ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.009

Image Tampering Localization based on Pyramid Multi-Scale Pooling Module

Caixia Meng[1,2,3], Bing Zhao[1,2], Kaijie Xi[1], Jiabao Zhang[1], Hongpeng Chu[1]

[1]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications

[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, China

[3]Xi 'an Key Laboratory of Big Data and Intelligent Computing, China

Email: mcxmcx@xupt.edu.cn, 1595819159@stu.xupt.edu.cn,{501247765, 392455199, 2325902720}@qq.com

Abstract. In this paper, we proposed a new two-stream image forgery localization method based on pyramid multi-scale pooling module. One stream learns the differences in JPEG compression rate between original regions and tampered regions by processed through ELA. In the second stream, the network learns the global image manipulation features from the input image RGB values. Next the fusion stage fuses multiple tampering features from the dual-channel feature extraction network to generate the shallow feature map, Coordinate Attention is applied to refine the predicted mask and further improve the localization accuracy. And then through the multi-scale pyramid pooling module, so the secondary fusion is carried out. Finally, the fused feature map restores the original image size by using UpSampling for Pixel level classification prediction. The proposed net can effectively decrease incorrect prediction since it makes better use of the contextual spatial information in images. And most of all, from two stages extraction fuses low-level and high-level information to refine the global manipulation features. The experimental results on several public datasets show that the proposed model outperforms some state-of-the-art methods.

Keywords: Forgery Localization, Pyramid Multi-Scale Pooling, Feature Fusion, Coordinate Attention, Pixel-level Classification Prediction.

1. Introduction

With the advances of artificial intelligence technology, image editing technology has evolved from using manual modification of Photoshop, GIMP, ACDsee and other editing software to using AI technology to achieve intelligent and automated image tampering. For example, in recent years, Generative Adversarial Networks (GAN) have been used to synthesize high-resolution false images. With the maturity of deep forgery technology, not only the process of image tampering is fast and convenient, but also the modification methods are ever-changing. After inserting the splicing area, post-processing operations such as blurring, smoothing, retouching and fusion will be used to cover up the traces of tampering, making the tampered image looks more realistic and natural. It can be seen that the authenticity of digital media content can not to be guaranteed, and its authenticity has been questioned more and more, multimedia authentication has become a research hotspot and difficulty in the field of information security.

At present, the issue of media forgery has become the focus attention of researchers. The purpose of forgery may be for entertainment (such as using Meitu Xiuxiu and other tools to beautify photos, AI face-changing technology, etc.) or maliciously changing the content of the image (such as deliberately modifying photos of important documents or exaggerating the severity of news events) and so on. In recent years, various incidents of multimedia forgery that have appeared also constantly remind people to pay attention to the security of media content. The example shows in Fig. 1, forged images such as "South China Tiger", "Square Peace Dove", "Qinghai Tibet Railway", "Tibetan Antelope" of Qinghai-Tibet Railway and "Baiyun Mountain Snowscape" of Guangzhou have seriously misled people's cognition. There are also numerous examples of multimedia forgery in the world, American field reporters' reports on the Iraq war using splicing photos has aroused public distrust, and Iran's photos of tampering with missile launches pose a threat to world security. Multimedia forgery has been involved in many fields such as politics, science, news, war and entertainment. In addition, the European Union issued the "ETHICS GUIDELINES FOR TRUSTWORTHY

AI" on April 8, 2019, which regards privacy and data management as one of the seven elements that reliable artificial intelligence needs to meet, and countries have also strengthened their supervision of this technology. At the same time, psychological research also shows that about 30% of people will be deceived by false information, which will seriously affect the public's view of things, and may even cause serious consequences.

Fig.1: Forged images shared over social media platforms:the left is "South China Tiger" event and the right is "Iran's photos of tampering with missile launches" event

2. Related Works

In the early digital image forensics research, researchers mainly used manual feature extraction algorithms to extract fingerprints that characterize the inherent features of the image, and compared it with the image fingerprints of the known devices. So as to identify the originality and integrity of the image to be detected, such as local noise analysis [1], CFA artifacts [2], illumination variance analysis [3], and double JPEG compression [4].

