applied sciences

MDPI

Article

Multiscale Content-Independent Feature Fusion Network for Source Camera Identification

Changhui You [1], Hong Zheng [2,*], Zhongyuan Guo [2], Tianyu Wang [2] and Xiongbin Wu [2]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan 430000, China; youchanghui@whu.edu.cn
[2] School of Electronic Information, Wuhan University, Wuhan 430000, China; guozhongyuan@whu.edu.cn (Z.G.); tianyuwangwww@163.com (T.W.); xbwu@whu.edu.cn (X.W.)
* Correspondence: zh@whu.edu.cn

Abstract: In recent years, source camera identification has become a research hotspot in the field of image forensics and has received increasing attention. It has high application value in combating the spread of pornographic photos, copyright authentication of art photos, image tampering forensics, and so on. Although the existing algorithms greatly promote the research progress of source camera identification, they still cannot effectively reduce the interference of image content with image forensics. To suppress the influence of image content on source camera identification, a multiscale content-independent feature fusion network (MCIFFN) is proposed to solve the problem of source camera identification. MCIFFN is composed of three parallel branch networks. Before the image is sent to the first two branch networks, an adaptive filtering module is needed to filter the image content and extract the noise features, and then the noise features are sent to the corresponding convolutional neural networks (CNN), respectively. In order to retain the information related to the image color, this paper does not preprocess the third branch network, but directly sends the image data to CNN. Finally, the content-independent features of different scales extracted from the three branch networks are fused, and the fused features are used for image source identification. The CNN feature extraction network in MCIFFN is a shallow network embedded with a squeeze and exception (SE) structure called SE-SCINet. The experimental results show that the proposed MCIFFN is effective and robust, and the classification accuracy is improved by approximately 2% compared with the SE-SCINet network.

Keywords: multiscale; content-independent; source camera identification; fusion network; multi branch

Citation: You, C.; Zheng, H.; Guo, Z.; Wang, T.; Wu, X. Multiscale Content-Independent Feature Fusion Network for Source Camera Identification. *Appl. Sci.* **2021**, *11*, 6752. https://doi.org/10.3390/app11156752

Academic Editor: Andrea Prati

Received: 27 June 2021
Accepted: 19 July 2021
Published: 22 July 2021

1. Introduction

With the rapid development of the new generation of information technology represented by the Internet, big data and artificial intelligence, networking, digitization and intellectualization have become the trend of the times, and digital images have been integrated into all aspects of social life. People can easily use mobile phones or cameras to capture pictures and then use commonly used image editing software to tamper with the image content to spread rumors, commit economic fraud, and other criminal activities. The spread of false pictures is becoming increasingly widespread. As a result, an increasing number of people are losing confidence in the authenticity of digital images and think that images are not reliable information carriers [1,2].

To fight against the crime of fake pictures and rebuild people's trust in image information, digital image forensics has become a research hotspot in recent years. Source camera identification is an important part of digital image forensics, which works to determine from which camera a digital image originated. In addition, source camera identification has a high application value in tracking the source of pornographic images, art photo copyright authentication, and so on. In the past decade, a large number of algorithms for source camera identification has emerged. Although their principles and methods

are different, they all have one thing in common: to extract some traces introduced by human or equipment defects in the image-shooting process and then determine the image acquisition equipment according to these traces. Therefore, let us briefly introduce the general forming process of digital images.

The image-forming process is shown in Figure 1. As shown in the Figure, the light on the surface of the object is projected to the surface of the photosensitive elements Charge Coupled Device (CCD) and Complementary Metal-Oxide (CMOS) through the lens. The light is decomposed into different colored lights by filters on the photosensitive elements. The colored lights are sensed by the corresponding photosensitive units of each filter, and analog current signals of different intensities are generated, which are then converted into digital signals by analog to digital conversion (ADC). Finally, digital signal processing (DSP) carries out color correction and white balance processing, and encodes and compresses these image data into digital images. The formation of digital images requires multiple information processing links involving a series of image-related software processing programs and optical components. However, there are differences in software algorithms and optical components used by different manufacturers or models of cameras. Researchers utilize different techniques to discover traces left by every hardware component or software process during image formation on the image content. These traces are known as intrinsic image artifacts.

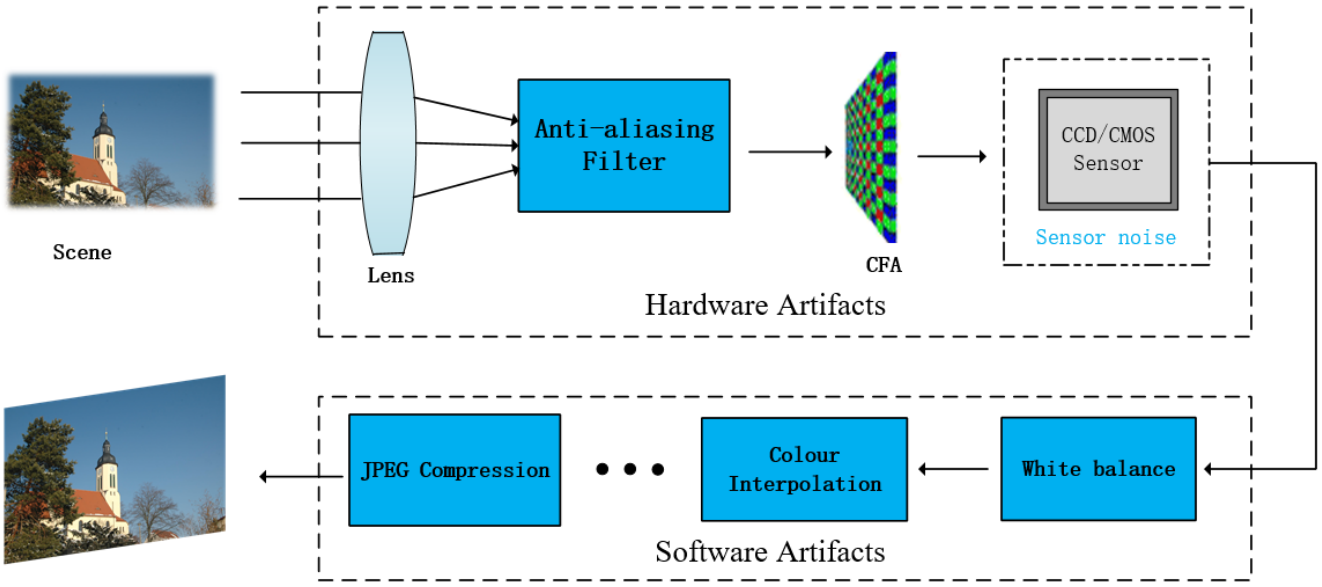

Figure 1. Image acquisition pipeline in typical camera devices.

