A Dual-Stream Input Faster-CNN Model for Image Forgery Detection

Lizhou Deng[1], Ji Peng[2], Wei Deng[3], Kang Liu[1], Zhonghua Cao[1], and Wenle Wang[1(✉)]

[1] School of Software, Jiangxi Normal University, Nanchang 330027, Jiangxi, China
wenlewang@jxnu.edu.cn
[2] College of Information and Computer Engineering, Pingxiang University, Pingxiang 337055, Jiangxi, China
[3] School of Intercultural Studies, Jiangxi Normal University, Nanchang 330022, Jiangxi, China

Abstract. With the development of multimedia technology, the difficulty of image tampering has been reduced in recent years. Propagation of tampered images brings many adverse effects so that the technology of image tamper detection needs to be urgently developed. A faster-rcnn based image tamper localization recognition method with dual-flow Discrete Cosine Transform (DCT) high-frequency and low-frequency input is presented. For capturing subtle transform edges not visible in RGB domain, we extract high-frequency features from the image as an additional data stream embedding model. Our network model uses low-frequency images as the subject data to detect object consistency in different regions, further complements high-rate streams to strengthen image region consistency detection, and complements duplicate stream object tampering detection. Extensive experiments are performed on the CASIA V2.0 image dataset. These results demonstrate that faster-rcnn-w outperforms existing mainstream image tampering detection methods in different evaluation indicators.

Keywords: Image Forgery Detection · Faster-CNN · Dual-Stream Input

1 Introduction

Images have been widely used as carriers of information for rapid dissemination in the network, but with that comes the test of integrity and authenticity of images. With development of GAN [1–3] and the popularity of various PS tools, the threshold for tampering and forging fake pictures with no visual traces has been greatly reduced, and retouching has become relatively simple, but at the same time, many tampered pictures have been used to spread rumors, fabricate false news, and illegally seek benefits and other problems. Thus it can be seen that image tampering detection technology is particularly important, and there is a growing need for this new technology in society and the general public. The current research in digital image tampering is not mature enough, however, especially the research on multiple tampering detection. As a result, research on digital image tampering detection technology is of great significance. There are three ways to divide the type of image alteration.

Y. Cao and X. Shao (Eds.): MONAMI 2022, LNICST 474, pp. 105–115, 2023.
https://doi.org/10.1007/978-3-031-32443-7_7

1. Copy-move [4, 5]: By copying a region on the image, move the copied region to a location other than the copied one.This method applies to a single image.
2. Splicing [6–8]: Image splicing copies regions from a genuine image and pastes them to other images.This method applies to multiple images.
3. Removal eliminates [9]: regions from an authentic image followed by inpainting. This method applies to a single image.

In Fig. 1, tampered image areas are mostly item objects in order to enhance the tampering reliability, i.e., objects are added or removed in the tampering.

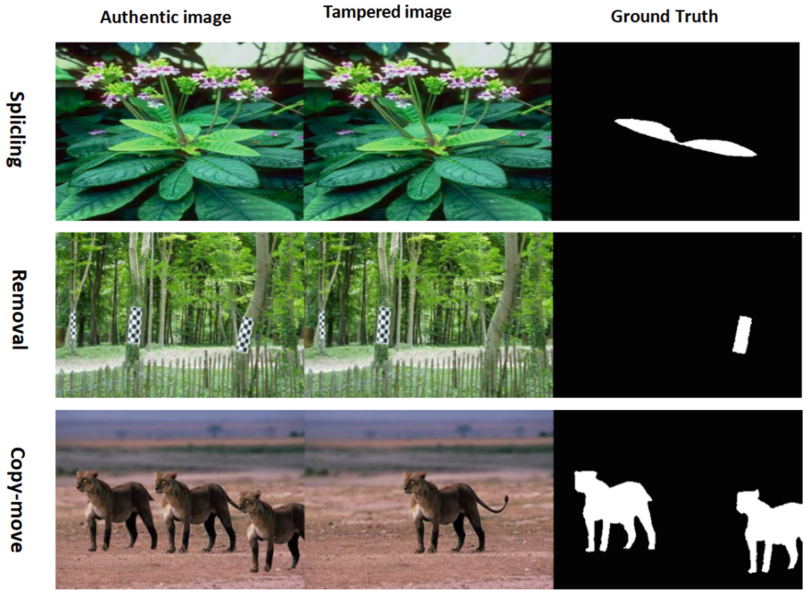

Fig. 1. Three ways of image tampering.