Adobe, the company of Photoshop, with Berkley jointly proposed a two-stream Faster R-CNN network [1] to detect the tampered regions given a manipulated image. Subsequently, there has been a wide focus on using deep learning technology to solve the problem of image tampering localization. The current detection methods for tampered images and deeply forged images mainly rely on the combination of feature extraction methods and deep learning technology, the hidden features of tampered and forged images are extracted through a variety of methods, and then detected and recognized through the learning of artificial neural network. RGB-N[5] adopts the steganalysis rich model and Faster R-CNN, but it can only provide bounding boxes instead of segmentation masks. ManTra-Net [6] learns features to distinguish 385 known manipulation types and treats the problem as anomaly detection, however, its fail to take full advantage of the spatial correlation and consequently have limited generalizability [7]. Bi proposed an end-to-end RRU-Net [8] module to achieve the splicing detection without any preprocessing and post-processing, the core idea of this approach was to improve the learning of CNN by the propagation and the feedback process of the residual in CNN [9]. However, in most of these approaches, the shallow characteristic information is often ignored, or only the high-frequency information is explored for detecting the manipulation, and the contextual spatial information is lost.

In the massive data analysis scenario, the end-to-end forensics model simplifies the complex process of the analysis model and makes data processing more efficient. In recent years, with the success of deep learning in various computer vision tasks, such as object detection [5, 10] and semantic segmentation [11, 12], many methods based on deep learning have been developed for image tampering detection and localization [13], the efficiency of neural network models in massive data processing and the accuracy of digital image fingerprint extraction have been continuously highlighted. Therefore, a large number of multimedia forensics analyses began to adopt end-to-end forensics framework based on the deep learning models. However, in practice, when facing unknown images in practical applications, the type of image tampering cannot be known in advance preventing the wide practical application of these restricted manipulation localization approaches. So how to find a lightweight network that not only meets the requirements of tampering localization accuracy, but also has universal adaptability is the top priority of many scholars.

For overcoming the drawbacks of current CNN-based detection methods. In this paper, we propose a novel image forgery detection approach that can automatically learn feature representations based on a lightweight deep learning framework. The primary contributions are summarized as follows: 1) we design a

dual-channel network that learns both the RGB features and images processed by ELA features for pixel-level forgery localization, so that our net fuses multiple tampering features and enables the localization of various types of manipulations, improve the generalization ability of models. 2) Two Stage feature fusion, Stage-1, in order to reduce the amount of model parameters, we using MobilenetV1 extracted above dual-channel shallow feature information respectively, then fuses two-channel information as Feature layer5(F5) to enhance the global feature representation and identify the manipulated regions for latter coarse localization. Stage-2, As prior knowledge, Stage-1 we extract the F5 to guide Stage-2, so we using pyramid multi-scale pooling module strengthen the contextual spatial information feature extraction network for final tampered region localization, this is the second stage feature fusion. 3) In order to further improve the accuracy of locating splicing tampered regions, we add the attention mechanism into the MobilenetV1.

3. Materials and Methods

In this section, we present the overall model architecture consists of three parts shows in Fig. 6, which include Feature preprocessing, the coarse manipulation localization (Stage-1: Shallow feature fusion extraction network), the fine manipulation localization (Stage-2: Strengthen feature fusion extraction network).

3.1. Feature preprocessing

For lossy images, any digital modification will make the modified area become unstable, and the Error Level Analysis (ELA) is for the lossy compression of JPEG images, by intentionally re-saving the image at a known error rate and then computing the difference between the images [14]. For JPEG Images, the entire picture will show the same error level potential, if the error level of a certain part of the image is obviously different, it may indicate that it has been digitally modified. ELA highlights the difference in JPEG compression rate. So in the ELA channel, we convert images of various storage formats into JPEG format. The steps of ELA are summarized below:

1) read in image as JPEG.

2) write image as JPEG with Quality lower level (e.g., 90)

3) read in compressed image (decompress)

4) take absolute value of the difference between the decompressed image in step 3) and the original image in step 1)

Examples of ELA results are shown in Fig. 2. Even with careful inspection, humans find it difficult to recognize the tampered regions. And after ELA processing, it can be observed that the compression difference between the tampered area and the non-tampered area is obviously, also compared with the ground truth mask, it shows that ELA can extract the features of the tampered image area very well.

