Deep Residual Convolutional Neural Network with Curriculum Learning for Source Camera Identification

I.O Animasahun, Kin-Man Lam
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China

ABSTRACT

Source camera identification is a fundamental area in forensic science, which deals with attributing a photo to the camera device that has captured it. It provides useful information for further forensic analysis, and also in the verification of evidential images involving child pornography cases. Source camera identification is a difficult task, especially in cases involving small-sized query images. Recently, many deep learning-based methods have been developed for camera identification, by learning the camera processing pipeline directly from the images of the camera under consideration. However, most of the proposed methods have considerably good identification accuracy for identifying the camera models, but less accurate results on individual or instance-based source camera identification. In this paper, we propose to train an accurate deep residual convolutional neural network (ResNet), with the use of curriculum learning (CL) and preprocessed noise residues of camera images, so as to suppress contamination of camera fingerprints and extract highly discriminative features for camera identification. The proposed ResNet consists of five convolutional layers and two fully connected layers with residual connections. For the curriculum learning in this paper, we propose a manual and an automatic curriculum learning algorithm. Furthermore, after training the proposed ResNet with CL, the flattened output of the last convolutional layer is extracted to form the deep features. The deep features are then used to learn one-vs-rest linear support vector machines for predicting the camera classes. Experimental results on 10 cameras from the Dresden database show the efficiency and accuracy of the proposed methods, when compared with some existing state-of-art-methods.

Keywords: Source camera identification, Deep learning, Photo-response non-uniformity, Residual convolutional neural network, Curriculum learning

1. INTRODUCTION

Source camera identification (SCI) is a fundamental area in forensic science, which deals with attributing a photo to the camera that has captured it. SCI can either be the identification of the model/brand of the camera or the identification of the individual camera, used to capture the photos, which are termed camera model identification and instance-based source camera identification, respectively. SCI can help us identify the owners and provide useful information to further aid the investigation of illicit images. Applications of SCI include pseudo-pornographic images and life-threating scenes source identification. Sensor pattern noises (SPN) or photo response non-uniformity (PRNU) has been used for instance-based SCI, due to its uniqueness and universality [1]. The extracted PRNU images are unique to the camera that captured them, i.e. different cameras of the same brand/model have their PRNU fingerprints different. For each camera, its PRNU fingerprint can first be constructed from a set of photos, containing the same noise residues, taken from the camera. The noise pattern of a testing photo is compared with the PRNU fingerprints or the reference SPNs of the cameras in a database to verify if the testing photo is captured by one of the cameras in the database. The verification can be achieved using different similarities measures, such as normalized correlation (NC) or peak to energy correlation (PCE) [2]. A high value of the correlation coefficient indicates that the chance of the photo being captured by a particular camera is high.

Instead of using NC or PCE, machine learning techniques, such as support vector machine (SVM) and neural networks (NNs) [3, 4], have also been used to learn the camera features. Another machine-learning technique, which has high discriminative power and is capable of achieving robust feature representation, is deep learning. Deep convolutional neural networks (CNNs) have been used in some research works [5-7] for SCI. Furthermore, the work in [8] proposed using a content-adaptive fusion residual network (CA-FRN) for SCI, based on image size of 64×64. The assumption in [5, 7, 8] is that, CNN is able to learn the features about the processing pipelines of cameras directly from the images, without using any handcrafted features. To achieve good identification accuracy for the proposed CNN-based methods, images are divided into patches for training. The works in [5, 7, 8] achieved considerably good performance in terms of the average identification accuracy for identifying the camera models, but lower average identification accuracy for individual instance-based SCI. A large amount of training images is also required for a good performance, unlike the PRNU-based approaches that use a single patch per image. Hence, high computational cost is required for the CNN-based methods. Furthermore, the work in [9], carried out a comparative study on the proposed CNN-based SCI and on a PRNU-based method using identical settings, concluded that the PRNU-based method has higher identification accuracy, at a lower computational cost, than the proposed CNN-based method. An interesting question is whether a more accurate deep network architecture can be designed for instance-based SCI, by using effective training algorithms. In addition, identification based on small-size patches using PRNU is even more challenging for camera identification. This is because the estimated PRNU fingerprint will be weakened, as the number of pixels reduces. However, identification from a small patch is important, especially in applications, such as splicing translocations and small-sized forgery detection.