According to the literature [11], the images are transformed from the rgb domain to the frequency domain by Discrete Cosine Transform(DCT) then the high(or low) frequency component in the frequency domain are obtained as the input stream by high(or low) pass filter. The high-frequency component streams are further connected with the low-frequency component streams to complement each other. Motivated by this, in [12] we applied a dual-stream input model, where low-frequency images and high-frequency images are inputted separately at the input and fed to the network for training. Two channels are simultaneously trained to exchange training features and learn features in different frequency domains to accomplish consistency across objects, and cross-focus on both channels to back-propagate the enhanced algorithm. By iterating the object features, a global feature representation is achieved for later use in detection operations. Last, the detection region size is synchronously mapped to the original image using the consistency of the image position. After extensive experiments on publicly available

datasets, we show that our faster-rcnn-w method outperforms most other methods in terms of evaluation metrics.

Our project currently contributes to the following:

(1) We innovatively ignore the rgb stream capture, but identify tamper artifacts by combining high-frequency transform with low-frequency features. This improves the prediction accuracy and reduces the speed of the coupled acceleration network.
(2) We make model improvements by performing a dual transformation on the input side and using a fusion pooling layer to reinforce the connection between the two channels. For the effect visualization, the results are mapped to the original image using image coherence with the tampered low frequency image position as a reference.
(3) We performed extensive experiments in multiple benchmark tests to demonstrate the advanced detection and localization properties of our method.

2 Related Works

In most image tampering detection, a single normal rgb stream image [13] is used as training object when image processing. Invisible image features and the intense intensity transform part often imply more tamper information [14]. This information is not visible in the rgb domain, so we propose a, multi-frequency domain joint modal approach and based on the faster-rcnn network model to find the image intense transformation information to achieve higher tampering recognition accuracy. At present, a lot of groundbreaking achievements have been made in image forgery detection. In the deep learning domain, in 2016 Bayar et al. Krishna et al. [15] innovated a novel convolutional layer structure to capture the correlation changes of adjacent pixels in the image when the image is tampered, and at the same time adaptively learn tampered features, and compress image content as much as possible. The detection effect. In 2017, a passive image forensics algorithm based on deep learning was proposed in the literature [10]. The image feature extraction part used CNN to learn features [16], and introduced the rich spatial model to initialize network parameters. Once the feature fusion is done, select The optimal features are used for classification and localization of tampered images and achieve high accuracy on public datasets. RGB-N [12] introduced a two-stream network for operation localization in 2018, where one stream extracted RGB features to capture visual artifacts, while the other exploited noise features to model the difference between tampered and unmodified regions. Inconsistencies between the two. It is the first method to use the dual-stream input model to complete the image detection problem, which further improves the tamper detection accuracy.

Image forgery is particularly critical in the medical field, in [17], We learned that in the Chest X-ray is a kind of medical image, [17] proposes a model for the detection of Chest-X-ray by Multi Attainment and Incorporating Background Information Model, Model focuses on how to improve the performance of decoders, for example, combining retrieval and generation for the template characteristics of the report. For image decoders, more mature convolutional neural networks (CNN) such as ResNet and DenseNet are used to extract features. We can learn from this that deep neural networks have a unique

advantage in processing the abnormal part of the image. This aligns with the idea of identifying areas with image tampering.

3 Method

We aim to improve the dual-stream input network, input the network with dual Discrete Cosine Transform (DCT) [11] high-frequency and low-frequency images, improve prediction accuracy, and map the low-rate detection frame to the original image via the original input image when the result is returned. Figure 2 shows the steps of our approach. This chapter explains the three parts of preprocessing, Input and Model refinement.

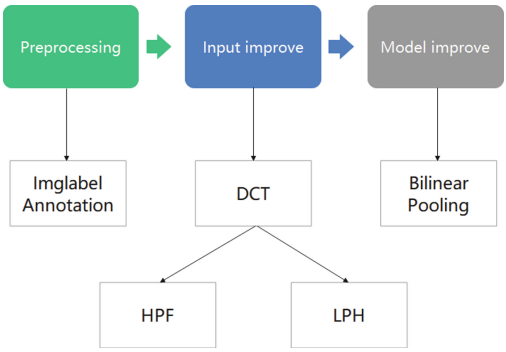

Fig. 2. Improvement steps of model.

