PRNU-based Image Classification of Origin Social Network with CNN

Roberto Caldelli[*†], Irene Amerini[*‡], Chang Tsun Li[‡§]

roberto.caldelli@unifi.it, irene.amerini@unifi.it, chli@csu.edu.au

[*]Media Integration and Communication Center (MICC), University of Florence, Florence, Italy
[†]National Interuniversity Consortium for Telecommunications (CNIT), Parma, Italy
[‡]School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, Australia
[§]Department of Computer Science, University of Warwick, Coventry, UK

Abstract—A huge amount of images are continuously shared on social networks (SNs) daily and, in most of cases, it is very difficult to reliably establish the SN of provenance of an image when it is recovered from a hard disk, a SD card or a smartphone memory. During an investigation, it could be crucial to be able to distinguish images coming directly from a photo-camera with respect to those downloaded from a social network and possibly, in this last circumstance, determining which is the SN among a defined group. It is well known that each SN leaves peculiar traces on each content during the upload-download process; such traces can be exploited to make image classification. In this work, the idea is to use the PRNU, embedded in every acquired images, as the "carrier" of the particular SN traces which diversely modulate the PRNU. We demonstrate, in this paper, that SN-modulated noise residual can be adopted as a feature to detect the social network of origin by means of a trained convolutional neural network (CNN).

I. INTRODUCTION

The pervasiveness of new ICT technologies has paved the way for new aggressive behaviors and cyber-violence. Many actions (e.g. harassment, violence instigation, cyber-bullying) are perpetrated online through social networks or messaging applications. In particular, the widespread usage of smartphones has intensified such a phenomenon: a picture is acquired and uploaded to different networks and illegal activities proliferate through misuses of such digital contents to achieve various malevolent objectives. This paper is aimed to deepen forensic analysis on images downloaded from social networks or spread via instant messaging apps going one step further in describing the "history" of the processing a digital image has undergone [1], [2]. In particular, in this paper a methodology able to identify the distinctive and permanent trace imprinted in digital media by the acquisition device is proposed in order to classify the photos according to the social network of provenance. In fact, it is well known that each SN leaves peculiar traces on each content during the upload-download process; such traces can be exploited to assist image classification. In this work, the idea is to use the PRNU (Photo Response Non-Uniformity) noise embedded in images by the source devices as a carrier which is uniquely modulated by each SN. The PRNU noise is usually used as a fingerprint to identify a specific digital camera in a dataset [3] or to perform image clustering [4]. The PRNU noise is a resilient fingerprint left in image by the sensor of the camera at the time when the image is taken and usually survives, under certain conditions, various processing the image is subjected to. For this reason, our idea is to adopt it as a signature to detect the social network of origin through the use of a trained Convolutional Neural Network (CNN). It is notorious that image resizing, cropping and JPEG compression make a reliable camera identification more difficult [5], [6], but in this particular case we are not interested in the source device identification task but in understanding if a SN will apply the unique modification consistently to the noise residual (the carrier) on every processed images. If such consistency and uniqueness are observed, we will know that the traces due to the modification can be exploited for social network identification.