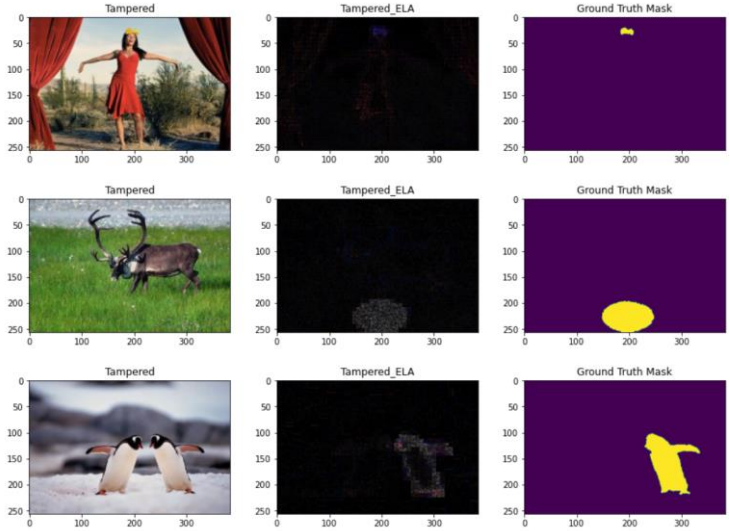

Fig. 2: Three tampered images (left) with their corresponding ELA processing images (medium) and ground-truth masks (right)

3.2. Stage-1: Shallow feature extraction network(MobilenetV1+Coordinate Attention)

Although the ELA captures rich manipulation clues, but some types of tampering still not to be identified well, so we extract the two-channel information which are ELA channel and RGB channel respectively, then fuses them to enhance the global feature representation and identify the manipulated regions for latter coarse localization.

In order to extract features well and reduce the amount of model parameters, we select the MobilenetV1 as the shallow feature extraction network. The MobileNetV1 model is based on Depthwise Separable Convolutions (show in Fig. 3) which is a form of factorized convolutions which factorize a standard convolution into a depthwise convolution and a 1×1 convolution called a pointwise convolution [15]. By this way can drastically reducing computation and model size.

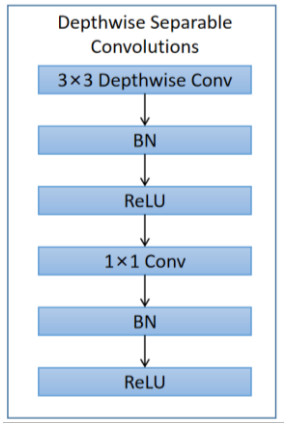

Fig. 3: Depthwise Separable Convolutions structure

Visual attention mechanism let feature extraction network pay more attention to a target (tampered regions) and suppress other useless information. However, the computational overhead brought by most attention mechanisms is not affordable for mobile networks, the most popular attention mechanism for mobile networks is still the SE attention and CBAM, but they generally neglect the positional information, which is important for generating spatially selective attention maps. Therefore, we adopt the Coordinate Attention (CA) [16], the CA block shows in Fig. 4. CA by embedding positional information into channel attention to enable mobile networks to attend over large regions while avoiding incurring significant computation overhead.

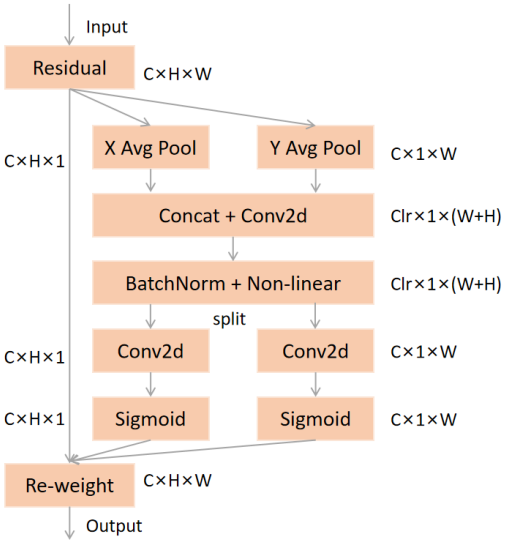

Fig. 4: Coordinate Attention Block

Both ELA channel and RGB channel through this network structure extract shallow features, the specific network is show in Fig. 5. Such as ELA channel, when we input a 576×576×3 image which are processed by ELA into the MobilenetV1, we could extract five feature layers through the multiple Depthwise Separable Convolutions, and add attention mechanism after the last three Feature layers (F3, F4, F5).