Therefore, in this paper, we propose curriculum learning algorithms, for training a deep residual CNN (ResNet) for instance-based SCI of small-sized images, using noise residues of cameras. The use of noise residues is to suppress the contamination of camera features by image contents. The main advantages of using ResNet are that it generates more robust representational bottlenecks and also tackles the problem of vanishing gradients through the smooth flow of data between networks [10]. For SCI, the PRNU fingerprint of a camera is difficult to detect from an image, if the image has complicated patterns. In other words, the extraction of the PRNU fingerprint from a smooth, plain image is much easier than from a cluttered or natural image. Based on this observation, we employ curriculum learning to train a deep ResNet for source camera identification. The major concept of curriculum learning (CL) [11] is to train a system from simple tasks to hard tasks. This can train a system with better performance than that with the simple and hard samples together. The use of curriculum learning can help to improve the speed of global convergence during training, and achieve a better local minimum [11]. The rest of the paper is organized as follows. Section 2 gives an overview of residual learning and curriculum learning for neural networks. Section 3 describes the framework of the proposed ResNet with curriculum learning. Section 4 discusses the experiments and results, while Section 5 presents the conclusion.

2. OVERVIEW OF RELATED METHODS

This section gives an introduction and the related theories of our proposed methods. We discuss the theoretical basis for deep residual learning for deep neural network, followed by a discussion on the concept of curriculum learning.

2.1 Deep residual learning

The work in [12] revealed that the network depth is crucial for a network to achieve high generalization power. However, an obstacle to increasing the depth of a network is the problem of vanishing or exploding gradients [13]. By vanishing gradient, it means that the gradients in the network loss function tend towards zero, and this makes the network optimization become difficult. To address the problem of degradation in accuracy when the number of convolutional layers increases, the residual neural network (ResNet) was proposed by He et al. [10]. This is motivated by the need to increase the depth of a deep neural network, without the problem of vanishing gradients. A deep neural network learns the underlying mapping of a given data x. Let $G(x)$ represent the mapping function, which is to be learned by a stack of either fully connected layers or convolutional layers. The work in [10] assumes that if $G(x)$ can learn from a stack of layers consisting of non-linear functions, then it can also approximate its residual functions. Assuming that the dimension of the input and output layers are the same, the residual function $F(x)$ can be expressed as $G(x) - x$. The assumption is that, instead of directly stacking additional layers, the additional layers can be added as identity mappings. Hence, a deeper network with smaller training loss can still be trained. The degradation of accuracy in a deep neural network without residual mapping can be attributed to the network optimisation algorithms, finding it difficult to approximate mappings, due to a stack of non-linear layers. However, this becomes easier with the use of residual mapping, because it helps the optimisation algorithms push the weights of the network layers towards zero. Residual mapping adds shortcut connections to the next layers.

2.2 Curriculum learning for neural networks

The idea of curriculum learning (CL) is motivated by the education system, where learning is introduced from simple concepts to hard concepts. Organizing the education system in this way helps the students to leverage hard concepts based on their understanding of the easy concepts. This same idea can be applied to the training of neural networks, where training is initiated to begin with those examples in a dataset that are easier to learn by a network, before introducing complicated examples to the network to learn. CL works based on a sequence of training subsets via progressive training. The distribution of the training subsets satisfies the requirement for curriculum learning, if the entropy of the training subsets monotonically increases. CL has been demonstrated in [11] that it can improve the global convergence speed during training and achieve a better local minimum. What easy and hard examples mean depending on the area of the application. In facial expression recognition [14], those face images with a high expression intensity can be considered easy examples, while those images with a low expression intensity are hard examples.

3. OUR PROPOSED METHOD

The PRNU image in a smooth or flat image is easier to learn than that in a natural image, which contains more complex textures. In our CL algorithm, flat images are used as easy, simple samples, while natural images are used as hard examples. The general framework for our proposed algorithm is shown in Figure 1. The noise residues in the images extracted by a wavelet-based denoising method [1] form the PRNU images, which are arranged according to the complexity of the images, from easy images, i.e. those smooth and flat images, to difficult images, i.e. those natural images, and are divided into d groups, denoted as $(D_1, D_2 \ldots\ldots D_d)$. In other words, d subsets are formed for curriculum learning. Assume that the images are generated by K cameras, i.e. there are K classes, and n_i is the number of samples in a subset D_i. Then, the total number of samples N in the camera dataset is given by $N = \sum_{i=1}^{d} n_i$. The denoising method applied to the camera images is based on a wavelet-based denoising filter [1], and the noise residues in an image captured by a specific camera are obtained by subtracting the filtered images from the original images. The noise residues in an image are also pre-processed by the zero-mean operation. This zero-mean operation is applied to row-by-row, followed by column-by-column, of each noise-residue image. This operation can help reduce the effect of linear patterns introduced into the noise residues, due to the color interpolation and pipeline processing operations of sensor and electronic circuits in cameras [15]. It also acts as a normalization process. Each training subset, D_1 to D_d, is used to train the ResNet sequentially. After all the subsets have been used for training, the features from the last convolutional layer of the trained ResNet are extracted to form the deep features of the input samples. These extracted deep features are then used to learn one-vs-rest linear support vector machines (SVMs) for predicting the camera classes. The use of the one-vs-rest linear SVM classifiers results in more training samples for the classifiers.

