different emphases for different types of manipulations. Therefore, we recalibrate feature maps on channels on the basis of "$+B_1+B_2$". "ASPP-Att" represents employing attention mechanism after ASPP module. "Concat-Att" means applying attention mechanism on the fusion feature F_{fuse}. From Table 2, it is obvious that the performance of "$+B_1+B_2$+Concat-Att" is better than that of "$+B_1+B_2$ + ASPP-Att". Since some channels of feature maps can be regarded as responses to one or more specific types of manipulations, whereas some are noise, "Concat-Att" can benefit from capturing richer manipulations-related features. The effectiveness of the feature recalibration module is demonstrated in Figure 5b. It can be clearly seen that the localization performance of "With Concat-Att" is better than that of "Without Concat-Att", especially in the bounding boxes.

Furthermore, we investigate the effect of auxiliary loss weight on the result. "aux-0.1", "aux-0.2" and "aux-0.3" indicate the auxiliary loss weight $\alpha = 0.1$, $\alpha = 0.2$ and $\alpha = 0.3$, respectively. Obviously, "$+B_1+B_2$ + Concat-Att + aux-0.2" gets the best result with 0.6360 IoU and 0.7775 F_1 score. As a result, we use it as the proposed method in the following experiments.

4.3. Comparison against Other Methods

To evaluate the performance of the proposed method, we compare it with a number of state-of-the-art methods, which are listed in Table 4. We select three kinds of competing algorithms. The first one is traditional forensics including CFA [16] and NOI [43]. The second one is deep learning based forensics including RRU-Net [36], ManTra-Net [8], MVSS-Net [44] and a top solution TianchiRank-3 [45] in the Tianchi competition, where the ManTra-Net model is pretrained on a private large scale dataset, the MVSS-Net model is pretrained on CASIA v2.0, and TianchiRank-3 is trained with a batch size of 10 and 300 epochs. The last one is semantic segmentation method EncNet [46].

Table 4. The competing methods.

Category	Method	Ref.	Feature/Network
Traditional forensics	CFA	[16]	CFA artifact
	NOI	[43]	Noise artifact
Deep learning based forensics	RRU-Net	[36]	U-Net based network
	ManTra-Net	[8]	Manipulation tracing network
	MVSS-Net	[44]	Multi-view multi-scale network
	TianchiRank-3	[45]	U-Net with Se-Resnext50
Semantic segmentation method	EncNet	[46]	Context encoding network

Table 5 lists the localization results on TIANCHI dataset compared with CFA [16], NOI [43], RRU-Net [36], ManTra-Net [8], MVSS-Net [44], TianchiRank-3 [45] and EncNet [46]. It can be seen that the proposed algorithm achieves the highest score in IoU and F_1 compared with other detection methods, outperforming the second best approach EncNet [46] by 9.49% IoU and 5.80% F_1. Some detection results are shown in Figure 6. It is clear that our proposed MFAN can localize tampered regions more accurately than other detection methods.

Table 5. The localization results on TIANCHI dataset.

Method	IoU	F_1
CFA [16]	0.0509	0.0969
NOI [43]	0.0782	0.1450
RRU-Net [36]	0.4114	0.5340
ManTra-Net [8]	0.0651	0.1223
MVSS-Net [44]	0.0037	0.0073
TianchiRank-3 [45]	0.1794	0.3042
EncNet [46]	0.5809	0.7349
MFAN	**0.6360**	**0.7775**

The good performance of our method in fake certificate image detection is benefited from two main factors. (1) In the feature extraction step, we not only take advantage of the last layer of backbone, but also concatenate it with other low-level features to form a richer feature. (2) The fusion feature was applied by attention mechanism on channels to pay more attention to channels that are associated with the tampered traces of manipulations.

In order to verify the robustness of the proposed method, experiments are carried out and compared with two hand-crafted methods (CFA [16], NOI [43]) and two best performing CNN-based methods (RRU-Net [36], EncNet [46]) under four common attacks including JPEG compression, Gaussian noise, resize and median blur. The details of attacking parameters are illuminated in Table 1. The comparative experiment results under different attacks are shown in Figure 7. Ordinates in Figure 7 represent the F_1 score. From all subfigures in Figure 7, we can clearly see that the proposed method has the best localization performance under different attacks. Figure 7a is the result under JPEG compression. It can be observed that the slopes of all lines are very small, which indicates that these approaches are robust against JPEG compression with quality factors varying from 60 to 100. Figure 7b exhibits the performance under Gaussian noise. With the increase of standard deviation, the F_1 scores of the proposed MFAN and EncNet [46] drop down gradually, and EncNet [46] degrades more rapidly than the proposed method. Figure 7c is the experiment result under resize. All approaches except for NOI [43] show small slopes. The localization result under median blur is demonstrated in Figure 7d. The performances of RRU-Net [36], EncNet [46] and the proposed method get worse when the kernel size becomes larger, among which RRU-Net [36] is less sensitive. The performance of CFA [16] in the Figure 7a–d is relatively stable. We consider that this is because CFA [16] is almost invalid for fake certificate detection, so the interferential influence on CFA [16] is very limited. From the above analysis, it can be found that the proposed MFAN has better and robust performance.