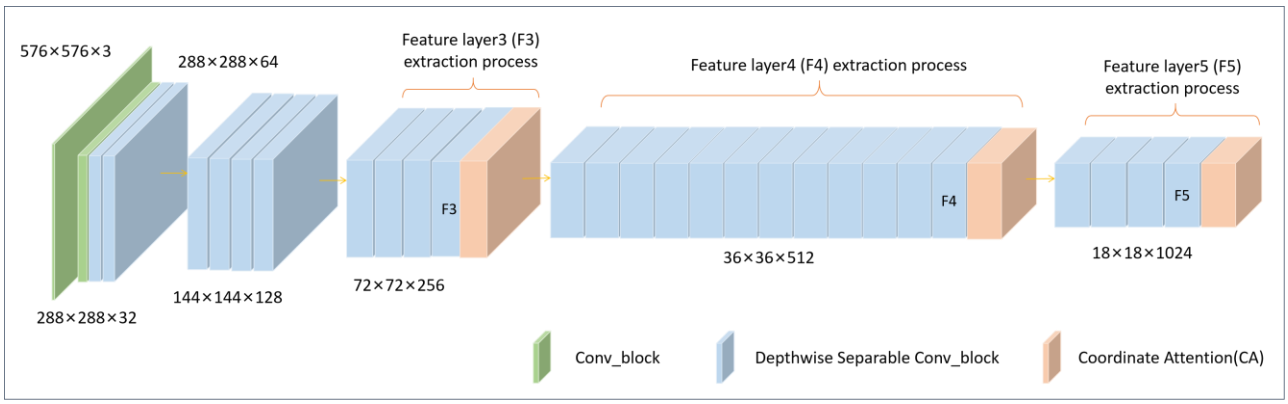

Fig. 5: Shallow feature extraction architecture

3.3. Stage-2: Strengthen feature extraction network(pyramid multi-scale pooling model)

In Stage-2 (Strengthen feature extraction network), the coarse localization information from Stage-1 (Shallow feature extraction network) guides the model to further refine local contextual spatial information features and performs manipulation localization on pixel level. The overall process is shown in the Fig. 6.

When we integrate feature maps information, the common operations are CONCAT, however, in order to let shallow feature extract network better combined the content information of the two channels and reduced the missing localization rate, we use ADD replace Contact to fuse two-channel information. ADD means the amount of information under each dimensional feature increases, but the dimension of the image does not increase. At the same time, compare with the CONCAT, ADD reduce a half number of parameters and calculations. Therefore, we using ADD operation to fuse the F5 from the MobilenetV1 of the two channels, this step show in Figure 6(c).

Usually, Many Copy-Move localization errors due to similar appearance of objects, and higher-layer feature contains more semantic meaning and less location information [17]. Therefore, we exploit the capability of global context information by different-region-based context aggregation through the pyramid multi-scale pooling module[23], extend the pixel-level feature to the global pyramid pooling. The local and global clues together make the final prediction more reliable. As show in Fig. 6(e), we adopted four different pyramid features scale, which are 1×1, 2×2, 3×3 and 6×6 size, and use the 1×1 convolution reduced the number of channels, next the low-dimensional feature map is directly UpSampled by bilinear interpolation, then different levels of features are concatenated as the final global feature of pyramid pooling. Finally, the final global feature map resizes the images shape to the original image size for pixel level prediction that generate two categories for each pixel (tampered and authentic regions).