The existing source camera identification algorithms can be divided into two categories. The first extrac features manually, and then compares similarities. The features related to hardware are artifacts caused by optical and sensor defects. Choi et al. discovered for the first time that the lens radial distortion (LRD) level is dissimilar in different lens manufacturing designs and that this value changes depending on the focal length of the camera lens [3]. A short focal length suffers from more barrel distortion, while a long focal length suffers from further pillow distortion. In [4,5], the authors employed the straight-line method to estimate the LRD parameters. They improved the accuracy of camera attributes by combining the estimated parameters with basic statistical features [6], trying to capture the photometric artifacts and geometric artifacts left by the color-processing algorithm in the image. In [7], the authors present source camera identification via image texture features that are extracted from well-selected color models and color channels, and the proposed method is superior in both detection accuracy and robustness than the other methods. In [8], by considering the image texture, the authors propose to design a

new classifier by adopting a weight function, leading to the remarkable reduction of the feature dimensionality.

In examining the differences in lens distortion parameters of different brands or models of cameras, Hwang et al. proposed a source camera identification method based on the lens distortion correction interpolation attribute. Sensor pattern noise (SPN) is the most serious sensor artifact [9]. It consists of two main parts: fixed pattern noise (FPN) and photo response non-uniformity (PRNU). The method presented in [10] uses a wavelet denoising filter to extract the pattern noise of images; the method presented in [11] is applied for estimation of camera fingerprints by averaging a large amount of reference image noise to suppress random noise components and contamination effects.

The features related to software are the color filter array (CFA) interpolation algorithm, joint photographic experts group (JPEG) compression algorithm, white balance algorithm, and gamma correction. The method presented in [12] established a search space with 36 possible CFA modes and estimated the interpolation coefficients by fitting a linear filtering model in various texture regions of the image for each CFA mode P in search space p. The method presented in [13] proposed a new method based on the basic principle of color interpolation to estimate the CFA mode of a digital camera from a single image. Through a detailed imaging model and its component analysis, the method presented in [14] estimated the intrinsic fingerprint of various camera processing operations.

Another category of source camera identification methods is based on deep learning, which uses CNNs to automatically extract useful features and then classify them using classifiers. The CNN's powerful feature extraction ability makes it outstanding in computer-vision-related tasks. Therefore, many researchers have attempted to apply deep learning methods to the field of image forensics and achieved good results. Luca Bondi et al. in [15] divided an image into several image patches and classified the source camera of each patch. Finally, according to the voting rule, the camera device with the most image patches was selected as the source camera of the image to be tested. Yang divided an image into three types (smooth, saturated, and others) according to the image content, and then used a content adaptive residual network to classify the image source to determine the camera equipment to which the image belongs [16]. Tuama et al. proposed a network similar to Alexnet for image source detection, which is superior to the classical networks Alexnet and Googlenet in camera model detection, and obtained a better detection effect [17].

After AlexNet was introduced in 2012, it won the championship of the Large-Scale Visual Recognition Challenge (ILSVRC). The CNN has attracted the attention of many researchers. In the following years, deep-learning made amazing achievements in image classification [18–20], object detection [21–23], image denoising [24,25], and information security [26]. Due to the strong feature extraction ability of convolutional neural networks and the excellent performance obtained by those techniques on many fields, researchers attempted to apply deep learning to image forensics and achieve better performance than the traditional artificial feature extraction algorithm. For the above reasons, we chose the deep learning scheme for camera source identification. The application of deep learning in image source forensics includes the following three aspects:

1. Using traditional neural networks or making appropriate improvements to the network for source camera identification tasks [27–31].
2. Data enhancement or image preprocessing to improve the data signal-to-noise ratio (SNR) [16,32–35].
3. Improvements in performance through network fusion [16,32].

Although these methods have made great breakthroughs in the field of image forensics, there are still many important problems to be solved, such as how to effectively remove the interference of image content in a forensics task. Digital image forensics is different from computer vision tasks, and the content of images is the largest interference factor. However, the existing convolutional neural network is used to solve computer-vision-related tasks. Therefore, how to effectively apply neural networks to the forensics field has been a difficult problem for researchers. In this paper, a multiscale content-independent feature fusion

network is proposed to reduce the interference of the image content to image forensics and improve the image signal-to-noise ratio. Firstly, we add a multiscale filtering module before each branch network to remove the content information in the image. In contrast to the previous single filter, we innovatively combine multiple scale filters, which can effectively suppress a variety of image content features. In addition, our network can be used as a general scheme, and traditional networks such as AlexNet and ResNet can be easily embedded in the MCIFFN so as to achieve great performance improvements. Experimental results show that the proposed algorithm can effectively suppress the interference of image content and greatly improve the performance of the CNN.

2. Methodology

Although deep learning has achieved excellent performance in computer-vision-related tasks, this does not mean that a traditional CNN can be directly applied to the field of image forensics. In contrast to visual tasks, the key features of image forensics are the noise artifacts left in the image during the image acquisition process, not the image content. By contrast, the image content is the largest interference factor affecting source camera identification. Therefore, to successfully apply the existing CNN to the field of image forensics, we must suppress the image-content-related features as much as possible. In this paper, a multiscale content-independent feature fusion network (MCIFFN) is proposed to solve the problem of source camera identification. In order to capture more comprehensive information of the images, three branch networks are paralleled together to construct the MCIFFN. The three branch networks are used to extract different types of image features by adding different preprocessing modules. The design of the three preprocessing modules is different from each other, which are used to filter different types of image content and extract the noise features. The preprocessing modules of the first two branches are composed of two adaptive filters with different scales, which are used to remove the image content information and extract the multiscale content-independent noise features related to the camera attributes. In order to retain the information related to the image color [13], this paper does not preprocess the third branch network, but directly sends the image data to CNN, so the preprocessing module of the third branch is set to be empty. The image data are first sent to the preprocessing module of each branch to remove the image content features, and then sent to the corresponding CNN feature extraction network. Finally, the CNN features of the three branches are fused, and the fused features are used for image source classification. The CNN in the MCIFFN structure is a shallow network with a squeeze and exception (SE) structure. The structure of MCIFFN is shown in Figure 2.