Regarding the improvement of the model, namely as faster-rcnn-w, we will introduce three aspects: high-frequency image feature extraction, low frequency feature extraction and bilinear pooling. The structure of faster-rcnn-w model is shown as Fig. 3.

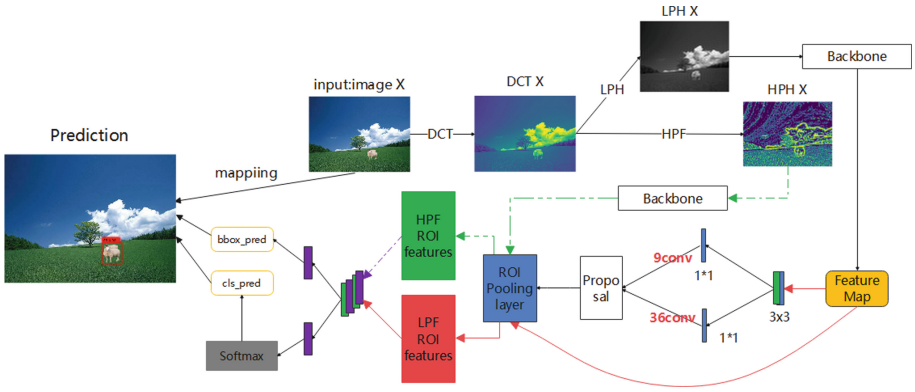

Fig. 3. Structure of faster-rcnn-w model.

3.1 High and Low Frequency Feature Extraction

Because manipulated images are often post-processed to hide tamper artifacts, capturing subtle tamper traces in RGB space is challenging. Thus, we extract features from the frequency domain to provide complementary cues for action sensing.

If an image X is taken as input, it is first converted from RGB to frequency domain using discrete cosine transform (DCT). The image processing process is shown in Fig. 4.

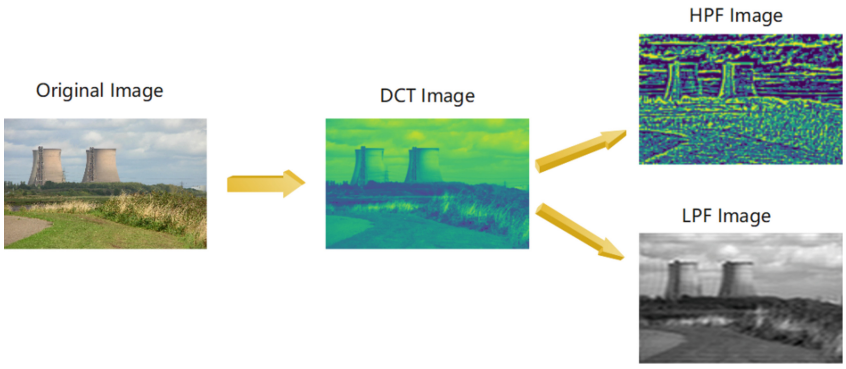

Fig. 4. Image Processing Flow

As there are no tamper marks hidden in normal images, capturing noise and image details in normal rgb space is challenging. With the goal of extracting more image features, we input the original image X of Fig. 3, first perform Discrete Cosine Transform (DCT) to convert It is transformed from the rgb domain to the Frequency Domain:

$$X = DCT(X) \tag{1}$$

where $X \in R^{H \times W \times 3}$.

The high frequency components are then obtained by a high pass filter to preserve the displacement invariance and local consistency of natural images:

$$DCT\ X = D^{-1}(H(X, a)) \tag{2}$$

where H is a high-pass filter, and a is a manually designed threshold that controls the low-frequency.

Similarly, we need to use the low frequency filter to let the DCT image enter and get the low frequency component:

$$DCT\ X = D^{-1}(L(X, b)) \tag{3}$$

where L is a low-pass filter, and b is a manually designed threshold that controls the high-frequency.

As shown in Fig. 4, we can see that the original image is converted into a DCT image, and the entire process of converting the DCT image into high frequency and low frequency.