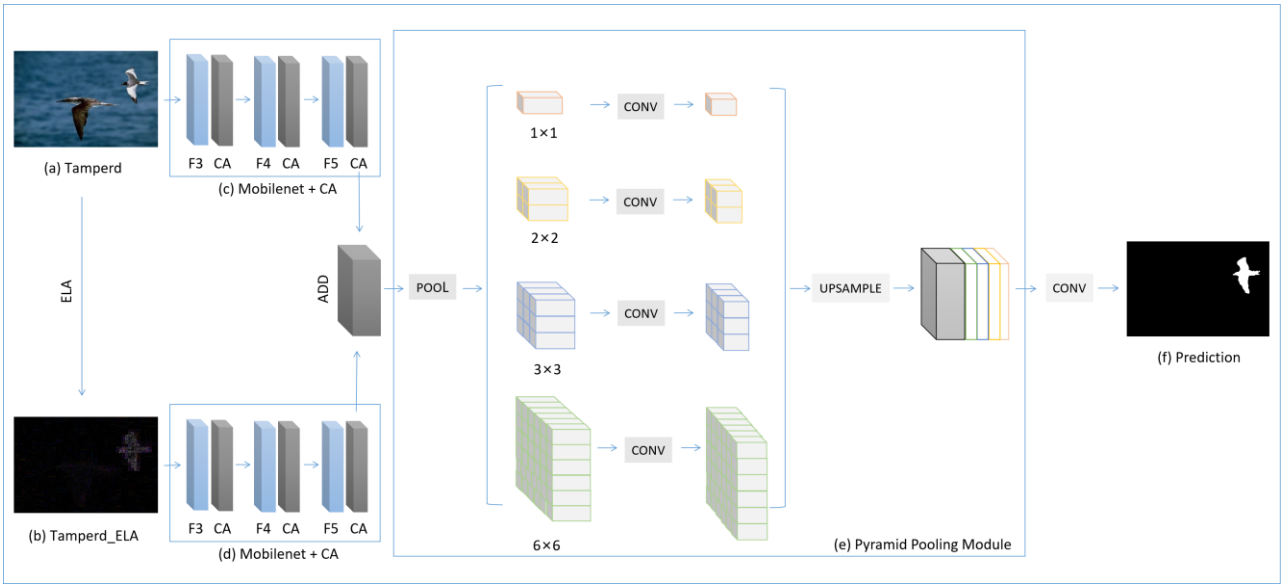

Fig. 6: Overview of the proposed architecture for image manipulation localization

4. Experimental Setting

4.1. Datasets

1) CASIAV1.0[18]: The tampered set contains 921 tampered images of size 384×256 pixels. The Corel Images can be roughly clustered into 8 categories according to image content (scene, animal, architecture, character, plant, article, nature and texture). The tampered images are generated by using crop-and-paste operation under Adobe Photoshop on these authentic images.

2) CASIAV2.0[18]: The tampered set contains 5123 tampered images ranging from 320×240 to 800×600 pixels. And besides splicing, in database V2.0 introduce blurring when manipulating the tampered image set.

3) IFS-TC: The set comprises several original images captured from different digital cameras with various scenes either indoor or outdoor. The "forged" images contain 450 tampered images and comprise a set of different manipulation techniques such as copy/pasting and splicing with different degrees of photorealism.

Based on the above data set, we selected 90% splicing tampered images as experimental data, on this basis, separate 5% from experimental data as the validation set, and others as test set.

4.2. Details

The proposed net is implemented using Tensorflow2.3 and conducted on a RTX2060MaxQ GPU. When set 50 epochs and every epoch size is 4, the time of every epoch is typically less than 10 minutes on our PC with AMDR9-4900HS CPU and 16GB RAM on CASIAV2 datasets. In addition, the network parameters are learned in the training process by the binary cross-entropy between the ground-truth and the predicted masks over a mini-batch of images. We train the network with Adam setting the initial learning rate to 1e-4, early stopping is introduced to avoid overfitting.

4.3. Evaluating indicator

In order to make a quantitative evaluation of the experimental results of this model, we use Accuracy, Precision, Recall, F1 score and AUC to evaluate the performance of the proposed splicing forgery localization methods in pixel level. Precision means how many of the predicted positive samples are true positive samples. Recall means how many positive examples in the sample are predicted correctly. F1 score is harmonic average of accuracy and recall. AUC is the area under ROC curve, and the FPR (False Positive Rate) is the horizontal axis of ROC curve, TPR (True Positive Rate) is the vertical axis of ROC curve. The higher all these indicators mean the better model performance. Above these are formulated in Eq. (1)(2)(3)(4)(5)(6) respectively.

$$\text{Accuracy} = \frac{n_{correct}}{n_{total}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

$$\text{TPR} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{6}$$

where TP (True Positive) denotes the number of correctly detected tampered pixels, FP (False Positive) denotes the number of incorrectly detected tampered pixels, FN (False Negative) denotes the number of

incorrectly detected un-tampered pixels and TN (True Negative) denotes the number of correctly detected untampered pixels.

5. Results

In this Section, experiments are carried out to demonstrate the effectiveness of our proposed a new two-stream image forgery localization method based on pyramid multi-scale pooling module.