2.1. MCIFFN Structure

As shown in Figure 2, MCIFFN is composed of three branch networks. The first two branch networks are composed of a preprocessing module and CNN feature extraction module. The function of the preprocessing module is to suppress the image content information and introduce the image forensics domain knowledge into the subsequent deep learning network. The third branch network directly sends the original image data to CNN without preprocessing. In Figure 2, the preprocessing module of the third branch network is NULL, which means no preprocessing. In the first branch, the dense information in the image is removed by a 3×3 adaptive filter to output feature map F_1, and then the sparse information in the image is removed by a 5×5 adaptive filter to output feature map F_2. Finally, the fusion features of F_1 and F_2 are sent to the CNN network. In the second branch, a 5×5 adaptive filter is used to remove the sparse information output feature map F_3, and then F_3 is sent to a 3×3 adaptive filter to remove the residual dense information output feature map F_4. Finally, F_3 and F_4 are fused and sent to the CNN network. The third branch does not preprocess the input data but directly sends the image data to the CNN network, mainly considering that some color information in the image is helpful for image forensics [17].

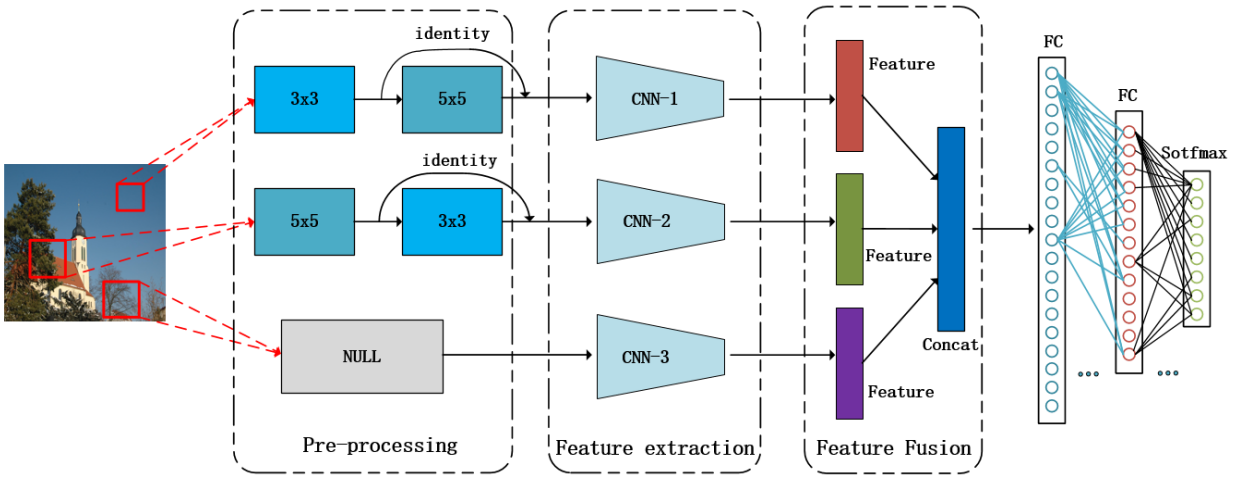

Figure 2. The schema of the MCIFFN structure.

2.1.1. Squeeze and Excitation (SE)

The convolution kernel, as the core of the CNN, is typically used to aggregate spatial information and channel-wise information in a local receptive field and finally obtain global information. A convolutional neural network is composed of a series of convolution layers, nonlinear layers, and down-sampling layers. These layers capture the image features from the global receptive field to describe the image. However, it is very difficult to learn a network and exhibit strong performance. SENet starts from the relationship between feature channels, hoping to explicitly model the interdependence between feature channels. In addition, instead of introducing a new spatial dimension to fuse feature channels, it adopts a new "feature recalibration" strategy. Specifically, it automatically obtains the importance of each feature channel through learning, enhances the key features, and suppresses the useless features according to importance. Generally, it allows the network to use global information to selectively enhance useful feature channels and suppress useless feature channels to realize the adaptive calibration of feature channels. Squeeze-and-Excitation is shown in Figure 3.

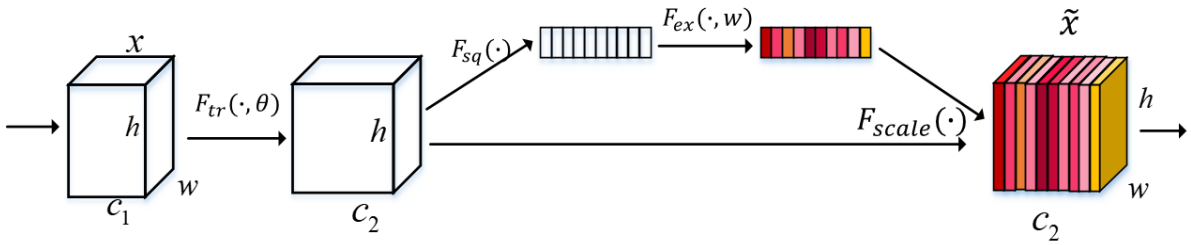

Figure 3. A Squeeze and Excitation block.

Figure 3 illustrates the working principle of the SE module. Given an input x, the number of characteristic channels is C_1. After a series of convolutions and other general transformations, a feature with the number of characteristic channels C_2 is obtained. Different from the traditional CNN, the following three operations are used to recalibrate the previous features:

- The first is the squeeze operation, which compresses the features along the spatial dimension, turning each two-dimensional feature channel into a real number that has a global receptive field to some extent. The output dimension matches the input feature channel number. It represents the global distribution of the response on the feature channel and makes the layer close to the input while also obtaining the global receptive field, which is very useful in many tasks.

- The second is the exception operation, which is similar to the gate mechanism in recurrent neural networks. A parameter W is used to generate weights for each feature channel, where the parameter W is learned to explicitly model the correlation between feature channels.
- The last is a reweight operation, which regards the weight of the output of exception as the importance of each feature channel after feature selection and then weighs the previous feature channel by channel through multiplication to complete the recalibration of the original feature on the channel dimension.

