Intelligent Automation
& Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.039805
Article

A New Method for Image Tamper Detection Based on an Improved U-Net

Jie Zhang, Jianxun Zhang*, Bowen Li, Jie Cao and Yifan Guo

Department of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 40005, China
*Corresponding Author: Jianxun Zhang. Email: zjx@cqut.edu.cn

Received: 17 February 2023; Accepted: 28 April 2023; Published: 11 September 2023

Abstract: With the improvement of image editing technology, the threshold of image tampering technology decreases, which leads to a decrease in the authenticity of image content. This has also driven research on image forgery detection techniques. In this paper, a U-Net with multiple sensory field feature extraction (MSCU-Net) for image forgery detection is proposed. The proposed MSCU-Net is an end-to-end image essential attribute segmentation network that can perform image forgery detection without any pre-processing or post-processing. MSCU-Net replaces the single-scale convolution module in the original network with an improved multiple perceptual field convolution module so that the decoder can synthesize the features of different perceptual fields use residual propagation and residual feedback to recall the input feature information and consolidate the input feature information to make the difference in image attributes between the untampered and tampered regions more obvious, and introduce the channel coordinate confusion attention mechanism (CCCA) in skip-connection to further improve the segmentation accuracy of the network. In this paper, extensive experiments are conducted on various mainstream datasets, and the results verify the effectiveness of the proposed method, which outperforms the state-of-the-art image forgery detection methods.

Keywords: Forgery detection; multiple receptive fields; cyclic residuals; U-Net; channel coordinate confusion attention

1 Introduction

The ever-evolving image processing technology and powerful image processing software are heavily integrated into our lives. Even users without professional image processing knowledge can very easily tamper with the image content, resulting in a reduction in the authenticity of the image content. There are three types of common image processing techniques, as shown in Fig. 1.

(1) Splicing: A splicing method that copies areas from one image and pastes them into other images.

(2) Copy-move: A method of copying certain areas of an image and moving them to cover other areas.
(3) Remove: A method to remove some regions from an image.

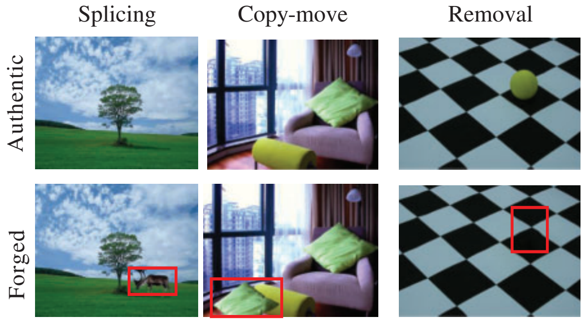

Figure 1: Examples of common forged image types (tampered areas are in red boxes)

The current forgery detection methods mainly identify the differences in image properties between the tampered and untampered regions of an image such as differences in lighting, shadows, sensor noise, and camera reflections.

Traditional tampering detection algorithms for clipping combinations are mainly based on the differences in attributes between tampered and non-tampered regions in images, and detection algorithms based on these attribute differences can be broadly classified into four categories:

(1) Detection methods based on the essential image attributes [1–3].

(2) Detection methods based on imaging device attributes [4–6].

(3) Detection method based on image compression attributes [7–11].

(4) Image hash-based detection methods [12,13].

These methods are used to obtain these attributes. A failed detection occurs when these attributes are not apparent or do not exist. Furthermore, postprocessing operations such as image blurring, JPEG compression, and subsampling can affect specific image attributes, reducing the detection efficiency of traditional detection methods.

To date, deep learning has led technological advances in areas such as computer vision. In this process, deep learning techniques have also been gradually introduced in the localization of image falsification. Liu [14] proposed a stitching tampering pixel-level localization framework based on full convolutional networks and conditional random fields that can easily overfit overfitting due to the limitation of the dataset. Bi et al. [15] proposed a circular residual U-net network to enhance the learning abilities of CNNs.

With [16–18] being subsequently proposed, pixel-level localization for stitching, copy-paste and image removal tampering is achieved. However, the current CNN-based feature extraction detection methods suffer from the loss of contextual information, gradient degradation of deeper networks, and single extracted features. In summary, to address the problems of existing CNN-dwelling detection methods, a multiple sensory field-scale feature extraction and channel coordinate confusion U-Net (MSCU-Net) is proposed in this paper to solve these problems. Our main contributions are as follows.