Firstly, we compare the effect of adding attention mechanism to which layer is better. Through experimental comparison, it is found that adding CA behind the Feature3, Feature4 and Feature5 layer is better than adding attention mechanism to Depthwise Separable Convolutions. The prediction results show in fig. 7. It is obviously showed the CA divide the edge more finely, further render the final result nearly the same as the ground truth.

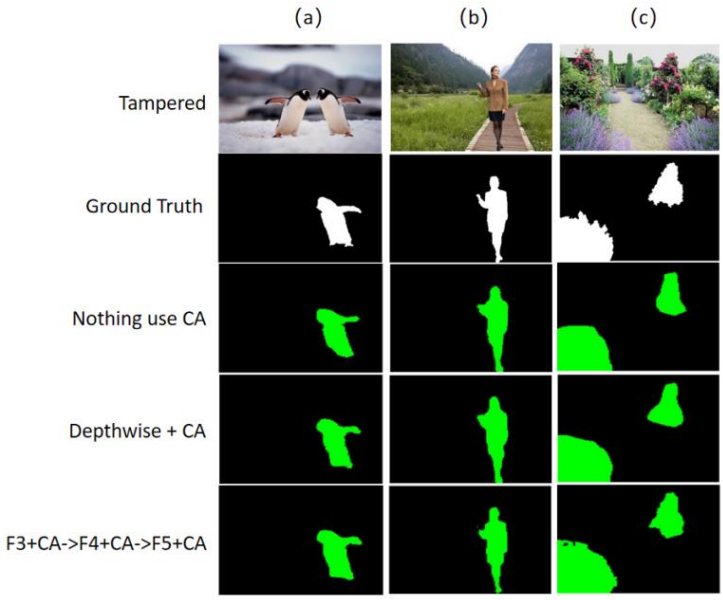

Fig.7: Localization performance of the proposed method with and without and where to add CA

There are generally more no tampered pixels than the tampered ones in a forged image. Many methods computed as the average over all the pixels, will be more biased towards the no tampered classes. Consider imbalances num in each category, we only describe the score of the predicted tamper regions. We also compare our method with other several state-of-the-art image forgery detection methods on public datasets. Finally, we choose three traditional feature extraction-based detection methods and several deep learning networks. TABLE Ⅰ shows the localization performance comparison of the proposed scheme with other methods on CASIAV2. It can be easily observed that ours is better than the others methods in F1-score, and AUC is slightly lower than the LSTM-EnDec-Skip, but better than the other methods.

TABLE I. Pixel level localization performance comparison of the proposed scheme with other methods

	F1-score	**AUC**
ELA [14]	0.214	0.613
NOI1 [22]	0.263	0.612
CFA1 [2]	0.207	0.522
RGB-N [5]	0.408	0.795
LSTM-EnDec-Skip [21]	0.432	**0.814**
CONSTRAINED R-CNN [20]	0.475	0.789
HybridArch [19]	0.525	-
ours	**0.543**	0.795

We compared the result between three public datasets in Accuracy, Precision, Recall, F1-score, AUC, respectively, the weight of the score calculated for each category is no longer 1/category, but the proportion

of each category calculate the score of the predicted regions, this means the weighted, the specific results show in TABLE II.

TABLE II. Pixel level localization performance comparison of the proposed method in different datasets

	Accuracy	Precision	Recall	F1-score	AUC
CASIA V1	0.8927	0.9053	0.8927	0.8813	0.5922
CASIA V2	0.9527	0.6139	0.5646	0.5428	0.7953
IFS-TC	0.9083	0.8537	0.9083	0.8692	0.5125

To further prove the effectiveness and robustness of our net, some visualizations are given in three public datasets. The prediction shows in Figure 8. The following can be observed from the figures: Our experiments are performed on 3 public datasets and demonstrate the effectiveness of our proposed approach for image forgery localization.

