Robust Camera Model Identification Based on Richer Convolutional Feature Network

Ze-Yu Zou[1, 2], Yun-Xia Liu[1, 2,*], Wen-Na Zhang[1, 2], Yue-Hui Chen[1, 2], Yun-Li Zang[4], Yang Yang[5] and Bonnie Ngai-Fong LAW[6]
[1]School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China
E-mail: zeyuzou@foxmail.com, ise_liuyx@ujn.edu.cn, ujn_zhangwn@qq.com, yhchen@ujn.edu.cn Tel: +86-531-82767500
[4]Integrated Electronic Systems Lab Co., Ltd, Jinan 250100, China
E-mail: zangyunli@ieslab.cn, Tel: +86-531-88018000
[5] School of Information Science and Engineering, Shandong University, Qingdao 266237, China
E-mail: yyang@sdu.edu.cn, Tel: +86-532-58630701
[6] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China
E-mail: ngai.fong.law@polyu.edu.hk Tel: +852-27664746

***Abstract*— Based on convolutional neural network (CNN), the problem of robust patch level camera model identification is studied in this paper. Firstly, an effective feature representation is proposed by concatenating a multiscale residual prediction module as well as the original RGB images. Motivated by exploration of multi-scale characteristic, the multiscale residual prediction module automatically learn the residual images to avoid the subsequent CNN being affected by the scene content. Color channel information is integrated for enhanced diversity of CNN inputs. Secondly, a modified richer convolutional feature network is presented for robust camera model identification by fully exploiting the learnt features. Finally, the effectiveness of the proposed method is verified by abundant experimental results at the patch level, which is more difficult than image level experiments.**

I. INTRODUCTION

Camera model identification has long been a hot topic in image forensic tasks. Given an image under investigation, camera model identification (CMI) can assist in identifying owners of illegal and disputed materials, as well as solve image copyright problems to some extent. The rationale for CMI is that there is sequence of operations performed inside the camera to obtain the digital images, such as lens aberration, demosaicing, white balance and so on. Each operation leaves an intrinsic and irreversible trace in the image, based on which stable CMI can be performed.

Traditional camera model identification methods rely on well-designed handcrafted features. There have been extensive features exploited for camera model identification, such as the CFA color features [1], de-mosaic trace [2] and features based on other image characteristics [3]. It is acknowledged that residual features usually lead to competitive performance for robust CMI [4]. As the CMI information is a relative weak signal compared with the image content, residual image is obtained by subtracting a denoised version from the original image for the first step. Based on estimated residual image [5], several features [6][7] can be constructed for CMI purpose. However, traditional residual based methods are usually influenced by image content due to the imperfection of denoising filters. Performance of smooth regions is often better than that in edge and texture areas, making these methods somewhat limited. Besides, method noise [8] related to certain denoising algorithm introduces artifacts, which will leave trace in estimated residual images, thus inevitably it will influence later CMI accuracy.

In the past few years, convolutional neural networks have demonstrated superior performance in many image forensics tasks [9][10]. Luca et.al. pioneered the first application of CNN in CMI [11], while a content-adaptive fusion residual network is proposed in [12] to achieve simultaneous brand level, model level, and device level source camera identification. Integrating with pre-trained model, the effectiveness of DenseNet in CMI is discussed in [13][14]. A high-pass filter is utilized for residual image estimation before CNN in [15]. The authors studied the constrained convolution layers as preprocess procedure in [16][17]. The intrinsic feature learning capability of CNN approaches have greatly improved CMI accuracy. However, there is always space for performance improvement by fully exploiting the learnt features.

In this paper, a CNN based robust camera model identification method is presented. Firstly, a multiscale residual prediction module with constrained convolutional layer is proposed to automatically integrate information in different local neighborhoods. Secondly, enhanced data diversity is achieved by simultaneous incorporating learnt residual images as well as the original images as inputs for network training. In this way, color interpolation information can also contribute to provide multiple cues for robust CMI. Thirdly, a modified richer convolutional feature network is proposed to perform camera model identification based on image patches. Finally, our experimental results are all based on small image patches, which is more practical.