We use the multiple perceptual field feature extraction module instead of the single convolution module in the traditional U-Net, enabling the model to extract more features and have a larger perceptual field.

Moreover, we introduce channel coordinates to confuse the attention mechanism in the skip-connection of the model to enhance the spatial information encoded by the CNN.

2 Related Work

Deep learning methods based on data-driven approaches have shown excellent performance on many computer vision and image processing tasks. Thus, many researchers have also applied deep learning to image tampering detection, and a few of the more popular directions are improving the detection accuracy by improving network structures, network models based on the correlation design of imaging devices, etc.

The U-Net network is essentially a convolutional neural network that was proposed by Ronneberger et al. [19] in 2015 and is widely used in medical fields.

It is widely used in the field of medical image processing and has improved both the detection rate and detection accuracy compared to the traditional neural structure segmentation algorithm.

U-Net uses successive convolutional layers and maximum pooling layers to obtain the contextual feature information in images uses a series of upsampling layers to interpolate and amplify the obtained feature information to obtain a high-resolution feature map, and finally uses the lateral propagation of features between layers to reduce the loss of feature details and accurately locate the tampered region. However, since the identifiable features between tampered and non-tampered regions in the image are more hidden and weak, these identifiable features will have gradient disappearance when the network structure is deeper. To solve this problem, He et al. proposed ResNet [20] in 2015. The principle is that by adding a constant mapping to the shallow network and directly skipping the intermediate layer to transfer to the later network layers, the input of a segment of the neural network is superimposed on its output as the input of the lower network through a shortcut connection, and the output $y(x) = F(x) + x$ converts the learning target from $F(x)$ to $F(x) + x$, which can ensure a deeper depth. It simplifies the network training and enhances the learning methods of CNNs.

Hu et al. [21] developed the first attention module used in computer vision and proposed the squeeze-and-excitation module (SE), which extracts both spatial and channel information of the feature map and is widely used in various tasks in the field of vision. Woo et al. [22] proposed the convolutional block attention module (CBAM) module, which is based on the SE module. Zhang et al. [23] proposed SA-Net to fuse the output feature maps at each scale with reference to the top-down characteristics of the human visual system. However, since tampering detection task features are difficult to identify, the channel coordinate confusion attention (CCCA) module is proposed in this paper to enhance the feature extraction capability of the network.

3 Multisensory Field and Channel Coordinate Obfuscation U-Net

3.1 Multiscale Feature Extraction U-Net Network

The multiscale feature extraction module solves the problem of traditional networks extracting single features, and the propagation of cyclic residuals both solves the problem of gradient degradation and amplifies the differences in image attributes between untampered and tampered regions. The CCCA attention mechanism enhances the spatial coordinate information of encoding, enhances meaningful features, and suppresses useless features. In summary, the cyclic multiscale feature extraction structure extracts features between the layers of the network while ensuring a more obvious discrimination of the essential image attribute features, which can achieve better and more detection performance compared with traditional feature extraction-based detection methods and existing CNN-based detection methods. The network architecture of MSCU-net is shown in Fig. 2.

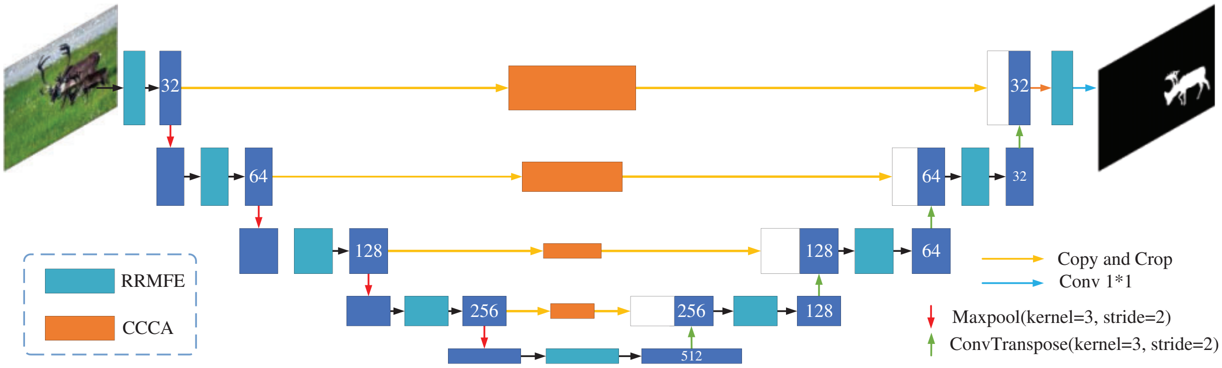
32
64
128
256
512
256
128
64
128
64
32
32
RRMFE
CCCA
Copy and Crop
Conv 1*1
Maxpool(kernel=3, stride=2)
ConvTranspose(kernel=3, stride=2)