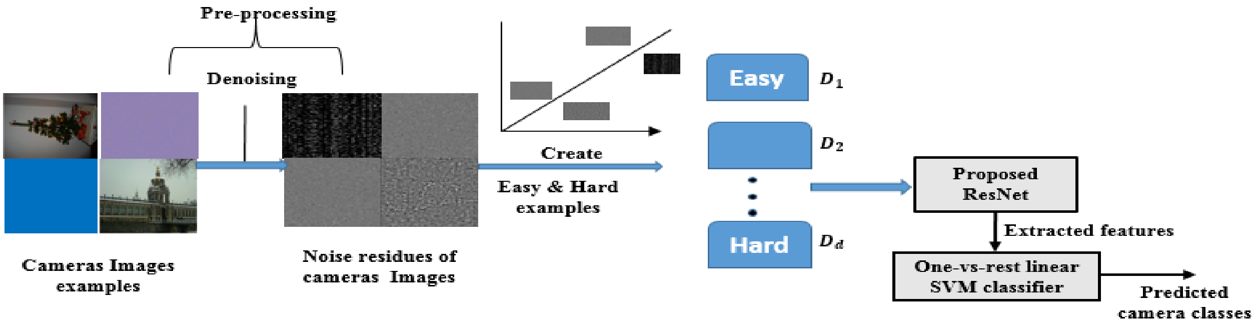

Figure 1. The general framework of our proposed deep ResNet with curriculum learning for SCI.

3.1 The proposed ResNet model and its training process

The architecture of the proposed deep ResNet model is shown in Figure 2, which consists of five convolutional layers, three residual connections, and two fully connected layers. The convolutional layers of the proposed CNN includes processing operations, such as convolution, stride, batch normalization (BN), and leaky rectilinear activation (Leaky ReLU). The size of the input data into the ResNet model is 64×64. Since the 2×2 stride is used in Conv5, 2×2 stride and 1×1 filter size are used in Conv4, to linearly downsample the output of Conv3, so that it can have the same feature-map size when concatenated with the output of Conv5, before given it as input to the two fully connected layers FC1 and FC2. Other details, including the values of the model parameters, are also shown in Figure 2. The model is trained using categorical cross-entropy loss and optimized using mini-batch stochastic gradient descent. The optimal performance of our proposed ResNet is obtained by using 20 epochs, a mini-batch size of 16, and the learning rate of 0.001 with a momentum of 0.9. Moreover, sparsity constraint and weight regularization methods are used with the regularization parameter of 10^{-5} each in the FC1, FC2 and softmax layers to further prevent model overfitting. Imposing sparsity constraint is a form of regularization, but not weight regularization. It regularizes the outputs of the layer rather than the weight of the layer. Weight regularization can be achieved by using $L2$ norm regularization, while the sparsity constraint is imposed by adding the absolute values of the true value of a layer into the loss function. Finally, we introduce the use of class weights to the training function. Class weights penalize under or over-represented classes in the training set. A class weight is computed as, $class\ weight = 0.15 \log (N/number\ of\ samples\ in\ a\ class)$. If the class weight is less than 1, the estimated weight for the class will be set at 1. This is then passed as a dictionary into the class weight parameter of the network training function in the Keras deep learning framework.