The rest of this paper is organized as following. Section 2 discusses the related works of residual image estimation and CNN structures. Details of the proposed CMI algorithm is discussed in Section 3. Section 4 presents the evaluation protocol and experimental results, while Section 5 concludes the work.

II. Related Works

A. Residual Image Estimation

Residual computation is the first prerequisite for many CMI algorithms, as the camera model specific information is often coded in weak signals as compared with the image content. Given an image *I*, Residual *R* is usually obtained by denoising as:

$$R = I - F(I), \tag{1}$$

where *F*(•) denotes the denoising algorithm. The wavelet domain adaptive denoiser adopted in [5] is a common choice for many CMI algorithms, whose performance is unsatisfying that severe artifacts can be observed in edge and texture regions in estimated residual images.

The idea of residual estimation has also been exploited in CNN based CMI algorithms. Amel et al. [15] employed a fixed high-pass filter for residual calculation

$$R = I * \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \tag{2}$$

before subsequent convolutional layers. The estimated residual images are then served as input for CNN structure, so that more complex camera model features could be automatically learnt by the data-driven model. However, the diversity of CNN features is relatively limited by using a predetermined high-pass filter.

Belhassen et al. [18] proposed a constrained convolutional layer to suppress the influence of image content. They fixed the center value of the convolution kernel weight to -1, and the surrounding weights summed to be 1, i.e.

$$\begin{cases} \omega_k^{(1)}(0,0) = -1 \\ \sum\limits_{m,n\neq 0}^{k} \omega_k^{(1)}(m,n) = 1 \end{cases}, \tag{3}$$

where $\omega_k^{(1)}(m,n)$ denotes the k^{th} filter coefficients in the first layer at corresponding position (*m*, *n*). In this way, these convolution kernels are constrained to be high-pass kernels and residual estimation can be achieved in fully end-to-end manner. However, only 3 constrained kernels with fixed size of 5×5 is adopted in the green channel of input RGB images.

B. CNN Architechture for CMI

Relatively simple CNN network structure is adopted in the first CNN attempt [11] for camera model identification. CNN serves as an efficient feature extractor where the CMI result is obtained by cascaded trained SVM classifier. A systematic performance evaluation protocol is proposed, where overall accuracy of 93% is reported on carefully selected 18 camera models in Dresden database [19]. Experiments prove that it can achieve satisfactory generalization ability.

Amel et al. [15] studied the scalability problem in CNN network design, and reported a small network consisting of three convolutional layers, one pooling layer and three fully connected layers. This network is based on modification of the AlexNet, which leads to comparable performance with 27 layers of GoogleNet model.

In [16], based on experimental evaluation of several CNN design principles, a network is proposed with constrained convolutional layer, ReLU activation and max pooling. The authors further enhanced the low-level feature extraction part by involving nonlinear median filtered residuals to construct augmented convolutional feature maps (ACFM) in [17]. Remarkable performance improvement demonstrated the significance of effective feature representation module design in CNN.

III. The Proposed CMI Algorithm

Camera model identification performance drops dramatically with the decrease of patch size. In this paper, we constrain all discussion into patch level evaluation that the input images are divided into non-overlapping small image patches. The framework of proposed method consists of two major parts, namely the feature representation module with multiscale residual prediction concatenated with original image patches, and the modified richer convolutional feature network for camera model identification.

A. Feature Representation Module with Multiscale Residual Prediction

Effective feature representation module should be able to provide abundant CMI related features, while excluding feature extraction method related noises which would confuse subsequent CNN layers. To this end, the data-driven constrained convolutional layer is adopted as a basic component for residual image estimation. The proposed feature representation module is illustrated in Fig.1.