Figure 2: The multisensory field and channel coordinate confused U-Net (the number in the box indicates the number of feature map channels)

3.2 Multisensory Field Feature Extraction Module

The U-Net network sends the feature maps with much detail in the shallow network to the corresponding decoders to assist in segmentation by adding a jump connection between the encoder and decoder. However, the decoders uniformly use the single-scale convolution module for the merged feature maps. The single-scale convolution module uses only fixed-size convolution kernels, which makes it difficult to make full use of the information at different scales in the merged feature maps. In this paper, we propose an improved feature extraction module with multiple sensory fields. As shown in Fig. 3, the single-scale convolution module in the decoder is replaced by the improved multiscale module, which enables the decoder to utilize the feature information of different receptive fields in a comprehensive manner and further improve the segmentation accuracy of the network. The improved multiscale convolution module is divided into two parts: multiscale and feature fusion. The multiscale part consists of a series of convolutional layers of different scales, through which the feature information of different sensory fields in the feature map is extracted, and the $1 * 1$ convolution is introduced in front of the multiscale convolutional layer to downscale the input feature map with reference to the Inception structure [24], which can ensure the feature extraction effect and at the same time effectively reduce the computation of the whole module and further improve the generalization ability of the model by introducing more nonlinearities. The feature fusion part uses a $3 * 3$ convolution to fuse and downscale the output feature information of the multiscale part, which further reduces the computational effort of the model while extracting more features of the tampered region without affecting the model effect.

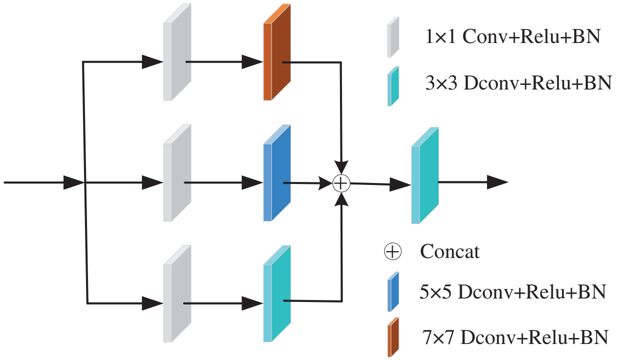

Figure 3: Multi-sensory field feature extraction module (MFE)

3.3 The Ringed Residual Module

To solve the vanishing gradient problem, we introduce the structure of cyclic residuals.

A cyclic residual module is shown in Fig. 4, which is subdivided into residual propagation and residual feedback, with the output of residual propagation defined as

$$y_f = F\left(x, \{W_i\}\right) + W_s * x \tag{1}$$

where x and y are the inputs and outputs of the module, W_i denotes the weights of the i th layer, and the function $F\left(x, \{W_i\}\right)$ denotes the residual mapping to be learned.

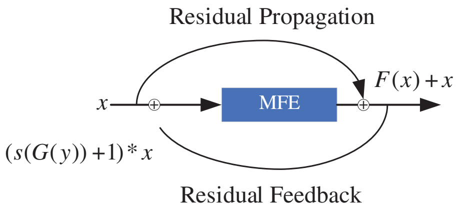

Figure 4: The ringed residual MFE (RRMFE)

Residual feedback, through sigmoid activation response values superimposed on the input information to amplify the difference in the essential properties of the image between the untampered and tampered regions, is defined by Eq. (2).

$$y_b = \left(s\left(G\left(y_f\right)\right) + 1\right) * x \tag{2}$$

where x is the input, y_f is the equation, and y_b is the augmented input. The function G is a linear mapping for changing the size of y_f. The function s is the sigmoid activation function. Residual feedback is analogous to the human need to repeat what has been learned and learning new knowledge in the process of constant repetition. Residual feedback can amplify the differences between image attributes. The structure of the circular residuals is shown in Fig. 4.