Although the problem of social network identification was only brought to the attention of multimedia forensic community recently, its importance has led to the introduction of new benchmarking datasets [7],[8],[9]. In [10], a preliminary work has proved that the process to upload images onto *Facebook* does leave unique and detectable traces in the content. The same authors later refined their idea in order to classify, using a K-NN classifier, different social networks based on the traces of resizing, compression, renaming and metadata alterations left during the upload/download procedure [9]. Furthermore, in [8] and [11], methods to differentiate social networks such as *Facebook*, *Flickr* and *Twitter* are exploited by adopting only content-based information recovered from DCT (Discrete Cosine Transform) histograms of JPEG images. In [8], social network identification is achieved by means of a Bagged Decision Tree Classifier (BDTC), while a CNN is used to perform classification in [11]. The use of CNN is particularly suited to solve this kind of problem because of its capability of automatically learning the best features to solve the classification task. CNN and machine learning are used extensively in many areas such as image classification and object detection and recently also in image forensics. CNN have also been employed to detect image manipulations by revealing single or double JPEG compressions [12],[13],[14] and to perform source camera identification [15],[16]. Deep learning has also been used very recently in [17] for classifying four types of global processing applied to an image namely low-pass filtering (blurring), high-pass filtering (sharpening), denoising
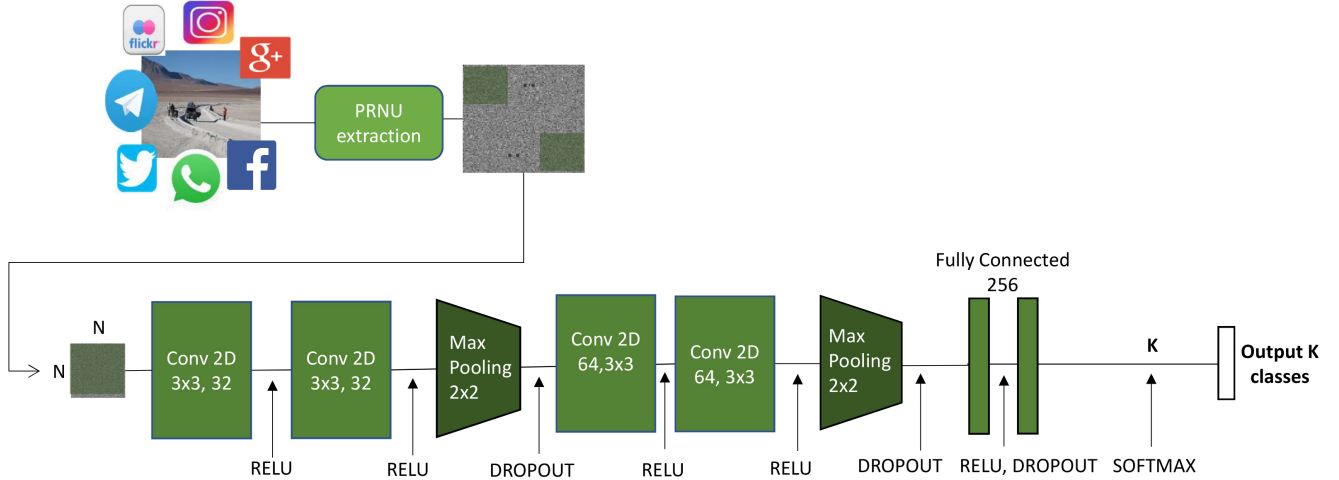
Fig. 1. The whole pipeline: PRNU extraction and the proposed CNN.

(content adaptive low-pass filtering), and tonal adjustment. Driven by the substantial increase of attention around this topic our objective is to study the behavior of the noise residual in the case of upload/downloading through a SN by using it as a feature for SN-based image classification.

The paper is organized as follows. Section II the proposed method is introduced describing the pre-processing phase (noise residual extraction and normalization) as well as the training and testing procedure of the CNN. Section III presents some of the main experimental results on different datasets, while Section IV draws final conclusions and suggests some possible future directions.

II. THE PROPOSED METHOD

The main assumption behind social media image classification is that each platform applies specific transformations during the process of upload-download. Some of these transformations are easy to infer (e.g. JPEG recompression, resizing), while some are unknown or not publicly available. However it is contended that they are unique to the extent that can be used to classify the social network of provenance of an image. According to this, different features have been proposed in the recent scientific literature and adopted as pre-processing of the input of a machine learning classifier. In this work, we have decided to investigate if these peculiar traces can be extracted from the image noise residual. In the subsection II-A, noise residual extraction is briefly introduced while in subsection II-B the whole procedure and the designed CNN is presented.

A. Noise residual extraction

Photo Response Non-Uniformity (PRNU) noise is very well-known in image forensic literature and has been used in many applications devoted to source/device identification. Such a noise is embedded within every digital image by

the camera sensor and represents a unique fingerprint of the device. Being PRNU part of the content, it will presumably be affected by the manipulations the image has been subjected to. PRNU is usually extracted from the image content through high-pass filtering (see Equation (1)) followed by an estimate operation, as formulated in Equation (2). The term W_i stands for the noise residual containing PRNU, while I_i and I_i^{den} represent the *i-th* image and its denoised version respectively. The PRNU fingerprint $\hat{K}$ is obtained through a minimum variance estimator as indicated in Equation (2) where M is the number of images utilized for the estimation.