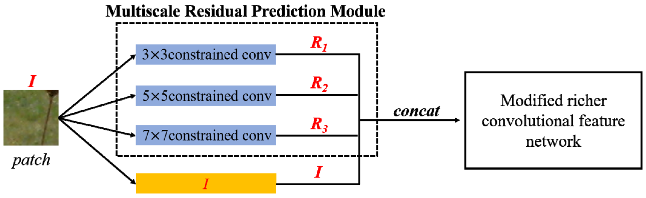
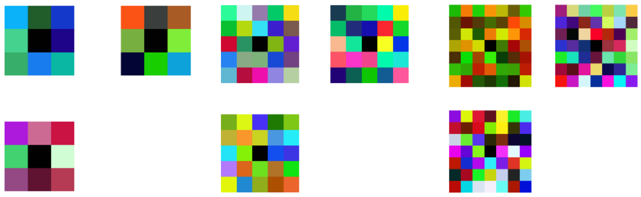

Multiscale Residual Prediction Module
I
patch
3×3constrained conv
5×5constrained conv
7×7constrained conv
I
R1
R2
R3
I
concat
Modified richer convolutional feature network

Fig. 1 Flowchart of the feature representation module.

A multiscale residual prediction module consisting of three parallel constrained convolutional layers with varying local sizes of 3×3×3, 5×5×3 and 7×7×3 is included to reduce the impact of scene content. Three constrained convolutional kernels are allowed in all pathways of R_1, R_2, and R_3. Parameter settings of in-camera operations employed by different manufacturers will leave traces in different local neighborhoods. In contrast to [16], where only green channel is passed into CNN, the constrained convolutional layers are applied to all RGB channels in the proposed scheme. Although the green channel carries a lot of information related to sensor pattern noise, the RGB three-channel images can reflect the interpolation trace, which could help CMI. Joint representation of these multiscale features in all color channels is expected to provide richer CMI related information.

It is worth mentioning that, the constrained convolutional layer can be trained just like a normal convolutional layer. During backpropagation, the weights of the convolution kernel are updated at each iteration by using a stochastic gradient descent algorithm. The constraints are added to constrained convolutional layer when updating the weights, so that the weight of the convolution kernel center is -1, and the sum of the other weights is 1. After completing the above constraints, the updated convolution kernel weights are returned.

We visualized the learnt constrained convolutional kernels in Fig 2. It can be found that the convolution kernels are all high-pass filters with the central coefficient shown in black, which can reduce the influence of image content to a certain extent. Meanwhile, kernels in different scales demonstrate varying properties, which are complementary to each other for providing rich features for subsequent CNN architecture.



3 ×3constrained conv 5 ×5constrained conv 7 ×7constrained conv

Fig. 2 Visualization of constrained convolutional kernels in multiscale residual prediction module.

Meanwhile, the original image patches also constitute an important part in effective feature representation that they contain color interpolation information which contributes to distinguishing different camera models. Therefore, original RGB patches are retained as the input for subsequent CNN layers.

B. *Richer Convolutional Feature Network*

The proposed CMI network is inspired by the richer convolutional feature (RCF) network [20]. Guided by the principle of making full use of learnt CNN features instead of seeking deeper network structure, RCF achieves state-of-the-art performance in edge detection. The principle of fully exploring richer CNN features naturally fits to our motivation for robust camera model identification.

The proposed network architecture is shown in Fig.3, where VGG16 is adopted as the backbone network. Depending on number of pooling layers applied, all convolutional layers are divided into five stages, each followed by ReLU and max pooling layers. In contrast to the local multi-scale features discussed in section 3.A, this fine to coarse decimation scheme provides another form of multi-scale feature representation in a global manner. A conv layer with kernel size 1×1 and channel depth 21 is utilized to accumulate features from multiple stages, where following eltwise layer (shown as $\sum$ in Fig.3) to attain hybrid features.

Since the CMI task is essentially a classification problem, the deconvolution layers of the RCF network are removed. To overpass the disadvantages of training difficulty and large parameter volume of fully convolutional layer, each eltwise layer is connected to a global average pooling (GAP) layer before they are passed into FC layers for classification. The number of nodes in FC layers are set to number of camera models to be discriminated and network parameters are given in detail in Fig.3. Final classification output is obtained by softmax of accumulated FC outputs by eltwise layer.