3.4 Channel Coordinates Confuse Attention

We propose channel coordinates confuse attention (CCCA) based on Spatial-Channel Squeeze & Excitation (scSE) [25]. CCCA consists of three channels that obtain the information of feature maps in spatial coordinates.

The input feature map of $\hat{U}_{cSE}$ is $U = [u_1, u_2 \ldots, u_c]$, where each channel $u_i \in R^{W*h}$, U is globally averaged to obtain the vector $z \in R^{1*1*c}$. Moreover, the k values at each position are

$$z_k = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} u_k\,(i,j) \tag{3}$$

The obtained vector passes through two fully connected layers.

$$\hat{Z} = W_1\,(\delta\,(W_2 z)) \tag{4}$$

where W_1 and W_2 are the weights of the fully connected layer. Afterward, after ReLU is performed, this process enhances the independence between each channel. For $\hat{Z}$ obtained after the sigmoid layer and normalized to between 0 and 1 to obtain $\sigma(\hat{Z})$, the whole calculation process can be expressed by the following equation:

$\hat{U}_{eCA}$ performs a global average pooling operation on the input feature map, performs a 1-dimensional convolution operation with a convolution kernel size of k, and obtains the weights w of each channel after the sigmoid activation function. The computation procedure is shown below:

$$\hat{U}_{eCA} = \left[\sigma\left(q_{1,1}\right)u^{1,1}, \ldots, \sigma\left(q_{i,j}\right)u^{i,j}, \sigma\left(q_{H,W}\right)u^{H,W}\right] \tag{5}$$

Then, the weights are multiplied with the corresponding elements of the original input feature map to obtain the final output feature map.

$\hat{U}_{cA}$ encodes the channel relationships and long-range dependencies by means of accurate location information. For the input X, each channel is first encoded along the horizontal and vertical coordinate directions using pooling kernels of dimensions $(H, 1)$ and $(1, W)$, and the captured location information is fully utilized to precisely locate the region of interest. The whole computational process can be expressed by the following equation:

$$y_c\,(i,j) = x_c\,(i,j) \times g_c^h\,(i) \times g_c^w\,(j) \tag{6}$$

CCCA is the combination of the above three, as shown in Fig. 5. By introducing the CCCA attention mechanism in skip-connection to enhance meaningful features and suppress useless features, the difference between tampered regions and untampered image attributes is amplified.

4 Experiments

To evaluate the performance of the proposed MSCU-Net, various experiments have been conducted to determine the effectiveness and robustness. Additionally, the method is compared with several other image forgery detection methods in different situations.

Experimental Dataset: We selected three public datasets, CASIA [26], NIST16 [27] and Columbia [5], for testing.

On CASIA, the witty vintage set contains images with the three tampering means of copy-shift, stitching, and removal.

On NIST16, the dataset provides images with three tampering means copy-shift, splice, and removal.

Columbia provides images of the stitching tampering means.

For the NIST2016 image dataset, there are two main subsets: training (404) and testing. Then, these subsets are randomly selected.