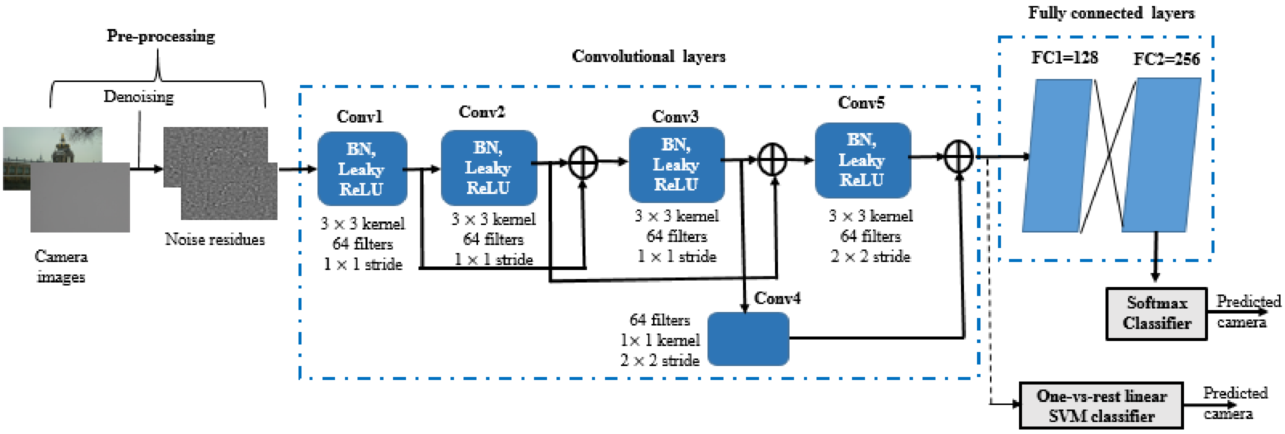

Figure 2. The layout of the proposed ResNet model for instance-based SCI.

3.2 The proposed curriculum learning algorithm

In this paper, we propose a manual and an automatic CL algorithm. For the manual CL algorithm, the easy and hard examples are selected manually prior to the training. For the automatic CL algorithm, the training data are sorted in the order of increasing complexity. The training procedures for the proposed manual and automatic CL algorithms are listed as follows.

Manual curriculum learning

i. The noise residues of flat images are first used to train the proposed ResNet model.
ii. The best-trained model is obtained.
iii. The best-trained ResNet model is trained with a decreased learning rate.
iv. The noise residues of natural images are now used to train the trained ResNet.
v. The best-trained model is then used for extracting deep features for camera identification.

The two data subsets are trained with 20 epochs. The learning rate is decreased as stated in (iii), so that the negative influence from the hard examples can be reduced. The learning rate is reduced from 0.001 to 0.0007.

Automatic curriculum learning

i. The input dataset contains all the training images (x) and the corresponding labels (y).
ii. The best version of ResNet is trained on the dataset (x, y).
iii. The predicted training features (X) of the softmax layer of the trained model are extracted.
iv. The softmax loss of X is then calculated. The softmax loss (p) can be expressed as,

$$p_i = \frac{e^{X^i}}{\sum_{j=1}^{K} e^{X^j}} \tag{1}$$

where $i = 1\!:\!N$, X^i is each training instance, K is number of camera classes, $j = 1\!:\!K$ and X^j is the value of X^i in j.

v. The cross-entropies for each instance in X are obtained. The cross-entropy loss can be defined as,

$$L_c\left(y, p_i\right) = -\sum_{i=1}^{N} y_i \log p_i \tag{2}$$

where y_i is the class label of a training instance, X^i. The value of y_i ranges from $1\!:\!K$.

vi. Training instances with smaller cross-entropies can be better optimized than training instances with larger cross-entropies, hence, the cross-entropies are sorted in ascending order. The indices of the sorted cross-entropies are used to re-order the original dataset, (x, y).
vii. The re-ordered dataset, (x, y), is used to train the ResNet model, and the trained ResNet is used to predict the camera classes.