2.1.2. SE-SCINet in MCIFFN Structure

Generally, the deeper the CNN network, the stronger its feature expression ability and the higher its classification accuracy. Deep networks, such as ResNet and DensNet are usually better than shallow networks, such as LeNet and AlexNet. However, for the task of image source forensics, although the detection accuracy of the deep network is higher, the shallow network can also achieve good accuracy, and the network complexity is smaller, and the network reasoning time is faster [16,17,36,37]. Therefore, a shallow network is usually selected for image source forensics.

The CNN in Figure 3 is a shallow network with an SE structure, and that structure is shown in Figure 4. The network proposed in this paper has five convolution layers, five pooling layers, an SE block, and a fully connected layer. The network input data are an $64 \times 64 \times 3$ image patch (64×64 pixels, 3 RGB color channels). As suggested in [38], in order to keep the computational complexity at bay, we use more convolutional layers with smaller kernel sizes instead of using large kernels and fewer convolutional layers. Therefore, all convolution layers in the network use convolution cores with a receptive field of 3×3. Because we still want our CNN to be able to model non-linear functions, we use a single ReLU layer towards the first fully connected layer of the network. This will make the CNN have a wide range of camera models due to the fact that the non-linearity can be helpful to capture non-trivial classes.

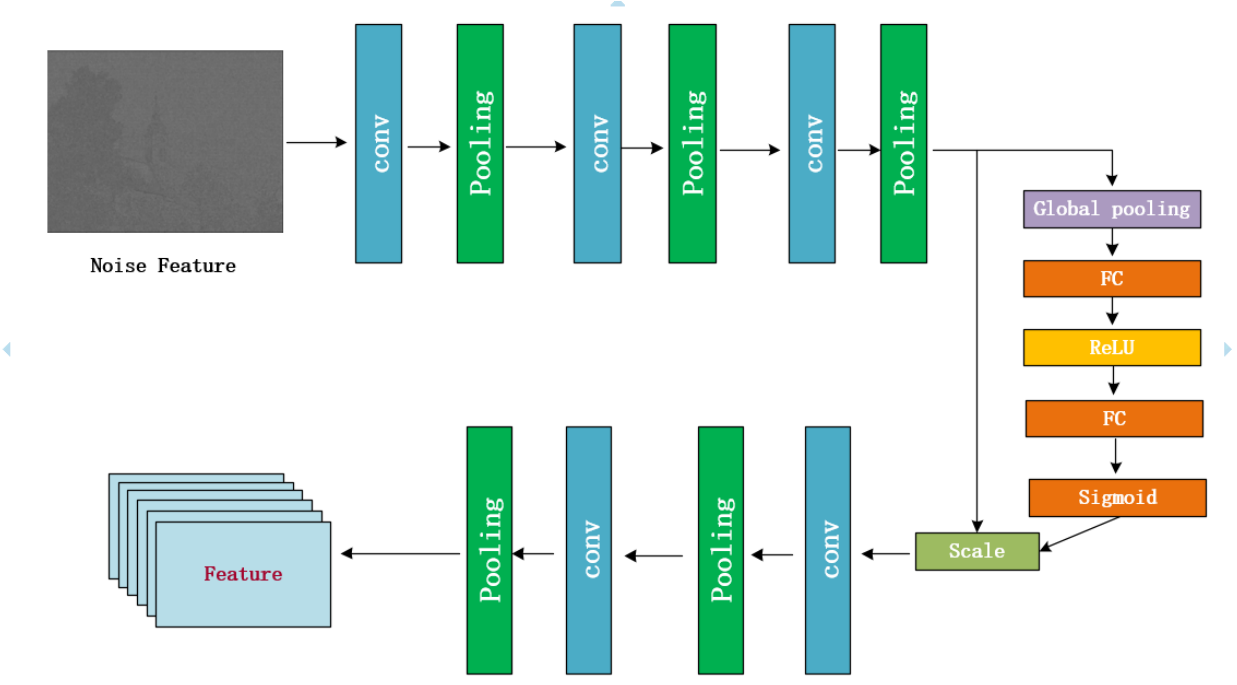

Figure 4. Structure of SE-SCINet network.

Finally, the output features of the fully connected layer are classified by the softmax classifier. In this paper, the standard nonlinear equation is $f(x) = max(0, x)$. Each convolution layer is followed by a max-pooling operation that helps to retain more texture

information and improve convergence performance. The network extracts 128-dimensional features, inputs them to the fully connected layer, and outputs the classification results through a softmax classifier. The convolution of the CNN can only fuse the spatial information of images, and then there is also correlation between channels of the CNN. To make full use of the information between channels, we embed an SE module in this CNN to explicitly model the information between channels. The network parameters are shown in Table 1.

Table 1. Network Parameters of SE-SCINet.

Layer Number	Layer Type	Kernel Size	Filters	Stride	Pad
1	Convolution	3×3	64	1	1
2	Max pooling	2×2	-	1	0
3	Convolution	3×3	32	1	1
4	Max pooling	2×2	-	2	0
5	Convolution	3×3	32	1	1
6	Max pooling	3×3	-	2	0
7	Convolution	3×3	64	1	1
8	Max pooling	2×2	-	2	0
9	Convolution	3×3	128	1	1
10	Max pooling	2×2	-	2	0
11	fc	-	128	-	-
12	ReLu	-	-	-	-
13	fc	-	23	-	-

2.1.3. Multiscale Fusion Analysis

The content and scenes of photos are rich and diverse. There are few pairs of photos that are identical. The same manufacturer or the same camera model might not capture the same or similar content. The scene taken by each camera is random. Therefore, it is impossible to track the camera through the content of the image. By contrast, the randomness and diversity of the image content are the largest interference factors of effective feature extraction. The traditional CNN is designed to solve the task of computer vision. The focus of the network is on the image content. Therefore, a preprocessing module should be added before the CNN to suppress features related to the image content. The preprocessing module is similar to a spatial filter G, which can suppress the content feature of image I and enlarge the image noise feature N.

$$N = I * G \tag{1}$$

The method presented in [39] added a constraint convolution layer to the front of the CNN to suppress image content and adaptively learn image-tampering features. The method presented in [40] used an SRM filter to extract local noise features and detected tampering traces through noise features. The method presented in [41] embedded a Laplacian filter into the first layer to improve the signal-to-noise ratio introduced by the recapture operation. The method presented in [35] designed a convolutional neural network similar to AlexNet for image source detection and preprocessed it with a local binary pattern (LBP). Although the preprocessing method above can suppress the image noise to some extent, because the filter function in the preprocessing layer is too single, it can only remove part of the image content, and improvement in network performance is limited.