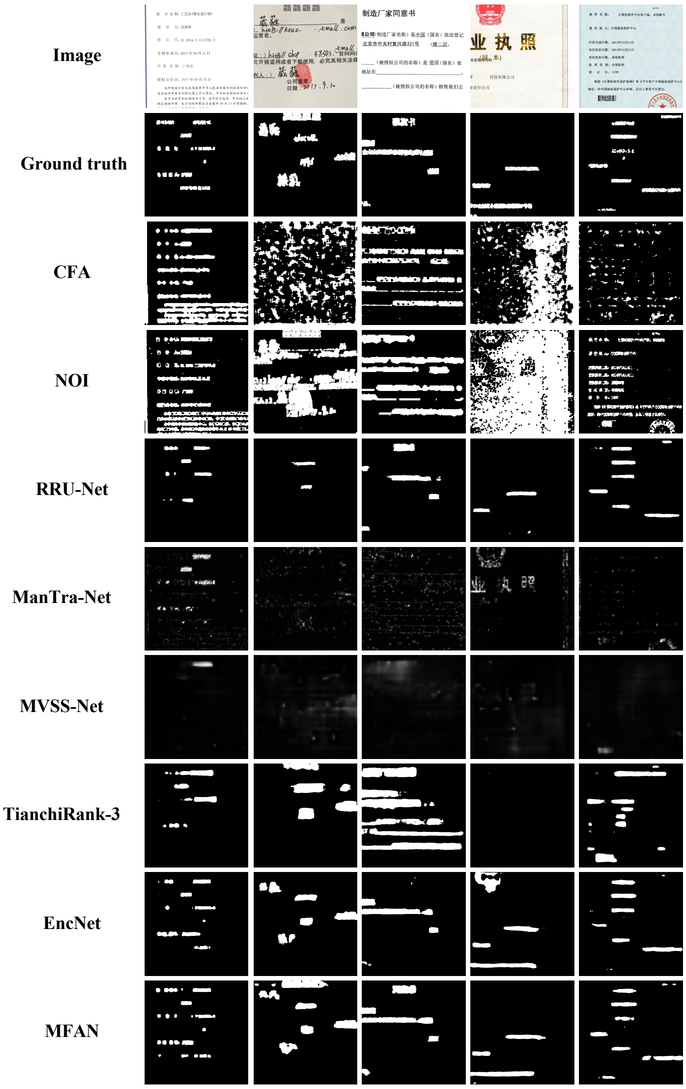


Figure 6. Some detection results on TIANCHI dataset. From top to bottom: tamper images, ground truth masks, CFA [16], NOI [43], RRU-Net [36], ManTra-Net [8], MVSS-Net [44], TianchiRank-3 [45], EncNet [46] and the proposed MFAN.

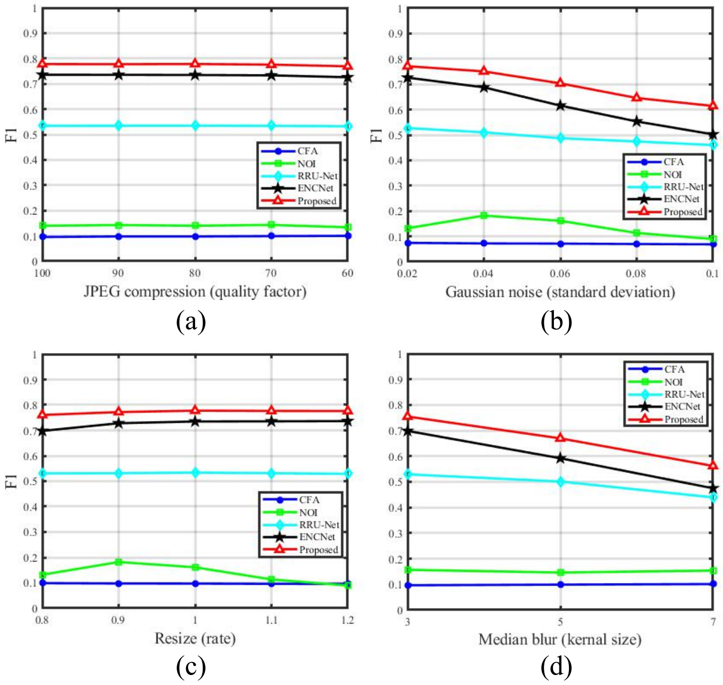


Figure 7. Comparison results under four attacks. JPEG compression: (**a**), Gaussian noise: (**b**), resize: (**c**) and median blur: (**d**). Ordinates represent the F_1 score.