4. EXPERIMENTS AND RESULTS

The well-known Dresden database [16] was used for testing our proposed method. Ten cameras are used in the Dresden database, and each camera has a number of flat and natural images. All the images of the cameras are used to evaluate the proposed methods, as tabulated in Table 1. All images are center-cropped to form image patches of size 64 × 64, with a single patch per image. All the flat images are used for training, while the natural images of each camera are separated randomly, such that 80% training samples and 20% testing samples. For the manual CL, flat images are used as easy, simple samples, while natural images are used as hard examples. For the automatic CL, both flat and natural images are combined, and easy and hard examples are sorted in the order of increasing complexity, as described in the automatic CL algorithm. In our experiments, we evaluate the performance of the ResNet, when it is trained with and without using CL, as well as anti-CL. Anti-CL means that the training is conducted by using hard examples first, followed by easy examples. All the experiments were carried out on a personal computer with 4.00GHz Intel (R) Core (TM) i7-6700k CPU and 1 terabyte memory. Extraction of the noise residues from all the images was carried out by using MATLAB 8.6.0. The deep learning implementation was carried out using Python and the Keras deep learning library with Theano backend and, Nvidia GTX 1080 Ti (11G memory). The Keras was also used with the Scikit-learn library to further leverage the power gain of the optimization of the model parameters. Table 2 shows the identification accuracies for both manual and automatic CL, using softmax classifier (ResNet-SC) and the one-vs-rest linear SVM classifier (ResNet-SVM). PRNU_F and PRNU_N denote the PRNU of flat and natural images, respectively. As shown in Table 2, the best performance is obtained by training the network with the PRNU of the flat images first, followed by the PRNU of both flat and natural images. Its overall identification accuracy using ResNet-SVM is 0.7% higher than that using PRNU of flat images followed by PRNU of natural images only, and 3.74% higher than the ResNet without using CL. Table 2 also shows that the anti-CL has the lowest identification accuracy. Furthermore, we conducted an experiment on the proposed automatic CL, and the experimental result, as shown in Table 2, has the identification accuracy 0.47% higher than that without using CL. This indicates that there is only little impact on the accuracy, with and without using the automatic CL. The confusion matrix for the ten cameras, with the best trained ResNet model, is shown in Table 3. The individual camera identification accuracies are shown in the diagonal of Table 3, and highlighted in bold. The average individual accuracy for the 10 cameras for images of size 64 × 64 is 78.66%.

Table 1. Details of the cameras used in the experiment, including the resolution and the number of flat and natural images.

S/N	Camera Brand	Resolution	Natural Images	Flat Images
1	Canon_Ixus70_0	2304 × 3072	171	50
2	Canon_Ixus70_1	2304 × 3072	179	50
3	Canon_Ixus70_2	2304 × 3072	171	50
4	Samsung_L74wide_0	2304 × 3072	229	50
5	Samsung_L74wide_1	2304 × 3072	224	50
6	Samsung_L74wide_2	2304 × 3072	231	50
7	Samsung_NV15_0	2304 × 3072	217	50
8	Samsung_NV15_1	2304 × 3072	214	50
9	Sony_DSC-H50_0	2736 × 3648	266	50
10	Sony_DSC-H50_1	2736 × 3648	234	50
Total number of Images			**2136**	**500**

Table 2. Overall accuracy for manual curriculum learning with combinations and orders of the flat images and natural images (%).

Curriculum Learning (CL)		
Different Progressions	**ResNet-SC**	**ResNet-SVM**
PRNU_F + PRNU_N	66.82	73.36
PRNU_F, PRNU_F + PRNU_N	67.76	77.10
PRNU_F + PRNU_N, PRNU_F	50.71	73.83
PRNU_F, PRNU_N	72. 43	76.40
PRNU_N, PRNU_F	57.94	72.43
Automatic Learning		
PRNU_F+ PRNU_N	65.89	73.83

We compare our best result in Table 2 (ResNet-SVM with manual CL) with three state-of-the-art methods for instance-based SCI. The methods are the maximum likelihood estimated SPN (MLE SPN) [17], Phase SPN [2], and the weighted averaging (WA) method [18]. The same experimental setup was used for all these state-of-the-art methods. For the purpose of fair comparison, in the three methods, the camera identification was achieved by using the peak to energy correlation.

Table 3: Identification accuracy (%) of the best result of the proposed method: **The average accuracy is 78.66%**

Camera Device		1	2	3	4	5	6	7	8	9	10
Canon_Ixus70_0	1	**76.32**	13.16	5.26	-	-	-	-	5.26	-	-
Canon_Ixus70_1	2	2.78	**83.33**	13.90	-	-	-	-	-	-	-
Canon_Ixus70_2	3	-	18.75	**78.12**	-	-	3.12	-	-	-	-
Samsung_L74wide_0	4	-	22.20	22.20	**64.44**	17.78	11.11	2.22	-	-	-
Samsung_L74wide_1	5	-	23.80	-	16.67	**61.90**	11.90	4.76	2.38	-	-
Samsung_L74wide_2	6	-	-	-	4.88	14.63	**75.61**	-	4.88	-	-
Samsung_NV15_0	7	-	-	-	-	-	-	**94.87**	5.13	-	-
Samsung_NV15_1	8	-	2.00	-	2.00	-	-	26.00	**68.00**	2.00	-
Sony_DSC-H50_0	9	-	-	-	-	-	-	-	-	**93.75**	6.25
Sony_DSC-H50_1	10	-	-	-	-	-	-	2.44	-	7.32	**90.24**