Then, in Fig. 3, the two input streams enter the backbone pure convolutional layer at the same time to extract features and then divert the streams. The low frequency needs to go through the RPN network and then enter the ROI Pooling layer, while the high frequency image directly enters the ROI Pooling layer and joins the low frequency.

3.2 Bilinear Pooling

The bilinear pooling layer [16, 18] fuses the information of the low-frequency image channel and the high-frequency image channel. The output of the compressed bilinear pooling layer is as follows:

$$x = f_l^T * f_h \tag{4}$$

where f_l represents the location-sensitive map features of the low-frequency image channel, and f_h represents the location-sensitive map features of the high-frequency image channel. The recombined bilinear feature vector will be used as the basis for subsequent scores to determine whether the region has been tampered with.

The classification of tampered regions is evaluated using a cross entropy loss, and a smooth loss function [19] is used to evaluate bounding box regression. Lastly, the overall model loss function is:

$$L_a = L_r + L_c(f_l, f_h) + L_b(f_l) \tag{5}$$

Among them, L_a represents the total loss of the model, L_r represents the RPN network loss function, L_c represents the final cross-entropy classification loss, which is jointly determined by the dual-channel features f_l and f_h through the bilinear pooling layer, and L_b represents the final Bounding box regression loss, which is determined only by features fl from low-frequency image channels.

4 Experimental

This section evaluates the performance of faster-rcnn-w, conducts experiments on multiple tampered image datasets, uses a variety of evaluation criteria to obtain experimental data and compares with previous popular models, draws experimental conclusions.

4.1 Data Set Selection

The dataset contains the original image, the tampered image, and the corresponding master image. Existing image tampering detection datasets: CASIA dataset, RTD dataset, Columbia Uncompressed Image Splicing Detection dataset, Coverage dataset.

(1) CASIA 2.0, including two types of tampering, copy-move and splice
(2) Columbia Uncompressed Image Splicing Detection, this dataset only contains splice tampering, the dataset is small, only 183 tampered pictures, and the image resolution is high.

(3) The Coverage dataset only contains copy-move tampering, including 100 pairs of tampered images and the original image, and the image resolution is generally.
(4) RTD dataset, including three types of copy-move, splice, and remove, and the resolution is high.

For these four datasets, we have divided the training and test sets. After taking 90% of each dataset as training set and 10% as test set, we set the RPN network preselection rules, and anchors are finally determined as 16, 32, 64, 128. As the surface generation network filters anchors, IoU threshold of positive samples (samples containing tampered areas) is set to 0.7, and IoU of negative samples (those samples that do not contain tampered regions) is fixed to 0,3. According to the NMS (Non-Maximum Suppression) algorithm, the Samples with an IoU value greater than 0.7 are classified as foreground samples (tampered features), samples with an IoU value less than 0.3 are classified as negative samples (non-tampered features), and samples that are not within the above two ranges are selected. During training, the batch training method of small batch is selected, the batch size is set to 2, the number of iterations is 21600, the basic learning rate (learning_rate = 0.001), and the learning rate decays to one-tenth of the previous time for each iteration of 3240 times. Training of the algorithm consists of initializing the model first, and then starting the training of the detection model after parameter configuration.

In this paper, CASIA and RTD datasets are mainly used. The datasets' division for training and test is shown as Table 1.

Table 1. Division of training set and test set.

Dataset	CASIA 2.0	Columbia	Coverage	RTD
Training set	458	164	90	198
test set	50	18	10	22

4.2 Display of Experimental Results

We conducted experiments on the fusion dataset, obtained the experimental results of three tampering methods (Cp, Rm, Sp) respectively, and compared with the single-stream rgb domain faster-rcnn model and the RGB-N method. As we can see from the visual effects at first, our method is the closest to the tampered position of the ground truth among the three detections and has very high confidence, the comparison is shown as Fig. 5.