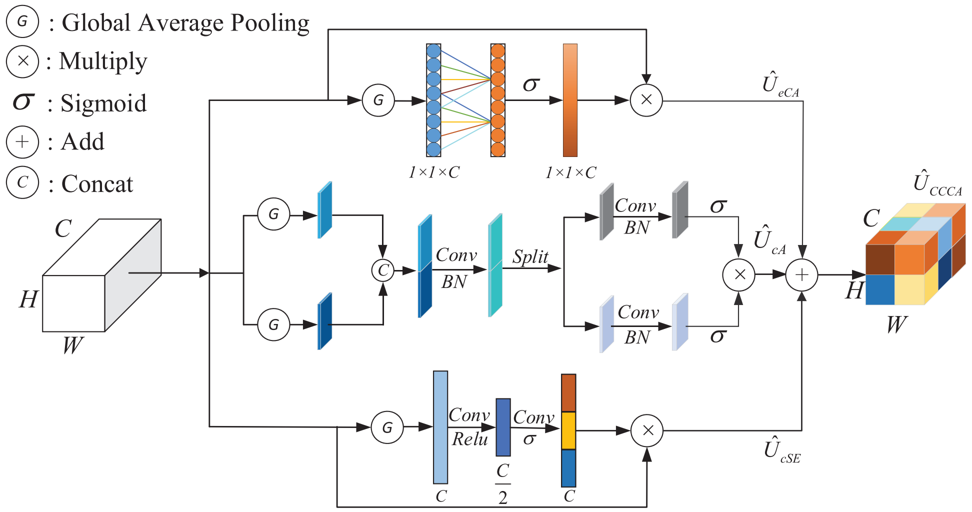

Figure 5: Channel coordinate confusion attention (CCCA). By adding $\hat{U}_{eCA}$ and $\hat{U}_{cSE}$ to obtain a more precisely calibrated spatial feature map, $\hat{U}_{cA}$ enables the network to focus on a large range of location information

For the CASIA image library, 5123 images from CASIA v2.0 were selected for training, 921 images from CASIA v1.0 were used for testing, and the entire Columbia dataset was used for testing. The division details are shown in Table 1. To better train MSCU-Net, the image size was uniformly processed to $384 * 256$.

Table 1: Training and testing split (number of images, the sign '-' in the table indicates that the item is not required)

	CASIA	Columbia	NIST16
Train	5123(CASIAv2.0)	-	404
Test	921(CASIAv1.0)	180	160

Evaluation Metrics: The evaluation metrics referred to in the comparison experiment section of this paper are *Precision*, *Recall*, and *F-measure*, and the evaluation metrics are the number of correctly detected tampered pixels (*TP*), the number of incorrectly detected tampered pixels (*FP*) and the number of incorrectly detected untampered pixels (*FN*). Among them, the accuracy rate is calculated as shown in Eq. (7), the recall rate is calculated as shown in Eq. (8), and the *F*-measure is calculated as shown in Eq. (9).

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

Compare detection methods: To evaluate the practical effectiveness of the algorithms proposed in this paper, the following models are selected for comparison.

The C2R-Net [28] algorithm can locate the tampered content in the image, but there are still a small number of false detections and more missed detections.

RGB-N [16] which is a two-stream Faster R-CNN network, is proposed and trained end-to-end to detect tampered regions of a given image.

SPAN [18] utilizes a pyramid architecture and models the dependency of image patches through self-attentive blocks.

RRU-Net [15] reinforces the CNN learning approach to amplify the difference between tampered and untampered regions.

Table 2 shows the average values of precision, recall, and F1 for the detection results of the proposed algorithm and the five compared algorithms on three publicly available datasets. From Table 2, it can be seen that the detection results of the proposed algorithm in this paper are better than the other five comparison algorithms in terms of precision and F1. However, the precision and recall on the CASIA dataset and Columbia dataset are slightly lower than those of the RRU-Net algorithm. The F-measure score metrics implemented on both image libraries indicate that the method in this paper achieves relatively high tampering localization results. In addition, images of the results of image tampering localization are provided. The test result picture displayed in Fig. 6 illustrates that our model delivers a better segmentation effect, as it can be observed.

Table 2: Training and testing split (number of images, the sign '-' in the table indicates that the item is not required, F1 in this table is F-measure)

Method	CASIA			NIST16			Columbia		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
C2R-Net	0.417	0.424	0.420	-	-	-	0.732	0.821	0.695
RGB-N	0.509	0.453	0.408	0.673	0.764	0.722	0.797	0.642	0.697
SPAN	-	-	0.382	-	-	0.582	-	-	0.815
U-Net	0.432	0.231	0.352	0.535	0.689	0.533	-	-	0.634
RRU-Net	**0.464**	0.440	0.397	0.892	0.862	0.862	0.882	**0.797**	0.863
Ours	0.451	**0.485**	**0.422**	**0.905**	**0.875**	**0.876**	**0.914**	0.773	**0.869**