Figure 2 shows the proposed multiscale feature fusion network architecture. MCIFFN is composed of three stream networks. There is a preprocessing module at the entrance of the first two stream networks to introduce domain knowledge. The third stream network does not preprocess to save image color information. Due to the randomness of the scene, the image has various scale feature information. The image content can be divided into

smoothing, saturation, and others. The frequency be divided into high-frequency and low-frequency information.

The degree of information density can be divided into sparse information and dense information. A single-scale filter cannot effectively suppress the multiclass content information in the image. Therefore, we add two kinds of receptive field scale adaptive filters to each preprocessing module: a 3×3 filter is mainly used to remove the dense information in the image, and a 5×5 filter is mainly used to remove the sparse information in the image. In previous preprocessing schemes such as the Laplacian filter and SRM filter, the filter parameters are manually set to suppress specific types of image content.

Image forensics tasks have a variety of key features, such as CFA, SPN, PRNU, and other complex features. Although the filter with fixed parameters can suppress the interference features (image content-related features), it may also destroy some key features. The adaptive filter in the preprocessing module of the MCIFFN structure learns the effective features to suppress the useless features and adjusts the filter parameters adaptively through a large amount of sample learning to suppress the useless features to the greatest extent and retain the effective features as much as possible.

In addition, inspired by the idea of feature fusion in ResNet [27], we fuse the features extracted by the two scale filters through identity mapping and send them to the CNN network. $F_3(\bullet)$ is a filter with receptive field 3, $F_5(\bullet)$ is a filter with receptive field 5, I is the input image, N_1 is the input noise of CNN in the first branch, N_2 is the input noise of CNN in the second branch, N_3 is the input noise of CNN in the third branch, and the input characteristics of the CNN in the three branches can be expressed as Formulas (2).

$$\begin{cases} N_1 = F_5(F_3(I)) + F_3(I) \\ N_2 = F_3(F_5(I)) + F_5(I) \\ N_3 = I \end{cases} \tag{2}$$

As shown in Formulas (2), the first branch is that image I first passes through a 3×3 filter to get the output feature $F_3(I)$, then $F_3(I)$ is sent through a 5×5 filter to get the output feature $F_3(F_5(\bullet))$, and finally, output features of the two filters are fused to get the input feature N_1 of the CNN-1 network. Different from the first branch, the second branch is image I, which first passes through a 5×5 filter to obtain the output feature $F_5(I)$, then sends $F_5(I)$ to a 3×3 filter to obtain the output feature $F_3(F_5(\bullet))$, and finally fuses the output features of the two filters to obtain the input N_2 of the CNN-2 network. Different from the first two branches, the third branch does not preprocess the input image, which can retain some color-related features. Therefore, the input of CNN-3 is the image I.

Finally, the MCIFFN fuses the multiscale features of CNN output from three streams and sends them to a softmax classifier for classification. The purpose of our proposed MCIFFN scheme is to provide a network structure suitable for source camera identification. Therefore, the CNN feature extraction network in the three branches of the MCIFFN can select the same or different convolutional neural networks according to the experimental task. In this experiment, the feature extraction network shown in Figure 3 is selected.

3. Experiment and Evaluation

3.1. Dataset

All experiments in this paper are based on the Dresden Image Database [42], which is the most commonly used database in the field of image source forensics and has the most complete types of cameras. Under controlled conditions, more than 14,000 indoor and outdoor scene images were collected from 73 digital cameras covering different camera settings, environments and specific scenes. It is helpful to strictly analyze the characteristics of manufacturers, models, and equipment and their relationship with other influencing factors.

3.2. Performance of MCIFFN

In this experiment, we will verify the rationality of the MCIFFN architecture from the filter size, network structure, and other aspects. We will select 23 camera models from the

Dresden dataset. Each camera model has 20 images. We cut each image into 64×64 pixel non-overlapping image patches, which constitute the dataset of this experiment. The dataset is split by assigning 4/6 of the images to a training set, 1/6 to a validation set, and 1/6 to a test set. The hyperparameter settings of MCIFFN are as follows: batch size is set to 64, training epoch is set to 30, and the number of iterations per epoch is 10,656. Therefore, a total of 319,680 iterations are performed. This can ensure that the training curve fully converges. The solver type is set to a stochastic gradient descent (SGD), the base learning rate is set to 0.001, the policy is set to exponential decay, gamma is set to 0.999; momentum is set to 0.9 and weight decay is set to 0.0001. MCIFFN test results are shown in Figure 5 and Tables 2 and 3.

Table 2. Universal Testing Result of MCIFFN.

Method	AlexNet	MCIFFN-AlexNet	ResNet18	MCIFFN-ResNet18
accuracy (ACC)	92.16	**98.32**	96.06	**97.14**

Table 3. Camera identification performance compared with previous methods.

Method	ACC	Time (ms)	Memory (kb)	Parameters
CAF-CNN [16]	98.2	2.032	2631	670,472
Laplacian-CNN [41]	89.74	1.2028	**1565**	**399,959**
LBP-CNN [35]	92.81	1.1569	5220	1,335,735
HP-CNN [17]	92.24	**1.1031**	5218	1,335,415
MCIFFN	**98.51**	1.3337	2471	632,295