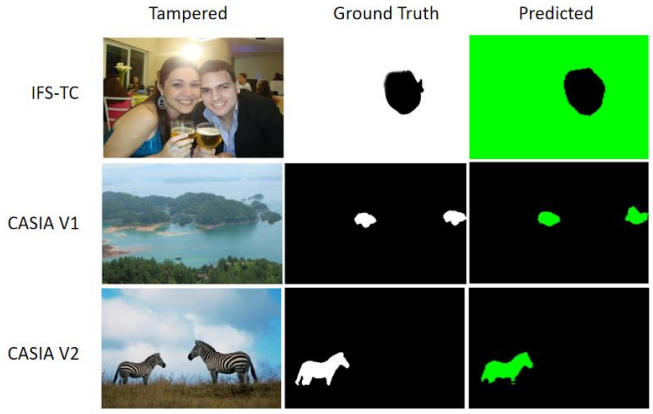

Fig. 8: Visual comparison in three public datasets

6. Conclusions And Discussions

Extensive experiments on several public datasets have been carried out, which demonstrates the superior performance of the proposed a new two-stream image forgery localization method based on pyramid multi-scale pooling module over other state-of-the-art image forgery detection methods and subsequently improve the integrity and accuracy of pixel level tampering localization. The generalization ability of the model is improved by the first feature fusion, and the global context information is collected through the second multi-scale feature fusion. Coordinate Attention is applied to refine the predicted mask and further improve the localization accuracy. Simultaneously, ELA is complementary to the RGB information, we also prove the validity of fusing two-channel in from experimental comparison in three different datasets.

7. Acknowledgement

This work was supported by the Special Construction Fund for Key Disciplines of Shaanxi Province Higher Education, the Natural Science Foundation of Shaanxi Province (Program No.2014JM8303).

8. References

[1] Cozzolino D, Gragnaniello D, Verdoliva L. Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques[C]//2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014: 5302-5306.R. Dewri, and N. Chakraborti. Simulating recrystallization through cellular automata and genetic algorithms. *Modelling Simul. Mater. Sci. Eng*. 2005, **13** (3): 173-183.

[2] Ferrara P, Bianchi T, De Rosa A, et al. Image forgery localization via fine-grained analysis of CFA artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(5): 1566-1577.

[3] De Carvalho T J, Riess C, Angelopoulou E, et al. Exposing digital image forgeries by illumination color classification[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(7): 1182-1194.

[4] Barni M, Bondi L, Bonettini N, et al. Aligned and non-aligned double JPEG detection using convolutional neural networks[J]. Journal of Visual Communication and Image Representation, 2017, 49: 153-163.

[5] Zhou, Peng, et al. "Learning rich features for image manipulation detection." Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition. 2018.

[6] Wu Y, AbdAlmageed W, Natarajan P. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9543-9552.

[7] Liu X, Liu Y, Chen J, et al. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization[J]. arXiv preprint arXiv:2103.10596, 2021.

[8] Bi X, Wei Y, Xiao B, et al. RRU-Net: The ringed residual U-Net for image splicing forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[9] Gao Z, Sun C, Cheng Z, et al. TBNet: Two-Stream Boundary-aware Network for Generic Image Manipulation Localization[J]. arXiv preprint arXiv:2108.04508, 2021.

[10] Yang C, Li H, Lin F, et al. Constrained R-CNN: A general image manipulation detection model[C]//2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020: 1-6.

[11] Chen H, Chang C, Shi Z, et al. Hybrid features and semantic reinforcement network for image forgery detection[J]. Multimedia Systems, 2021: 1-12.

[12] Zhang R, Ni J. A dense u-net with cross-layer intersection for detection and localization of image forgery[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 2982-2986.

[13] Zhuang P, Li H, Tan S, et al. Image Tampering Localization Using a Dense Fully Convolutional Network[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2986-2999.

[14] Krawetz N, Solutions H F. A picture’s worth[J]. Hacker Factor Solutions, 2007, 6(2): 2.

[15] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[16] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13713-13722.

[17] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[18] Dong J, Wang W, Tan T. Casia image tampering detection evaluation database[C]//2013 IEEE China Summit and International Conference on Signal and Information Processing. IEEE, 2013: 422-426.

[19] Zhang Y, Zhang J, Xu S. A hybrid convolutional architecture for accurate image manipulation localization at the pixel-level[J]. Multimedia Tools and Applications, 2021: 1-16.

[20] Yang C, Li H, Lin F, et al. Constrained R-CNN: A general image manipulation detection model[C]//2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020: 1-6.

[21] Mazaheri G, Mithun N C, Bappy J H, et al. A Skip Connection Architecture for Localization of Image Manipulations[C]//CVPR Workshops. 2019: 119-129.

[22] Mahdian B, Saic S. Using noise inconsistencies for blind image forensics[J]. Image and Vision Computing, 2009, 27(10): 1497-1503.

[23] Yoo D, Park S, Lee J Y, et al. Multi-scale pyramid pooling for deep convolutional representation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 71-80.