4.4. Performance on Natural Content Image

In addition, more experiments are conducted in this subsection to evaluate the universality of our method on natural content images. Three image forgery benchmark datasets including CASIA v1.0, CASIA v2.0 and Columbia are used. Some examples of detection results are shown in Figure 8. From a subjective perspective, the proposed algorithm localizes tampered areas better than the other four detection approaches.

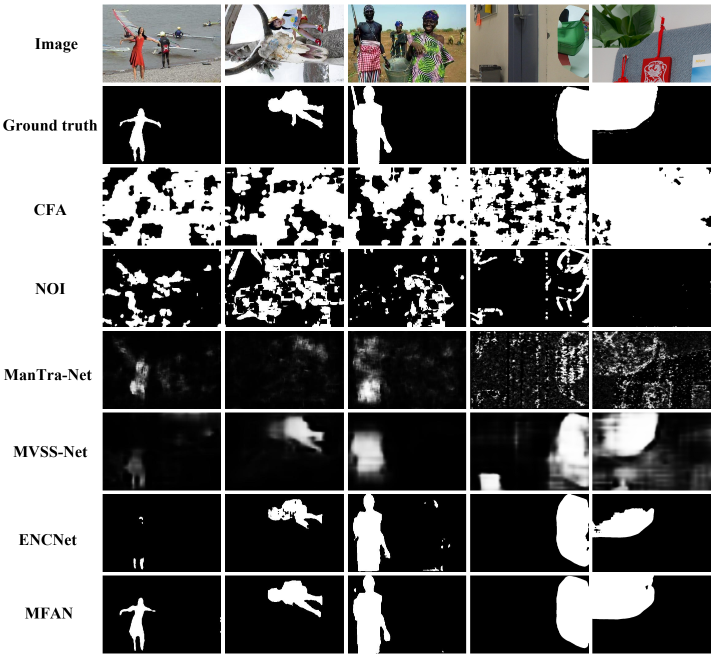


Figure 8. Some detection results on natural content images. From top to bottom: tamper images, ground truth masks, CFA [16], NOI [43], ManTra-Net [8], MVSS-Net [44], EncNet [46] and the proposed MFAN.

In order to accurately evaluate the effectiveness of algorithms, experimental results compared with CFA [16], NOI [43], ManTra-Net [8], MVSS-Net [44] and EncNet [46] on IoU and F_1 score are listed in Tables 6–9. Tables 6–9 report the detection performance on CASIA v1.0, CASIA v2.0, Columbia and NC2016, respectively. As illustrated in Tables 6–9, it can be observed that the proposed method performs better than others in terms of IoU and F_1 score. The traces of splicing areas in Columbia dataset are very obvious, and all images are uncompressed. As a result, forgery detection on Columbia is easier than other datasets. To identify tampered regions, CFA [16] employs CFA artifacts, and NOI [43] takes advantages of noise consistency. Both of them are based on hand-crafted methods that are hard to generate discriminative features. From the above analysis, it can be clearly seen that the proposed MFAN can detect not only fake certificate images well, but also be effective for natural content images. In the future work, we will consider applying our algorithm to images manipulated from GANs and extend it for detecting manipulation in videos.

Table 6. The localization results on CASIA v1.0 dataset.

Method	IoU	F_1
CFA [16]	0.0721	0.1346
NOI [43]	0.0833	0.1538
ManTra-Net [8]	0.1517	0.2635
MVSS-Net [44]	0.2511	0.4015
EncNet [46]	0.3733	0.5439
MFAN	**0.4257**	**0.5972**

Table 7. The localization results on CASIA v2.0 dataset.

Method	IoU	F_1
CFA [16]	0.0803	0.1486
NOI [43]	0.0614	0.1158
ManTra-Net [8]	0.0910	0.1669
EncNet [46]	0.4584	0.6286
MFAN	**0.4790**	**0.6478**

Table 8. The localization results on Columbia dataset.