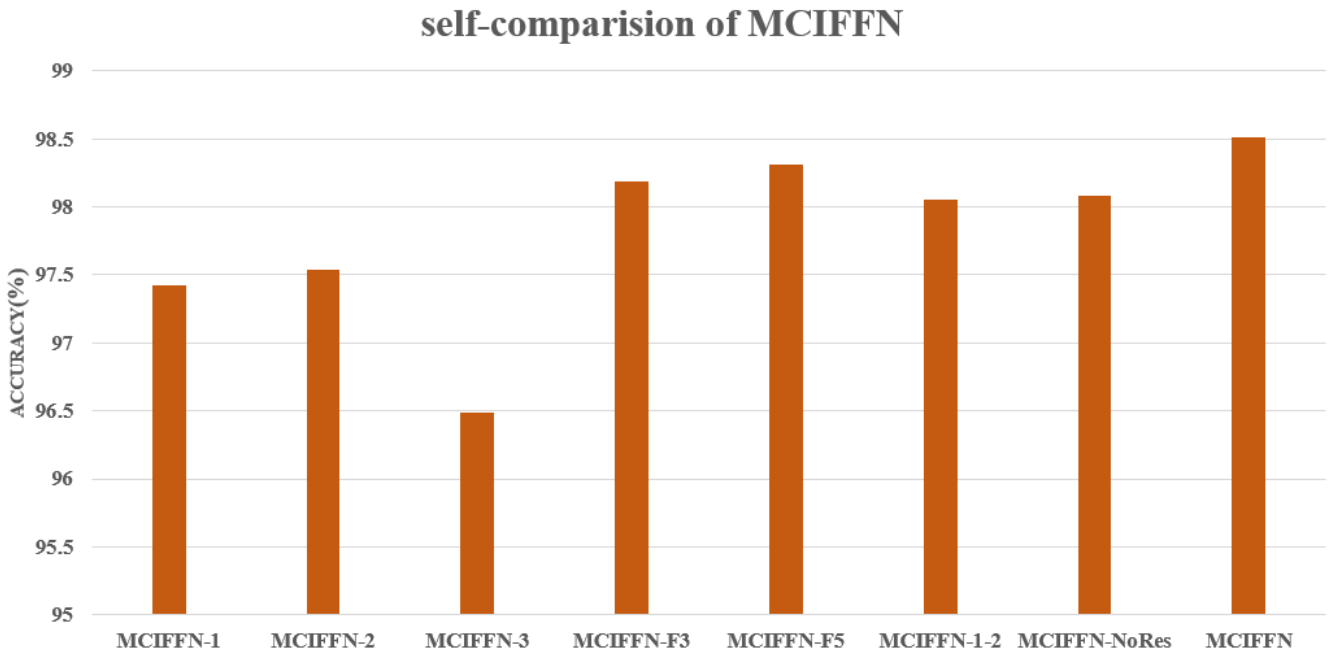

Figure 5. The accuracy of MCIFFN self-comparisons.

To verify the rationality of the algorithm, we make a variety of changes to the MCIFFN and then compare the test results. Before analyzing the experimental results in Figure 5, we need to introduce the different MCIFFN networks in detail. The MCIFFN algorithm is the fusion network shown in Figure 2. MCIFFN-1 is the top branch replaced by the MCIFFN network in Figure 2, MCIFFN-2 is the middle branch of the MCIFFN network, MCIFFN-3 is the bottom branch replaced by the MCIFFN network in Figure 2, MCIFFN-F3 is the network in which the sizes of all filters of the preprocessing module in the MCIFFN are 3×3, MCIFFN-F5 is the network in which the sizes of all filters of the preprocessing module

in the MCIFFN are 5 × 5, MCIFFN-1-2 represents the converged network of branches 1 and 2 of the MCIFFN network and MCIFFN-NoRes is the network that removes the identity mapping between the 3 × 3 and 5 × 5 filters of the MCIFFN network.

The test accuracies of MCIFFN-1, MCIFFN-2, and MCIFFN-3 are lower than that of MCIFFN, which proves that the multibranch fusion scheme can effectively combine a variety of key features and that the network can learn more abundant noise features. From the test results of MCIFFN-F3 and MCIFFN-F5, we can see that there are a variety of image content-related interference features in the image, and the combination of multiscale filters can better suppress the image content. From the test results of MCIFFN-NoRes, it can be seen that the network preprocessing module adds a direct channel to fuse the noise extracted by the two size filters, which can effectively extract a variety of key forensic information.

From the test results of MCIFFN-1-2, it can be seen that although the image content interferes with the extraction of key features, there are still some features related to the source camera in the color-related information. Therefore, our design scheme still retains information flow without preprocessing. MCIFFN-NoRes test results show that adding a direct channel between the two filters can effectively suppress different types of image content information. From the test time of a single image, although the multibranch fusion scheme is more time-consuming, the time difference is not large. From the comprehensive test accuracy and the test time of a single image, the MCIFFN architecture is the best scheme.

Table 2 shows the performance test results of MCIFFN embedded in traditional networks. MCIFFN-AlexNet and MCIFFN-ResNet18 are MCIFFN networks whose CNN is replaced by AlexNet and ResNet18. The test results of MCIFFN-AlexNet and MCIFFN-ResNet18 show that the MCIFFN framework is also suitable for traditional feature extraction networks. The test results of AlexNet and ResNet18 are much lower than those of MCIFFN-AlexNet and MCIFFN-ResNet18, which indicates that the traditional shallow network is not suitable for image forensics tasks directly, and an image content suppression module needs to be added to achieve better results.

To test the performance of the MCIFFN network, we compare the MCIFFN with other existing preprocessing methods, and the results are shown in Table 3. The table records the detection accuracy of each algorithm and the time required to test a single 64 × 64 pixels image patch. This time is obtained by averaging 160,455 test images in the test set. The classification accuracy of LBP-CNN, Laplacian CNN, and HP-CNN is far lower than that of MCIFFN and CAF-CNN. Although CAF-CNN and MCIFFN are close in classification accuracy, their network complexity and the network test time of a single image are far greater than those of MCIFFN. In summary, the MCIFFN has the best classification performance.

To show the classification performance of the MCIFFN more clearly, we group the classification accuracy of each class of cameras in the form of a confusion matrix, and the results are shown in Figure 6. The Figure shows the brand and model information of all cameras involved in this experiment and their classification accuracy. The test result also shows that it is more difficult to distinguish between cameras with the same brand whose feature similarity is higher than that of cameras with different brands, but overall, the classification accuracy can meet the needs of industrialization.