4.3 Evaluation Criteria and Description

The evaluation criteria used list as follow:

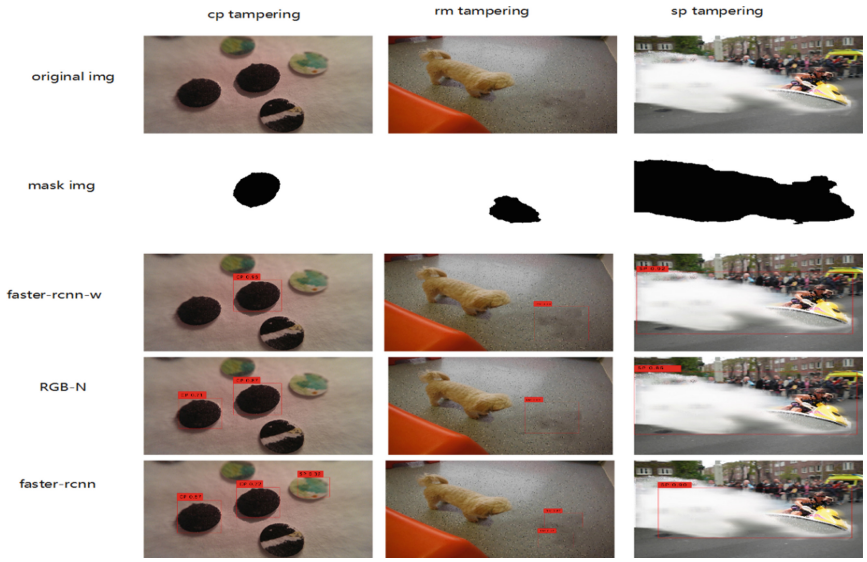

Fig. 5. Comparison with other models.

(1) F1 score precision and recall are generally used at the same time, and F1-score neutralizes the evaluation of both:

$$\mathrm{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

(2) AP indicator PR curve concept: P in the PR curve represents precision, and R represents recall, which represents the relationship between precision and recall. The recall is set to the abscissa, and the precision is set to the ordinate. The AP index is the area enclosed by the PR curve and the x-axis.

(3) The AUC index roc curve refers to the false alarm probability P(y/N) obtained by the subjects under different judgment criteria under specific stimulation conditions as the abscissa, and the hit probability P(y/SN) as the ordinate, connecting the drawn points. AUC (Area Under Curve) is defined as the area under the ROC curve enclosed by the coordinate axis.

First, we compared our own model in three tampering methods, as shown in Table 2.

Table 2. Experimental evaluation diagram.

Tarmperway/Evaluation criteria	F1-score	AUC
Cp	35.4	45.4
Rm	53.4	74.4
Sp	63.4	87.4

Based on the data in the table, it can be seen that this method can realize three tamper detection schemes. This method is more effective among them for splicing detection, and the AUC value reaches 87.4, followed by the AU value of removal 74.4. The detection result of this method for copy-paste averages is not ideal, however, and the AUC value is only 45.4. According to the AUC and F1-score results, we can see that the method has better classification performance in detecting the effects of splicing and deletion. A F1-socre score for splicing and tainting is 63.4, and an F1 score for tainter removal is 60.4. Copy-paste detection results are not perfect.

Subsequently, we compared the performance of the model in this paper with the rgb single-stream faster-rcnn model and RGB-N on the same dataset, and evaluated with the above three evaluation indicators, as shown in Table 3.

Table 3. Comparison of evaluation indicators of different models.

Method	Tampering form	F1-score	AUC
Faster-rcnn-w	Cp	35.4	45.4
	Rm	60.4	74.4
	Sp	63.4	87.4
RGB-N	Cp	41.2	84.3
	Rm	58.6	78.4
	Sp	49.6	81.2
Faster-rcnn (rgb)	Cp	40.3	45.4
	Rm	35.4	50.3
	Sp	37.4	49.6

Table 3 shows that our dual-stream input model outperforms the faster-rcnn single-stream model by a large margin in each evaluation index. We conduct experiments on the fusion dataset, and benchmark the single-stream rgb input faster-rcnn model. As standard single-stream model, we not only add an input source to the input, but also perform image processing on the double-end, which is beneficial for the dct operation, and add high-frequency and low-frequency image filters to obtain high-rate and low frequency images of the dual-stream input end. The standard rgb stream represents the overall content of the image, while the high frequency side contains detail and noise. With the action of the high frequency input side, our network can analyze unseen details and noise in the rgb domain, which greatly improves the accuracy of model prediction.