Method	IoU	F_1
CFA [16]	0.2292	0.3729
NOI [43]	0.1131	0.2033
ManTra-Net [8]	0.2484	0.3982
MVSS-Net [44]	0.2511	0.4015
EncNet [46]	0.8980	0.9463
MFAN	**0.9303**	**0.9639**

Table 9. The localization results on NC2016 dataset.

Method	IoU	F_1
CFA [16]	0.0506	0.0964
NOI [43]	0.0102	0.0202
ManTra-Net [8]	0.0984	0.1791
MVSS-Net [44]	0.1632	0.2807
EncNet [46]	0.8383	0.9120
MFAN	**0.8480**	**0.9177**

5. Conclusions

In this paper, we pay attention to the security of certificate images that are directly related to people's rights and interests, and propose an effective network MFAN for fake certificate image detection. The proposed model is built following the encoder–decoder structure. In the encoder, a pre-trained residual network is used as the backbone to extract rich features by three steps. Firstly, we employ ASPP module on the final layer of residual network to capture contextual information at different scales. Secondly, to keep the information of small objects from being lost, multi-level features are generated by fusing the ASPP feature and two low-level layers from the backbone. Finally, we recalibrate the feature maps on channels to suppress irrelevant information and highlight the tampered regions. In the decoder, the attentive features are fed into a convolutional network with an upsample layer to recover the original size and generate the localization mask. Extensive experiments are conducted to show the effectiveness and robustness of the proposed method in fake certificate image detection.

Author Contributions: Methodology, Y.S. and R.N.; software, Y.S.; writing and editing: Y.S.; analysis: Y.S., R.N. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China (2018YFC0807306), National NSF of China (U1936212), and Beijing Municipal NSF (4222014).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Liu, J.; Wu, S.; Xu, X. A logarithmic quantization-based image watermarking using information entropy in the wavelet domain. *Entropy* **2018**, *20*, 945. [CrossRef]
2. Vukotić, V.; Chappelier, V.; Furon, T. Are classification deep neural networks good for blind image watermarking? *Entropy* **2020**, *22*, 198. [CrossRef]
3. Li, H.; Li, B.; Tan, S.; Huang, J. Identification of deep network generated images using disparities in color components. *Signal Process.* **2020**, *174*, 107616. [CrossRef]
4. Zhang, W.; Zhao, C.; Li, Y. A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. *Entropy* **2020**, *22*, 249. [CrossRef] [PubMed]
5. Li, Y.; Zhou, J. Fast and effective image copy-move forgery detection via hierarchical feature point matching. *IEEE Trans. Inf. Forensics Secur.* **2018**, *14*, 1307–1322. [CrossRef]
6. Barni, M.; Phan, Q.T.; Tondi, B. Copy move source-target disambiguation through multi-branch CNNs. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 1825–1840. [CrossRef]
7. Jalab, H.A.; Subramaniam, T.; Ibrahim, R.W.; Kahtan, H.; Noor, N.F.M. New texture descriptor based on modified fractional entropy for digital image splicing forgery detection. *Entropy* **2019**, *21*, 371. [CrossRef]
8. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9543–9552.
9. 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]
10. Chen, Y.; Lyu, Z.X.; Kang, X.; Wang, Z.J. A rotation-invariant convolutional neural network for image enhancement forensics. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 2111–2115.
11. Wang, Y.; Nian, F.; Li, T.; Meng, Z.; Wang, K. Robust face anti-spoofing with depth information. *J. Vis. Commun. Image Represent.* **2017**, *49*, 332–337. [CrossRef]
12. Caldelli, R.; Galteri, L.; Amerini, I.; Del Bimbo, A. Optical Flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognit. Lett.* **2021**, *146*, 31–37. [CrossRef]
13. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
14. Yuan, R.; Ni, J.; Xie, H. Multi-semantic CRF-based attention model for image forgery detection and localization. *Signal Process.* **2021**, *183*, 108051.