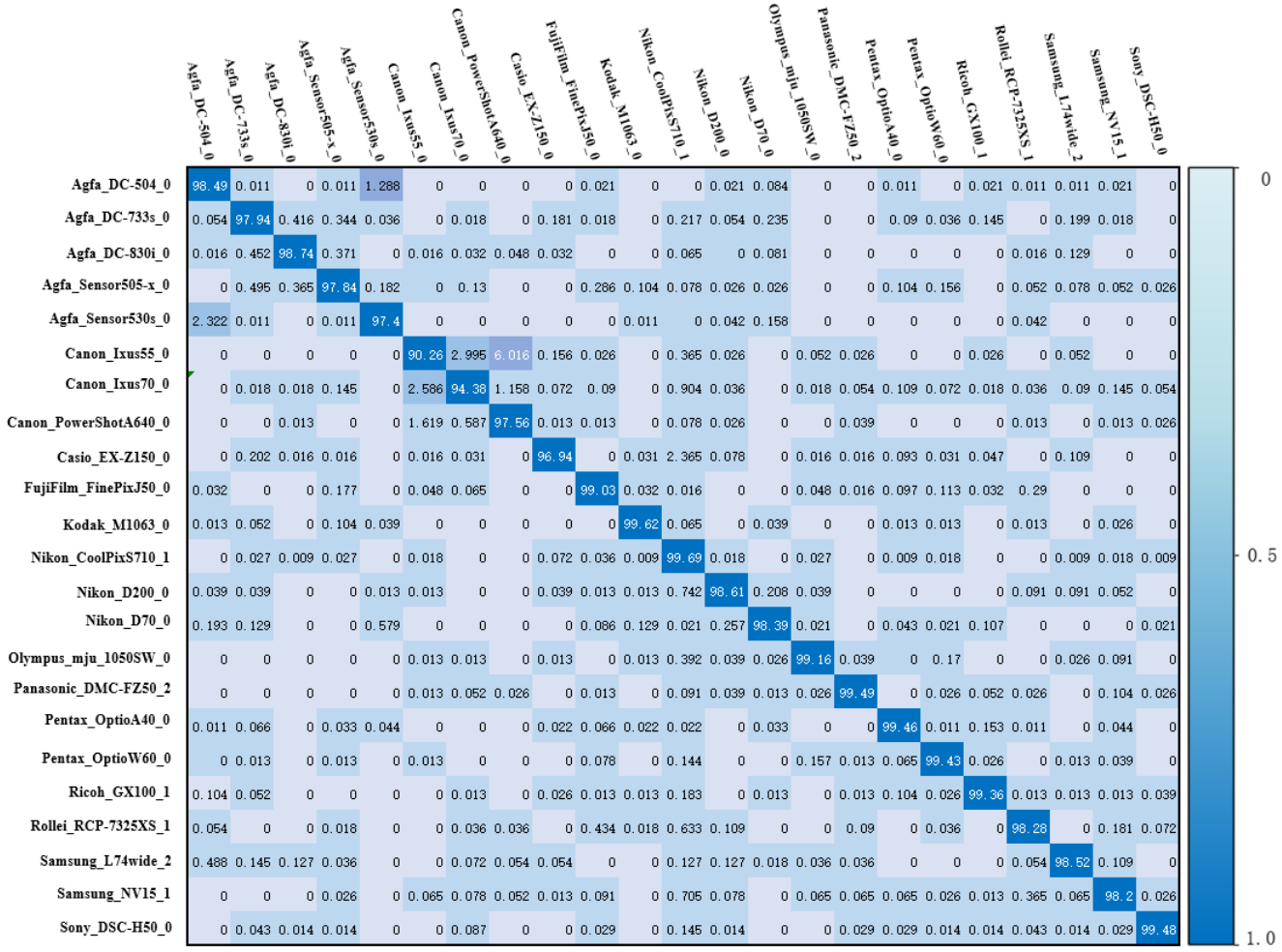

Figure 6. Confusion matrix for identification with 23 different camera models.

4. Conclusions

In this paper, we proposed a multiscale feature fusion network called MCIFFN for source camera identification. To suppress the image content, the MCIFFN uses two sizes of filters to extract camera attribute noise and fuses the two sizes of filter noise through identity mapping. The fused noise can retain more types of camera attribute-related noise. To extract different types of features as much as possible, we used multiple CNNs to extract image features and fused the features extracted from each branch network. Finally, the network selected the useful features by itself. Experimental results showed that the proposed MCIFFN can effectively suppress image content and extract multiscale source camera-related features. Compared with the original SE-SCINet, the classification accuracy improved by more than 2%. In addition, traditional networks such as AlexNet and ResNet can be easily embedded in the MCIFFN so as to achieve great performance improvements. Although our algorithm has been greatly improved in speed and accuracy, it still cannot meet the requirements of an engineering application for model size and running speed. Therefore, our next work will be to further simplify the network structure and realize the engineering application of camera source detection.

Author Contributions: Conceptualization, C.Y. and H.Z.; methodology, C.Y.; software, C.Y.; validation, Z.G.; formal analysis, Z.G.; investigation, T.W.; resources, Z.G.; data curation, T.W.; writing, C.Y. and H.Z.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Supported by the National Key Research and Development Program of China under Grant No. 2020YFF0304902 and the National Natural Science Foundation of China under Grant No. 61771352.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this research can be provided upon request.