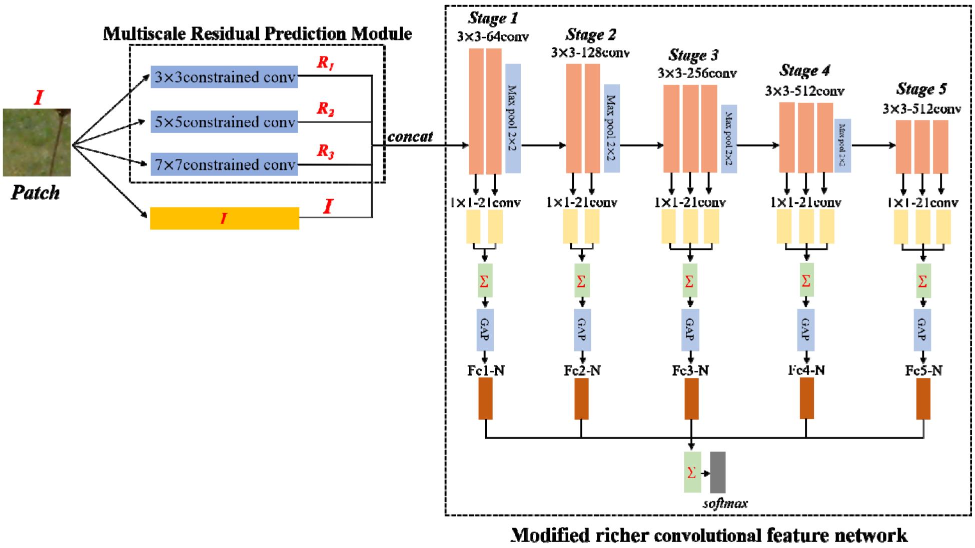


Fig. 3 The architecture of proposed CNN network.

IV. Experimental Results and discussion

A. Experimental Setup

Abundant experiments are carried out on selected camera models from the largest public image forensic Dresden database [19]. Camera models with only one camera device is excluded to avoid the influence of a single camera device on camera model features. Totally 12 camera models are selected according to this principle, whose detail information is given in Table I. When dividing the 7650 images into training set, validation set, and test set, we restrict their scene content to be inconsistent with each other. To further increase the evaluation difficulty, the test set are consisted of images coming from camera devices never appeared in the training set nor the validation set. Finally, the training set contains a total of 403,072 image patches and the validation set consists of 66,800 image patches. There are approximately 600,000 image patches in the test set.

TABLE I
Camera models used in the experiments.

No.	Camera model	Original Resolution	No. images
0	Canon_Ixus70	3072×2304	363
1	Casio_EX-Z150	3264×2448	692
2	Kodak_M1063	3664×2748	1698
3	Nikon_CoolPixS710	4352×3264	695
4	Nikon_D200	3872×2592	373
5	Olympus_mju-1050SW	3648×2736	782
6	Praktica_DCZ5.9	2560×1920	766
7	Ricoh_GX100	3648×2736	559
8	Rollei_RCP-7325XS	3072×2304	377
9	Samsung_L74wide	3072×2304	441
10	Samsung_NV15	3648×2736	412
11	Sony_DSC-T77	3648×2736	492

All evaluation are conducted on a patch level that all training and test images are divided into non-overlapping patches of 64×64 pixels, while 64 patches are randomly selected from each of the training and test images. All CMI accuracy experimental results are based on patch-level evaluation:

$$Accuracy = \frac{\text{No. of Correctly classified patches}}{\text{Total No. of test patches}} \times 100\%. \tag{4}$$

The patch-level evaluation is closer to the real CMI application scenario and is much more difficult as compared with the image-level classification where major voting strategy is adopted [11].

The experiments are conducted on a PC with Intel(R) Core(TM) i5-8500 CPU @ 3.00 GHz, equipped with a NVIDIA GTX 1080Ti GPU on Ubuntu 16.04 operating system under the caffe framework. The learning rate is initialized to 10-3, while the weight decay and momentum are set to 0.00075 and 0.9, respectively. Stochastic gradient descent (SGD) optimization method is utilized.

B. Experiment 1: Comparison of color channel input

In [16-18], only green channel is utilized as it carries most of the sensor pattern noise information and the least interpolation noise. However, all color channels are utilized in the proposed multiscale feature representation module. The abandonment of red and blue color information may have impact on the final classification results as interpolation trace is also useful in camera model identification.