15. Islam, A.; Long, C.; Basharat, A.; Hoogs, A. DOA-GAN: Dual-Order Attentive Generative Adversarial Network for Image copy-move Forgery Detection and Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4676–4685.
16. Ferrara, P.; Bianchi, T.; De Rosa, A.; Piva, A. Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1566–1577. [CrossRef]
17. Huh, A.M.; Liu, A.O.; Efros, A.A. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 101–117.
18. Chang, I.C.; Yu, J.C.; Chang, C.C. A forgery detection algorithm for exemplar-based inpainting images using multi-region relation. *Image Vis. Comput.* **2013**, *31*, 57–71. [CrossRef]
19. 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]
20. Mahmood, T.; Mehmood, Z.; Shah, M.; Saba, T. A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform. *J. Vis. Commun. Image Represent.* **2018**, *53*, 202–214. [CrossRef]
21. Fadl, S.M.; Semary, N.A. Robust copy–move forgery revealing in digital images using polar coordinate system. *Neurocomputing* **2017**, *265*, 57–65. [CrossRef]
22. Bashar, M.; Noda, K.; Ohnishi, N.; Mori, K. Exploring Duplicated Regions in Natural Images. *IEEE Trans. Image Process.* **2010**, 1–40. [CrossRef]
23. Yu, S.; Rongrong, N.; Yao, Z. Nonoverlapping Blocks Based Copy-Move Forgery Detection. *Secur. Commun. Netw.* **2018**, *2018*, 1–11.
24. Amerini, I.; Ballan, L.; Caldelli, R.; Bimbo, A.D.; Serra, G. A SIFT-Based Forensic Method for copy-move Attack Detection and Transformation Recovery. *IEEE Trans. Inf. Forensics Secur.* **2011**, *6*, 1099–1110. [CrossRef]
25. Shivakumar, B.L.; Santhosh Baboo, S. Detection of Region Duplication Forgery in Digital Images Using SURF. *Int. J. Comput. Sci. Issues* **2011**, *8*, 199–205.
26. Li, L.; Li, S.; Zhu, H.; Chu, S.C.; Pan, J.S. An efficient scheme for detecting copy-move forged images by local binary patterns. *J. Inf. Hiding Multimed. Signal Process.* **2013**, *4*, 46–56.
27. Guo, J.M.; Liu, Y.F.; Wu, Z.J. Duplication forgery detection using improved DAISY descriptor. *Expert Syst. Appl.* **2013**, *40*, 707–714. [CrossRef]
28. Wu, Y.; Abd-Almageed, W.; Natarajan, P. Busternet: Detecting copy-move image forgery with source/target localization. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 168–184.
29. Zhang, W.; Tang, X.; Yang, Z.; Niu, S. Multi-scale segmentation strategies in PRNU-based image tampering localization. *Multimed. Tools Appl.* **2019**, *78*, 20113–20132. [CrossRef]
30. Liu, B.; Pun, C.M. Locating splicing forgery by adaptive-SVD noise estimation and vicinity noise descriptor. *Neurocomputing* **2020**, *387*, 172–187. [CrossRef]
31. Mayer, O.; Stamm, M.C. Learned forensic source similarity for unknown camera models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 2012–2016.
32. Choi, H.Y.; Jang, H.U.; Kim, D.; Son, J.; Mun, S.M.; Choi, S.; Lee, H.K. Detecting composite image manipulation based on deep neural networks. In Proceedings of the International Conference on Systems, Signals and Image Processing, Poznan, Poland, 22–24 May 2017; pp. 1–5.
33. Wen, L.; Qi, H.; Lyu, S. Contrast Enhancement Estimation for Digital Image Forensics. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 49:1–49:21. [CrossRef]
34. Bayar, B.; Stamm, M.C. Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2691–2706. [CrossRef]
35. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Learning rich features for image manipulation detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1053–1061.
36. Bi, X.; Wei, Y.; Xiao, B.; Li, W. RRU-Net: The Ringed Residual U-Net for Image Splicing Forgery Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
37. Shang, Z.; Xie, H.; Zha, Z.; Yu, L.; Li, Y.; Zhang, Y. PRRNet: Pixel-Region relation network for face forgery detection. *Pattern Recognit.* **2021**, *116*, 107950. [CrossRef]
38. Hu, J.; Shen, L.; Sun, G.; Albanie, S. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
39. Tianchi. Available online: https://tianchi.aliyun.com/competition/entrance/531812/information (accessed on 13 December 2021).
40. CASIA. Available online: http://forensics.idealtest.org (accessed on 13 December 2021).
41. Hsu, Y.F.; Chang, S.F. Detecting Image Splicing using Geometry Invariants and Camera Characteristics Consistency. In Proceedings of the IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 9–12 July 2006; pp. 549–552.
42. NC2016. Available online: https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation (accessed on 13 December 2021).
43. Lyu, S.; Pan, X.; Zhang, X. Exposing Region Splicing Forgeries with Blind Local Noise Estimation. *Int. J. Comput. Vis.* **2014**, *110*, 202–221. [CrossRef]
44. Chen, X.; Dong, C.; Ji, J.; Cao, j.; Li, X. Image Manipulation Detection by Multi-View Multi-Scale Supervision. In Proceedings of the The IEEE International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021.

45. TianchiRank3. Available online: https://github.com/HighwayWu/Tianchi-FFT2?spm=5176.21852664.0.0.242327c1aOy8EI (accessed on 13 December 2021).

46. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.