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

References

1. Zhu, B.B.; Swanson, M.D. When seeing isn't believing [multimedia authentication technologies]. *IEEE Signal Process. Mag.* **2004**, *21*, 40–49. [CrossRef]
2. Farid, H. Digital doctoring: How to tell the real from the fake. *Significance* **2006**, *3*, 162–166. [CrossRef]
3. Choi, K.S.; Lam, E.Y.; Wong, K.K. Source camera identification using footprints from lens aberration. In Proceedings of the SPIE, San Jose, CA, USA, 9 October 2006.
4. San Choi, K.; Lam, E.Y.; Wong, K.K. Feature selection in source camera identification. In Proceedings of the EEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 8–11 October 2006; pp. 3176–3180.
5. San Choi, K.; Lam, E.Y.; Wong, K.K. Automatic source camera identification using the intrinsic lens radial distortion. *Opt. Express* **2006**, *14*, 11551–11565. [CrossRef] [PubMed]
6. Kharrazi, M.; Sencar, H.T.; Memon, N. Blind source camera identification. In Proceedings of the IEEE International Conference on Image Processing, Singapore, 24–27 October 2004; pp. 709–712.
7. Xu, B.; Wang, X.; Zhou, X.; Xi, J.; Wang, S. Source camera identification from image texture features—ScienceDirect. *Neurocomputing* **2016**, *207*, 131–140. [CrossRef]
8. Zhao, Y.; Zheng, N.; Qiao, T.; Xu, M. Source camera identification via low dimensional PRNU features. *Multimed. Tools Appl.* **2018**, *78*, 8247–8269. [CrossRef]
9. Hwang, M.G.; Park, H.J.; Har, D.H. Determining digital image origin using sensor imperfections. *Aust. J. Forensic Sci.* **2014**, *46*, 98–110. [CrossRef]
10. Lukas, J.; Fridrich, J.; Goljan, M. Source camera identification based on interpolation via lens distortion correction. *Image Video Commun. Process.* **2005**, *2005*, 249–260.
11. Lukas, J.; Fridrich, J.; Goljan, M. Digital camera identification from sensor pattern noise. *IEEE Trans. Inf. Forensics Secur.* **2006**, *1*, 205–214. [CrossRef]
12. Swaminathan, A.; Wu, M.; Liu, K.J.R. Nonintrusive component forensics of visual sensors using output images. *IEEE Trans. Inf. Forensics Secur.* **2007**, *2*, 91–106. [CrossRef]
13. Choi, C.-H.; Choi, J.-H.; Lee, H.-K. CFA Pattern Identification of Digital Cameras Using Intermediate Value Counting. In Proceedings of the Thirteenth ACM Multimedia Workshop on Multimedia and Security, Buffalo, NY, USA, 13 September 2011.
14. Swaminathan, A.; Wu, M.; Liu, K.J. Digital Image Forensics via Intrinsic Fingerprints. *IEEE Trans. Inf. Forensics Secur.* **2008**, *3*, 101–117. [CrossRef]
15. Bondi, L.; Baroffio, L.; Güera, D.; Bestagini, P.; Delp, E.J.; Tubaro, S. First Steps Toward Camera Model Identification With Convolutional Neural Networks. *IEEE Signal Process. Lett.* **2017**, *24*, 259–263. [CrossRef]
16. Yang, P.; Ni, R.; Zhao, Y.; Zhao, W. Source Camera Identification Based On Content-Adaptive Fusion Network. *Pattern Recognit. Lett.* **2017**, *119*, 195–204. [CrossRef]
17. Tuama, A.; Frédéric, C.; Chaumont, M. Camera Model Identification With The Use of Deep Convolutional Neural Networks. In Proceedings of the IEEE International Workshop on Information Forensics and Security, Abu Dhabi, United Arab Emirates, 4–7 December 2016.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
19. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
20. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *IEEE Comput. Soc.* **2016**, arXiv:1608.06993.
21. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *IEEE Comput. Soc.* **2013**, arXiv:1311.2524.
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
23. Kai, Z.; Zuo, W.; Chen, Y. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
24. Redmon, J.; Divvala, S.; Girshick, R. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **2016**, *26*, 3142–3155.

25. Xie, J.; Xu, L.; Chen, E. Image Denoising and Inpainting with Deep Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NA, USA, 3–6 December 2012.
26. Chowdhary, C.L.; Patel, P.V.; Kathrotia, K.J.; Attique, M.; Perumal, K.; Ijaz, M.F. Analytical Study of Hybrid Techniques for Image Encryption and Decryption. *Sensors* **2020**, *18*, 5162. [CrossRef]
27. Bondi, L.; Baroffio, L.; Güera, D.; Bestagini, P.; Delp, E.J.; Tubaro, S. Camera identification with deep convolutional networks. *arXiv* **2016**, arXiv:1603.01068.
28. Freire-Obregón, D.; Narducci, F.; Barra, S.; Castrillón-Santana, M. Deep learning for source camera identification on mobile devices. *Pattern Recognit. Lett.* **2019**, *126*, 86–91. [CrossRef]
29. Huang, N.; He, J.; Zhu, N.; Xuan, X.; Liu, G.; Chang, C. Identification of the source camera of images based on convolutional neural network. *Digit. Investig.* **2018**, *26*, 72–80. [CrossRef]
30. Yao, H.; Qiao, T.; Xu, M.; Zheng, N. Robust multi-classifier for camera model identification based on convolution neural network. *IEEE Access* **2018**, *6*, 24973–24982. [CrossRef]
31. Marra, F.; Gragnaniello, D.; Verdoliva, L. On the vulnerability of deep learning to adversarial attacks for camera model identification. *Signal Process. Image Commun.* **2018**, *65*, 240–248. [CrossRef]
32. Kamal, U.; Rafi, A.M.; Hoque, R.; Das, S.; Abrar, A. Application of DenseNet in Camera Model Identification and Post-processing Detection. *arXiv* **2018**, arXiv:1809.00576.
33. Bayar, B.; Stamm, M.C. Augmented convolutional feature maps for robust cnn-based camera model identification. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 4098–4102.
34. Zuo, Z. Camera Model Identification with Convolutional Neural Networks and Image Noise Pattern. 2018. Available online: http://hdl.handle.net/2142/100123 (accessed on 2 July 2018).
35. Wang, B.; Yin, J.; Tan, S.; Li, Y.; Li, M. Source camera model identification based on convolutional neural networks with local binary patterns coding. *Signal Process. Image Commun.* **2018**, *68*, 162–168. [CrossRef]
36. Bondi, L.; Güera, D.; Baroffio, L.; Bestagini, P.; Delp, E.J.; Tubaro, S. A Preliminary Study on Convolutional Neural Networks for Camera Model Identification. *Electron. Imaging* **2017**, *7*, 67–76. [CrossRef]
37. Yang, P.; Baracchi, D.; Ni, R.; Zhao, Y.; Argenti, F.; Piva, A. A Survey of Deep Learning-Based Source Image Forensics. *J. Imaging* **2020**, *6*, 9. [CrossRef]
38. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
39. 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]
40. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Learning Rich Features for Image Manipulation Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
41. Yang, P.; Ni, R.; Zhao, Y. Recapture image forensics based on Laplacian convolutional neural networks. In *International Workshop on Digital Watermarking*; Springer: Berlin, Germany, 2016; pp. 119–128.
42. Gloe, T.; BóHme, R. The Dresden Image Database for Benchmarking Digital Image Forensics. *J. Digit. Forensic Pract.* **2010**, *3*, 150–159. [CrossRef]