When comparing with RGB-N, we observed that RGB-NN has lower F1-score metrics (49.6) and AUC metrics (81.2) on splicing (sp) than our model. RGB model – N only inputs a noise stream feature as an additional input, but the additional high frequency stream of our model input not only has noise features, but also changes edge details, which complements the insufficient noise stream and increases model accuracy. We replace the rgb stream with the low frequency data stream. This has the benefit of reducing coupling. We repeat the high frequency content repeated in the rgb stream with

the high rate stream. The image is obtained in the low frequency domain with the high frequency content removed as the main body of the template. We show that the input model of the rgb domain can reduce coupling and improve model precision. However, in terms of copy and paste and tamper removal, RGB-N has a better effect than ours. This may be because when our model discards the rgb stream, it discards some content, causing some tampered regions belonging to high frequency content to be lost. Often when the tampered area is missing, the copy and paste content will be inconsistent with the copied area of the original image due to missing content, so faster-rcnn-w performs poorly in tampering cp. Similarly, under rm removal and tampering, our model and RGB-N have little difference in F1 and AUC indicators. It is because edge tearing is evident in tampered area, and the gray-scale contrast is strong whether it is high frequency input or SRM filtering. As the noise point of the image is enlarged, and the tearing edge is filled with such noise, under rm the performance of both is the same.

5 Conclusion

In today's society, image tamper detection is urgently needed to be developed, and our proposed faster-rcnn-w model is not inferior to the current mainstream image tampered detection templates. Above, complete dctization and extract high and low frequency features as input of dual-stream input, and inside the model, use dual-channel input and dual-pooling layer structure to complete the whole model improvement, no longer a simple overlay of convolution and rpn network, After the whole pattern is upgraded in three dimensions, the final result is presented on the low frequency image. We add rgb images at the output end to map the tamper area of low frequency images to complete the final result. Without changing the network hierarchy, this method improves the load speed, coupling and accuracy. At the same time, we have flaws. Out of the three tamper conditions we target, the tamper effect of cp copy-paste is not perfect. This is mainly because the tampered edges of the low frequency cp images are not consistent with the original image. Further research is warranted in future work. Enhance the tamper accuracy cp.

Acknowledgment. This work has been supported in part by the Natural Science Foundation of China under grant No. 62202211, the project supported by National Social Science Foundation under Grant No. 19CTJ014, the Science and Technology Research Project of Jiangxi Provincial Department of Education (No. GJJ170234).

Data Availability. The data in the experiments, used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest. The authors declare that they have no conflicts of interest.

References

1. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)

2. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
3. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
4. Cozzolino, D., Poggi, G., Verdoliva, L.: Efficient dense-field copy–move forgery detection. IEEE Trans. Inform. Forensic. Secur. **10**(11), 2284–2297 (2015). https://doi.org/10.1109/TIFS.2015.2455334
5. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: WIFS (2016)
6. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: image splice detection via learned self-consistency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 106–124. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_7
7. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: a new blind image splicing detector. In: WIFS (2015)
8. Kniaz, V.V., Knyaz, V., Remondino, F.: The point where reality meets fantasy: Mixed adversarial generators for image splice detection (2019)
9. Zhu, X., Qian, Y., Zhao, X., Sun, B., Sun, Y.: A deep learning approach to patch-based image in-painting forensics. Signal Process. Image Commun. **67**, 90–99 (2018)
10. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE Computer Society, Abu Dhabi (2017)
11. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital imagcs. IEEE Trans. Lnform. Forens. Secur. **7**, 868–882 (2012). https://doi.org/10.1109/TIFS.2012.2190402
12. Zhou, L.-N., Wang, D.-M.: Digital Image Forensics. Beijing University ol Posts and Telecommunications Press, Beijing (2008). (in Chinese)
13. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: AAAI (2021)
14. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV (2020)
15. Wang, J., Wu, Z., Chen, J., Jiang, Y.-G.: M2tr: Multi-modal multi-scale transformers for deep-fake detection. arXiv preprint arXiv:2104.09770 (2021)
16. Bianchi, T., Rosa, A.D., Piva, A.: Improved DCT coefficient analysis of forgery localization in JPEG images. In: Proceedings of the IEEE International Conference on Acoustics? Speech and Signal Processing (ICASSP), pp. 2444–2447. Prague, Czech Republic (2011)
17. Huang, X., Yan, F., Xu, W., Li, M.: Multi-attention and incorporating background information model for chest x-ray image report generation. IEEE Access **7**, 154808–154817 (2019)
18. Lin, T.-Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: ICCV (2015)
19. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: CVPR (2016)