$$W_i = I_i - I_i^{den} \qquad (1)$$

$$\hat{K} = \frac{\sum_{i=1}^{M} W_i I_i}{\sum_{i=1}^{M} (I_i)^2} \qquad (2)$$

Different kinds of denoising filters have been introduced to improve noise residual extraction. In this work, the often-used approach (wavelet-based) described in [18] has been adopted. Being a noise due to the acquisition sensor, it is spread all over the image and consequently has the same dimension as that of the host image (e.g. $R \times C$). The noise intensities are signed real numbers distributed around zero and, furthermore, an estimate operation (see Equation (2)) cannot be performed in the case at hand because there will not be M pictures at disposal to refine PRNU estimation. To address this issue, only noise residual W_i, as in Equation (1) is considered and recovered at full-frame size. After that, the intensity of each noise residual pixel $W_i(r, c)$ is scaled and normalized to the range $[0 : 1]$ and finally it is subdivided into non-overlapping squared patches of $N \times N$ size (in the following experiments $N = 64$), in order to consistently provide the CNN with the same number of to-be-learnt features (Figure 1).

B. The procedure and the CNN

For each image, noise residual is computed as described in the previous subsection and each $N \times N$ squared patch is fed to a convolutional neural network as illustrated in Figure 1. This particular net is comprised of two convolutional blocks and two fully connected layers.

Each convolutional block is composed by two convolutional layers with the ReLU activation function ($g(x) = max(0; x)$) followed by a pooling layer. The kernel size of all convolutional layers is 3×3 while the size of the pooling layer kernel is 2×2. Dropout [19] is used to prevent overfitting by randomly dropping units at training time from the fully connected layers. The first fully connected layer has a dimension of 256 and it employs dropout that, during preparatory analysis, proved to be a suitable solution. The dimensionality of the final layer is K and its outputs are sent to a softmax layer in such a way that the final output is a probability distribution for each of the K classes corresponding to the considered social networks. A categorical cross-entropy function [20] is employed as loss function to guide the training process of the classification problem. The prediction is obtained at the patch level after processing each image patch with the CNN. Majority voting is used at image level to assign the SN class label with the highest score among the patches of the image.

III. EXPERIMENTAL RESULTS

This section introduces some of the experimental tests carried out to understand if the proposed methodology and specifically the use of noise residual can be adopted to reliably track the social network of provenance of a certain image. In subsection III-A all the used datasets will be described, while in subsection III-B different kinds of experiments will be presented and discussed.

A. Datasets

Three different datasets have been used in our experiments and they are all publicly available on the web. The first two sets were already employed in previous works and they are used as reference: *UCID social* [1] and the dataset named *IPLAB* [2]. The *UCID* (Uncompressed Colour Image Database) dataset [21] is composed by 1338 pictures of 512×384 pixels acquired in the raw format by a single digital colour camera, a *Minolta Dimage 5*. The *UCID social* is based on the *UCID* dataset; it consists of JPEG compressed images generated at different quality factors $QF = 50 : 95$ (step 5) starting from the uncompressed ones and then uploaded/downloaded to/from three selected social network (*Flickr*, *Facebook* and *Twitter*) by using available API. In total, *UCID social* is composed by 40140 images (1338 images$\times$10 QFs$\times$3 social networks).

In the case of *IPLAB* database, 8 different classes are present: 5 social networks (*Facebook*, *Flickr*, *Google+*, *Instagram* and *Twitter*), 2 instant messaging apps (*WhatsApp* and *Telegram*) and one class of unshared (just taken by the camera) images. The considered pictures have been acquired by using

4 cameras with two (high and low) quality resolution allowed by each device (picture resolutions range from 640×480 to 5184×3456 depending on the camera). The devices involved in the creation of the *IPLAB* dataset are the following: *Canon 650D*, *QUMOX SJ4000*, *Samsung Note3 Neo* and *Sony Powershot A2300*. Each of the 4 devices contributed with 30 pictures at two different resolutions yielding 240 images in total. Consequently, with 8 SNs involved, the dataset consists of 1920 images (240×8 classes).

The third dataset is the *VISION* dataset presented in [7]. It is a quite comprehensive dataset composed of videos and images so we select a sub-set of it for our purpose. The following 10 smartphones (or tablets) have been taken into account: *Samsung Galaxy S3 mini*, *Huawei P9*, *LG D290*, *Apple iPhone5c*, *Apple iPhone6*, *Lenovo P70A*, *Samsung GalaxyTab3*, *Apple Iphone4* and 2 models of *Apple iPhone4s*. The amount of pictures per device is different and finally we got an ensemble of 2135 images. All of them have been shared on *Facebook* and then downloaded in high and low quality, and through *WhatsApp*, resulting in a total of 6405 (2135×3) images.