The comparison results are shown in Table 4. We can see that the overall average individual camera accuracy of the proposed method is 15.93%, 5.71%, and 15.51% higher than that of the MLE SPN, Phase SPN and WA methods, respectively. Therefore, our proposed deep learning-based method has better average identification accuracy than the compared methods. Hence, an accurately designed deep network, based on good training algorithms, can achieve better performance than the conventional or PRNU-based SCI methods, for small-query images using the same number of training and testing examples.

Table 4. Comparison with some state-of-the-art methods.

Methods	**Accuracy** (%)
Proposed method	78.66
MLE SPN [17]	62.73
Phase SPN [2]	72.95
WA method [18]	63.15

5. CONCLUSION

In this paper, deep residual convolutional neural networks (ResNet) with curriculum learning algorithms are proposed for source camera identification, by using sensor pattern noise of cameras. Experiments were conducted with different orders of the camera subsets (simple and hard subsets) from the Dresden database, with our manual and automatic curriculum learning algorithms. For the proposed manual curriculum learning, experimental results show that training based on easy training examples before hard examples will result in the best identification accuracy, based on the proposed ResNet model. Furthermore, our proposed automatic curriculum learning approach shows better identification accuracy compared to training without applying curriculum learning. In conclusion, our proposed deep learning methods for instance-based SCI can achieve better performance than the compared state-of-the-art methods using the same settings. For future works, we will focus on using suitable pre-processing operations to generate simple examples from the original training set so as to further increase the number of training subsets. Moreover, a better automatic curriculum learning approach will be explored, so that the learning efficiency can be improved, and hence increase the camera detection accuracies.

REFERENCES

[1] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, pp. 205-214, June 2006.

[2] X. Kang, Y. Li, Z. Qu, and H. J., "Enhancing source camera identification performance with a camera reference phase sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 393-402, April 2012.

[3] Y. Huang, J. Zhang, and H. Huang, "Camera model identification with unknown models," *Information Forensics and Security, IEEE Transactions on*, vol. 10, pp. 2692-2704, 2015.

[4] M. Kharrazi, H. T. Sencar, and N. Memon, "Blind source camera identification," in *International Conference on Image Processing , ICIP'04.*, 2004, pp. 709-712.

[5] L. Baroffio, L. Bondi, P. Bestagini, and S. Tubaro, "Camera identification with deep convolutional networks," *arXiv preprint arXiv:1603.01068*, 2016.

[6] A. Tuama, F. Comby, and M. Chaumont, "Camera Model Identification With The Use of Deep Convolutional Neural Networks," in *IEEE International Workshop on Information Forensics and Security*, 2016, pp. 1-6.

[7] L. Bondi, L. Baroffio, D. Guera, P. Bestagini, E. J. Delp, and S. Tubaro, "First Steps Towards Camera Model Identification with Convolutional Neural Networks," *IEEE Signal Processing Letters*, pp. 1-5, 2016.

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

[9] F. Ahmed, F. Khelifi, A. Lawgalv, and A. Bouridane, "Comparative Analysis of a Deep Convolutional Neural Network for Source Camera Identification," in *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*, 2019, pp. 1-6.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[11] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41-48.

[12] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377-2385.

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Aistats*, 2010, pp. 249-256.

[14] L. Gui, T. Baltrušaitis, and L.-P. Morency, "Curriculum learning for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 505-511.

[15] B.-b. Liu, X. Wei, and J. Yan, "Enhancing sensor pattern noise for source camera identification: An empirical evaluation," in *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, 2015, pp. 85-90.

[16] T. Gloe and R. Böhme, "The dresden image database for benchmarking digital image forensics," *Journal of Digital Forensic Practice*, vol. 3, pp. 150-159, 2010.

[17] M. Chen, J. Fridrich, M. Goljan, and J. Lukáš, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, pp. 74-90, March 2008.

[18] A. Lawgaly, F. Khelifi, and A. Bouridane, "Weighted averaging-based sensor pattern noise estimation for source camera identification," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014, pp. 5357-5361.