In order to verify that the three color channels can contribute to camera model classification, we performed a comparative experiment based on the CNN model in [11]. With fixed network structure, the only difference between two comparative scheme is the color channel input (which may slightly affect the depth of the first CNN layer kernels).

TABLE II
CMI accuracy comparison of different color channel input.

Methods	Accuracy (%)
G channel image	81.64
RGB channel image	**90.93**

As shown in Table II that, average classification accuracy is improved by 9.29% when after adopting the three color channel inputs, verifying that the RGB three channel information trained together can indeed facilitate the CMI application.

C. Experiment2: Comparison of multiscale residual prediction module

The proposed feature representation module is consisted of multiscale residual prediction module and RGB information. In order to determine the optimal structure of the multiscale residual prediction module, we have experimentally tested several combination structures. Since the constrained convolutional layer adopted in [16] is 5×5, it is employed as the benchmark. Modified RCF network structure is fixed for fair comparison, but network structure are trained respectively for camera model identification.

TABLE III
CMI accuracy Comparison of different
multiscale residual prediction module structures.

Structures of the residual prediction module	Accuracy (%)
5×5	96.52
3×3+5×5	96.67
3×3+5×5+7×7	97.03
3×3+5×5+7×7+ I	**98.05**

From comparison results in Table III, we see that the proposed multiscale residual prediction module achieves best CMI accuracy as expected. Furthermore, the integration of RGB channels brings additional 1.02% performance improvement.

D. Experiment 3: Comparison of network structures

In order to test the effectiveness of the proposed modified RCF network, we report CMI accuracy with fixed feature representation module in Table IV. Only the CNN structure in [11] is designed for 64×64 patches, which is the same as our setting.

TABLE IV
CMI ACCURACY COMPARISON OF NETWORK STRUCTURES.

Methods	Accuracy (%)
Feature representation module + CNN	87.16
Feature representation module + RCF	**98.05**

Prominet performance improvement can be observed in Table IV, demonstrating the effectiveness of the proposed RCF network that more camera model related features and underlying interrelation can be learnt. One may be noticed the gap between the 'Feature representation module + CNN' method and the 'RGB channel image' method in Table II that, the adoption of multiscale prediction module causes performance decrease. As we have observed convergence difficulty during our experiments, possible explanation is that the simple CNN network structrure in [11] is unable to capture the internal relationship between features and camera models. Noticing the 98.05% accuracy of the proposed method, this happens to add evidence for the effectiveness of the modified RCF structure.

E. Experiment 4: Comparison with other CMI methods

For the last experiment, we compare the proposed method with state-of-the-art methods, including Luca bondi [11] method, Amel TUAMA[15] method and two methods [16][17] from Belhassen Bayar. However, this comparison is not easily performed as the database, selected camera models, patch sizes all varies with each other. For instance, the networks in [15-17] are initially designed for 256×256 patches, which is easier as compared to the setting in this paper. To make them compatible to 64×64 patches, we adjusted the network structure by:

- Set the stride of conv3 layer to 1 in [15];
- Change the stride of conv2 layer from 2 to 1 in [16-17];
- Increase the padding parameter by 1 for conv2, conv3 and conv4 layers in [16-17].

We followed other default parameters settings suggested by authors to make a fair comparison of different methods.

TABLE V
ACCURACY COMPARISON OF DIFFERENT CMI METHODS.

Methods	Accuracy (%)
Luca bondi [11]	90.93
Amel TUAMA[15]	93.77
Belhassen Bayar [16]	87.42
Belhassen Bayar [17]	96.78
Our proposed methods	**98.05**

From the comparison results in Table V, it can be observed that the proposed method achieves the best CMI accuracy. Furthermore, it is necessary to emphasize that the proposed method is a fully data-driven method that is convenient in application. There is no need to go through additional cumbersome procedures, such as high-pass filtering and median filtered image feature extraction, etc.

V. CONCLUSIONS

A CNN based robust camera model identification method is proposed in this paper. An effective feature representation module is proposed to provide richer camera model related features, by employing a multiscale residual prediction module to ease the influence of scene content, as well as integrating the RGB color channels to provide color interpolation information. A modified richer convolutional network is proposed to make fully exploitation of learnt features. The effectiveness of the proposed method is verified with large scale patch-level experiments which is designed to mimic the real CMI applications.