Ablation Experiments: To verify the effectiveness of the CCCA and RRMFE modules, we perform ablation experiments on the NIST16 dataset, as shown in Table 3. The terms in bold font indicate the optimal performance value, $\surd$ indicates that the experimental model contains the corresponding module, and × indicates that the experimental model does not contain the corresponding module. As shown in the second row of the table, after directly fusing the spatial location coordinate information of the images, the experimental results are improved by 0.013 for precision, 0.003 for recall, and 0.067 for F-measure compared with the original network. To further amplify the attribute differences between tampered and untampered regions, the RRMFE module for different scales and different perceptual fields of feature fusion are proposed in this paper. The experimental results show that the *F*-measure and precision performance are optimal with the feature fusion of CCCA+RRMFE.

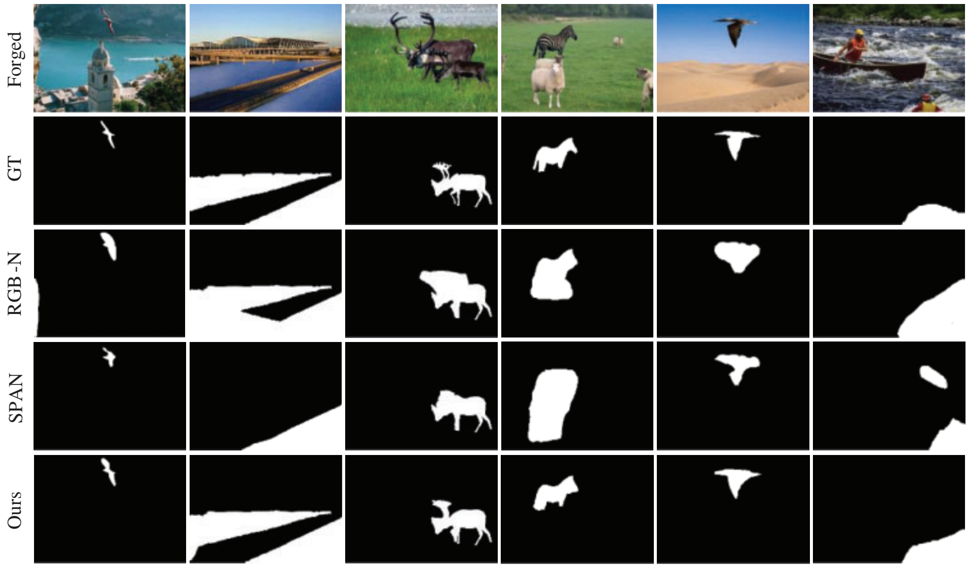

Figure 6: Visualization of our prediction results, showing the forged images, GT, RGB-N prediction results, SPAN prediction results and our prediction results from top to bottom. From the figure, our network model has better results in the details of segmentation

Table 3: Comparison of the model validity ablation experimental results (F1 in this table is the *F*-measure)

Baseline	CCCA	RRMFE	Precision	Recall	F1
√	×	×	0.892	0.8623	0.8628
√	√	×	0.9005	0.8626	0.8695
√	×	√	0.8928	**0.8799**	0.8741
√	√	√	**0.9052**	0.8751	**0.8766**

Robustness experiments: In real scenarios, tampered images mostly undergo various types of postprocessing operations, such as network transmission compression and noise. Therefore, a tamper detection and localization framework that is robust against postprocessing operations is particularly important. In this paper, we design robustness experiments for 2 types of postprocessing operations. The specific types and parameters are shown in Table 4. The experimental results of MSCU-Net are smooth and robust under the postprocessing operations of each parameter, and the experimental results are shown in Figs. 7 and 8.