The CNN is trained by subdividing each of the three datasets into training, validation and test sets with the following proportion 80%, 10% and 10% respectively. Images belonging to the three subsets are randomly selected each time. The input of the net is a 64×64 matrix and the outputs are K SN classes which vary according to the dataset under analysis and/or the kind of experiments. The number of elements of each class (i.e. social network) is not the same which gives rise to an unbalance training that mimics real application scenarios. The neural network, described in section II-B, is optimized by using the AdaDelta method [22]; the training phase ends when the loss function reaches its minimum on the validation set, usually less than 50 epochs.

B. Description of the experiments

This subsection presents an extended set of experiments devoted to test the proposed method in different operational conditions; in particular, it has been investigated the performance stability with respect to a diverse number of cameras (i.e. diverse PRNU), types and number of social networks and messaging applications and, above all, types of image datasets.

1) Results on UCID social dataset: In this first series of experiments, the *UCID social* dataset composed of 1338 images at 10 different JPEG quality factors (50:95 with a step of 5) and uploaded-downloaded on three social networks (*Facebook*, *Flickr* and *Twitter*) has been considered. Being all the pictures taken by a single camera, they are expected to contain the same PRNU fingerprint and, consequently, such a noise signal should be similar when extracted from the image content. Moreover, the distortions inflicted on the noise residual by the process of image upload-download on a social network, should be, at least ideally, similar. This circumstance should permit to leave out, at least at this stage, the possible effects due to the presence of different cameras (i.e. different noise residuals) to investigate the effectiveness of the proposed approach. In

Classification (%) vs SNs	Facebook	Flickr	Google+	Instagram	Original	Telegram	Twitter	WhatsApp
Facebook	**77.41**	9.52	0.80	8.69	0.95	1.56	0.80	0.27
Flickr	5.49	**69.85**	1.65	3.70	2.18	11.48	3.68	1.97
Google+	0.14	0.70	**87.08**	1.29	3.29	1.48	0.02	6.00
Instagram	6.93	15.07	2.07	**54.00**	1.61	7.84	0.30	12.18
Original	0.00	0.00	1.01	0.03	**97.73**	0.15	0.00	1.07
Telegram	0.63	3.50	0.50	1.76	0.34	**91.14**	1.59	0.54
Twitter	3.95	31.13	1.66	0.74	1.59	38.19	**21.92**	0.81
WhatsApp	0.03	1.68	2.39	12.74	4.30	4.30	0.31	**74.25**

TABLE I
PRECISION ON UCID DATASET FOR
Facebook (FB), *Flickr* (FL) AND *Twitter* (TW).

	Patch level			Image level		
QF	**FB**	**FL**	**TW**	**FB**	**FL**	**TW**
95	79.29	82.95	79.39	84.09	90.90	96.96
90	79.14	80.95	91.47	79.54	94.69	100.00
85	84.53	82.01	93.64	95.45	90.15	100.00
80	58.41	93.29	77.05	62.87	98.48	94.69
75	70.80	97.99	62.68	96.24	100.00	84.86
70	71.08	99.06	45.02	91.66	100.00	100.00
65	76.51	97.68	61.71	96.96	100.00	62.12
60	69.83	99.18	87.00	87.87	100.00	81.06
55	73.68	99.06	69.81	94.69	100.00	90.90
50	64.72	99.31	57.14	84.09	100.00	66.66
Avg	**72.80**	**93.15**	**72.49**	**87.35**	**97.42**	**87.73**

Table I, results concerning performances in terms of precision are presented with respect to all the 10 JPEG quality factors at both patch and image levels (a majority voting criterion is adopted in this case). Encouraging experimental results show that, regardless the diversity of quality factors, good distinction among social networks can be observed.