ACKNOWLEDGMENT

This work was supported by National key Research and Development Program (No. 2018YFC0831105 and 2018YFC0831100), the National Nature Science Foundation of China (No. 61305015, No. 61203269), the National Natural Science Foundation of Shandong Province (No. ZR201702160307), and Shandong Province Key Research and Development Program, China (No. 2016GGX101022).

REFERENCES

[1] S. Gao, G. Xu, and R. M. Hu, "Camera model identification based on the characteristic of CFA and interpolation." *International Workshop on Digital Watermarking*. Springer, Berlin, Heidelberg, 2011, pp. 268-280.

[2] C. Chen, and M. C. Stamm, "Camera model identification framework using an ensemble of demosaicing features." *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, Rome, Italy, 2015, pp. 1-6.

[3] X. Lin, J. H. Li, S. L. Wang, F. Cheng, and X. S. Huang, "Recent advances in passive digital image security forensics: A brief review." *Engineering* 4(1), pp. 29-39, 2018.

[4] F. Marra, G. Poggi, C. Sansone, and L. Verdoliva, "Evaluation of residual-based local features for camera model identification." *International Conference on Image Analysis and Processing*. Springer, Cham, 2015, pp. 11-18.

[5] M. Chen, J. Fridrich, M. Goljan, and J. Lukás, "Determining image origin and integrity using sensor noise." *IEEE Transactions on information forensics and security* 3(1), pp. 74-90, 2008.

[6] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification based machine learning approach with high order statistics features." *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, Budapest, Hungary, 2016, pp. 1183-1187.

[7] K. R. Akshatha, A. K. Karunakar, H. Anitha, U. Raghavendra, and D. Shetty, "Digital camera identification using PRNU: A feature based approach." *Digital Investigation* 19, 2016, pp. 69-77.

[8] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising." *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2, pp. 60-65, IEEE, 2005.

[9] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images." *IEEE Transactions on Information Forensics and Security* 14(5), pp. 1181-1193, 2018.

[10] W. Quan, K. Wang, D. M. Yan, and X. Zhang, "Distinguishing between natural and computer-generated images using convolutional neural networks." *IEEE Transactions on Information Forensics and Security* 13(11), pp. 2772-2787, 2018.

[11] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks." *IEEE Signal Processing Letters* 24(3), pp. 259-263, 2016.

[12] P. Yang, R. Ni, Y. Zhao, and W. Zhao, "Source camera identification based on content-adaptive fusion residual networks." *Pattern Recognition Letters,* 2017.

[13] U. Kamal, A. M. Rafi, R. Hoque, S. Das, A. Abrar, and M. Hasan, "Application of DenseNet in Camera Model Identification and Post-processing Detection." *arXiv preprint arXiv:1809.00576,* 2018.

[14] A. Kuzin, A. Fattakhov, I. Kibardin, V. I. Iglovikov, and R. Dautov, "Camera Model Identification Using Convolutional Neural Networks." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, Seattle, WA, USA, 2018, pp. 3107-3110.

[15] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks." *2016 IEEE International workshop on information forensics and security (WIFS)*. IEEE, Abu Dhabi, UAE, 2016, pp. 1-6.

[16] B. Bayar, and M. C. Stamm, "Design principles of convolutional neural networks for multimedia forensics." *Electronic Imaging* 2017(7), pp. 77-86, 2017.

[17] B. Bayar, and M. C. Stamm, "Augmented convolutional feature maps for robust cnn-based camera model identification." *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Beijing, China, 2017, pp. 4098-4102.

[18] B. Bayar, and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection." *IEEE Transactions on Information Forensics and Security* 13(11), pp. 2691-2706, 2018.

[19] T. Gloe, and R. Böhme, "The Dresden Image Database for benchmarking digital image forensics." *Proceedings of the 2010 ACM Symposium on Applied Computing*. Acm, Sierre, Switzerland, 2010, pp. 1584-1590.

[20] Y. Liu, M. M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, Hawaii, 2017, pp. 3000-3009.