Table 4: Robust experimental post-processing operations and their parameter settings

Post-processing means	Parameter values
JPEG compression	{100,90,80,70,60,50}
Gaussian noise	{0.01,0.008,0.006,0.004,0.002}

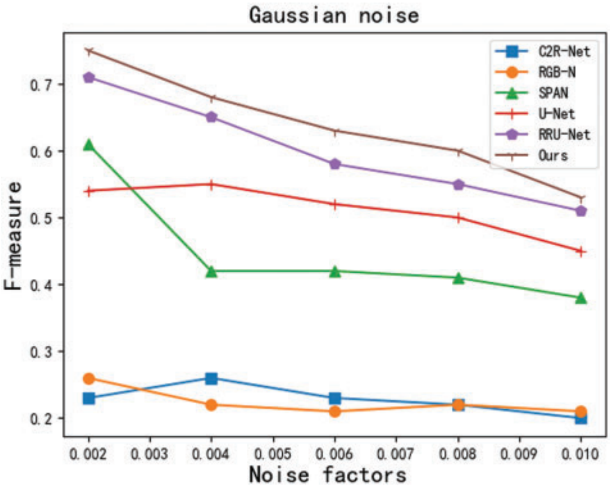

Figure 7: *F*-measure after adding Gaussian noise

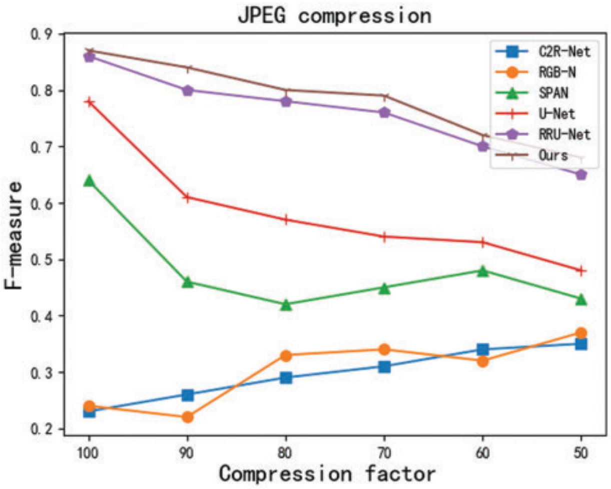

Figure 8: *F*-measure value after JPEG compression

Experimental details: The MSCU-Net detection method was run on a computer with an Intel(R) Core(TM) i5-10400F CPU and an NVIDIA A6000 GPU. In the training process of MSCU-Net, we use the cross-entropy loss function [29] and group normalization (GN) [30] to normalize the scattered data in high-dimensional space. We use random values for the initial parameters and stochastic gradient descent with a batch size of 10 samples, a momentum of 0.9, a weight decay of 0.0005, and an initial learning rate of 0.01.

5 Conclusion

In this paper, an image tampering detection algorithm based on U-shaped detection network is proposed. First, the suspected tampered regions in the image are detected using the U-shaped detection network. Then, to further optimize the detection results of the fine U-shaped network, the final detection results are obtained using the multi-feature extraction module and the coordinate channel confusion attention mechanism. To evaluate the effectiveness and practicality of the proposed algorithm in this paper, current deep learning-based detection algorithms are experimentally compared and experimentally visualized simultaneously on the tampering standard dataset, showing that the model pair in this paper has better detection and localization performance. In the task of tampering detection, it is important to extract image information at multiple scales, which can improve the detection ability of the model to a certain extent for targets in various sizes of regions and to facilitate the expression of image tampering features using attention modules. The experimental results show that the proposed algorithm in this paper outperforms several other comparative algorithms in detection.

Acknowledgement: The authors extend their appreciation to the National Natural Science Foundation of China.

Funding Statement: This work is supported in part by the National Natural Science Foundation of China (Grant Number 61971078) and Chongqing University of Technology Graduate Innovation Foundation (Grant Number gzlcx20222064).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. Chen, Y. Shi and W. su, "Image splicing detection using 2D phase congruency and statistical moments of characteristic function," *Electronic Imaging*, vol. 6505, pp. 65050R, 2007.

[2] W. Wang, J. Dong and T. Tan, "Effective image splicing detection based on image chroma," in *IEEE Int. Conf. on Image Processing (ICIP)*, Cairo, Egypt, pp. 1257–1260, 2009.

[3] X. D. Zhao, J. H. Li, S. H. Li and S. L. Wang, "Detecting digital image splicing in chroma spaces," *International Workshop on Digital Watermarking*, vol. 6526, pp. 12–22, 2011.

[4] H. Gou, A. Swaminathan and M. Wu, "Noise features for image tampering detection and steganalysis IEEE," in *Int. Conf. on Image Processing*, San Antonio, TX, USA, pp. VI-97–VI-100, 2007.

[5] Y. F. Hsu and S. -F. Chang, "Detecting image splicing using geometry invariants and camera characteristics consistency," in *IEEE Int. Conf. on Multimedia and Expo*, Toronto, ON, Canada, pp. 549–552, 2006.

[6] B. Mahdian and S. Saic, "Detection of resampling supplemented with noise inconsistencies analysis for image forensics," in *Int. Conf. on Computational Sciences and its Applications*, Perugia, Italy, pp. 546–556, 2008.

[7] M. K. Johnson and H. Farid, "Exposing digital forgeries in complex lighting environments in," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 450–461, 2007. https://doi.org/10.1109/TIFS.2007.903848

[8] S. Ye, Q. Sun and E. C. Chang, "Detecting digital image forgeries by measuring inconsistencies of blocking artifact," in *IEEE Int. Conf. on Multimedia and Expo*, Beijing, China, pp. 12–15, 2007.

[9] Z. H. Lin, J. F. He, X. Tang and C. K. Tang, "Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis," *Pattern Recognition*, vol. 42, no. 11, pp. 2492–2501, 2009. https://doi.org/10.1016/j.patcog.2009.03.019

[10] W. Luo, J. Huang and G. Qiu, "JPEG error analysis and its applications to digital image forensics," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 480–491, 2010. https://doi.org/10.1109/TIFS.2010.2051426

[11] T. Bianchi, A. de Rosa and A. Piva, "Improved DCT coefficient analysis for forgery localization in JPEG images," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, pp. 2444–2447, 2011.

[12] X. Wang, K. Pang, X. Zhou, Y. Zhou, J. Xue *et al.,* "A visual model-based perceptual image hash for content authentication," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1336–1349, 2015. https://doi.org/10.1109/TIFS.2015.2407698

[13] C. P. Yan, C. M. Pun and X. C. Yuan, "Quaternion-Based image hashing for adaptive tampering localization," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2664–2677, 2016.

[14] B. Liu and C. M. Pun, "Locating splicing forgery by fully convolutional networks and conditional random field," *Signal Processing: Image Communication*, vol. 66, pp. 103–112, 2018.

[15] X. Bi, Y. Wei, B. Xiao and W. Li, "RRU-Net: The ringed residual U-Net for image splicing forgery detection," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, pp. 30–39, 2019.

[16] P. Zhou, X. Han, V. I. Morariu and L. S. Davis, "Learning rich features for image manipulation detection," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1053–1061, 2018.

[17] Y. Wu, W. AbdAlmageed and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 9535–9544, 2019.

[18] X. F. Hu, Z. H. Zhang, Z. Y. Jiang and S. Chaudhuri, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Computer Vision—ECCV 2020*, European Conference, Glasgow, UK, pp. 312–328, 2020.

[19] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image xegmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, Springer International Publishing, 2015.

[20] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[21] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.

[22] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conf. on Computer Vision*, Cham, Springer, pp. 3–19, 2018.

[23] Q. L. Zhang and Y. B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *ICASSP 2021—2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, pp. 2235–2239, 2021.

[24] C. Szegedy, S. Ioffe and V. Vanhoucke, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2016. https://doi.org/10.1609/aaai.v31i1.11231

[25] A. G. Roy, N. Navab and C. Wachinger, *Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks*. Cham: Springer, 2018.

[26] J. Dong, W. Wang and T. Tan, "CASIA image tampering detection evaluation database," in *2013 IEEE China Summit and Int. Conf. on Signal and Information Processing*, Beijing, China, pp. 422–426, 2013.

[27] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates *et al.,* "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, pp. 63–72, 2019.

[28] Y. Wei, X. Bi and B. Xiao, "C2R Net: The coarse to refined network for image forgery detection," in *IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications/12th IEEE Int. Conf. on Big Data Science and Engineering (TrustCom/BigDataSE)*, New York, NY, USA, pp. 1656–1659, 2018.

[29] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980. https://doi.org/10.1109/TIT.1980.1056144

[30] Y. Wu and K. He, "Group normalization," in *Computer Vision—ECCV 2018*, Munich, Germany, pp. 3–19, 2018.