2) Results on IPLAB dataset: In this subsection, another dataset, publicly available on the web (see Section III-A for details), has been taken into account. In this case 8 different classes are present (7 among social networks or instant messaging apps and one class of original unshared image) and the considered pictures have been acquired by using 4 cameras (e.g. 4 diverse fingerprints) with two different kind of resolution (low and high). In Table II, the confusion matrix is presented in terms of precision percentages at the patch level. the sum of the values on the rows yields to 100% and the bold numbers along the diagonal indicate correct detections. It can be observed that for most of the 8 social networks the classification performances are satisfactory, in particular, they are very good for the case of *Original* images where the precision of 97.73% is achieved. Two cases are not so good: *Instagram* and *Twitter*. In the first one, though the obtained precision is not so acceptable (54.00%), it can be said that the errors are quite distributed over all the other classes and *Instagram* is not confused with another specific social network. This misclassification could be due to the cropping

performed by *Instagram* causing high variation among noise residuals which, in turn affects the classification. On the contrary, in the second case, the correct classification rate of the pictures from *Twitter* is only about 22% and there is a strong mis-classification towards *Flickr* and *Telegram* (more than 30% for each). It is worthy pointing out that this is a very challenging dataset composed of images with extremely different image resolutions (from the lower to the maximum possible resolution of the camera).

In Figure 2 the results obtained with the *IPLAB* dataset when propagating the classification from the patch level to the image level is depicted. The average correct classification rate of this case is 83.85% and the *Instagram* identification is around 75%. On the contrary the *Twitter* case remains misclassified demonstrating the difficulty in the classification task for this particular SN. This issue will be investigating in future works.
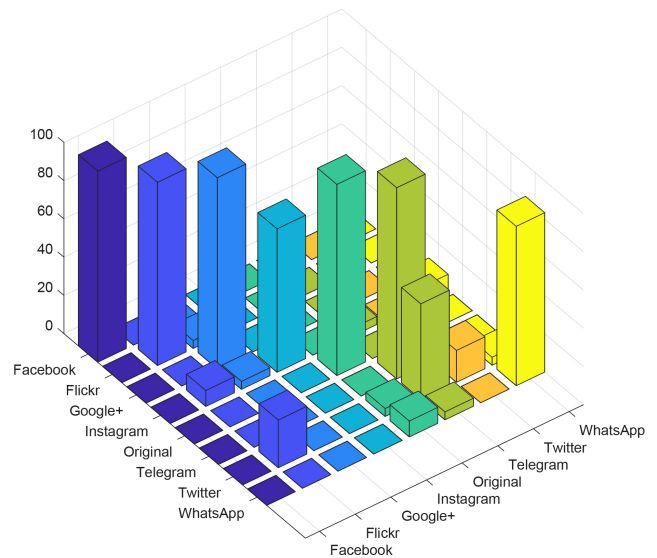

Fig. 2. Confusion matrix for the *IPLAB* dataset at image level.

3) Results on VISION dataset: In this subsection, experiments devoted to investigate the influence of a higher number of devices (i.e. many kinds of noise residuals) have been carried out on a selection of the *VISION* dataset. Various tests gradually increasing the number of devices involved

have been done, only some of the most significant ones are reported hereafter. In particular, Table III presents the patch level performance in terms of precision associated with the cases of 5 and 10 different smartphones (left and right side of the table respectively) with images subdivided in three classes: *Facebook* (two diverse kinds of resolutions, low and high are grouped together), *WhatsApp* and *Original*. By observing the

TABLE III
PRECISION ON VISION DATASET FOR
Facebook (FB), *WhatsApp* (WA) AND *Original* (ORIG).

	5 smartphones			10 smartphones		
	FB	**WA**	**Orig**	**FB**	**WA**	**Orig**
CNN						
PRNU	99.712	95.799	99.982	97.860	97.972	99.792
DCT	99.691	98.543	99.925	97.767	98.613	99.992

first row of Table III, it can be pointed out that doubling the number of devices involved in the experiments does not give rise to a decrement of the performances that are very high. Moreover, a comparison between the proposed technique and another methodology which uses DCT-based features [11] to train the same CNN is provided. It can be observed that results are very good for both the techniques and consequently comparable. Results on *VISION* dataset at the image level are straightforward by obtaining an error-free classification for each of the three classes.

IV. CONCLUSION

In this paper, the idea to use the modifications noise residual undergoes when an image is uploaded/downloaded to/from a social network as a distinctive feature to classify the provenance social network of that image has been proposed. Experimental tests on different datasets, number and types of devices, as well as types of social platforms have been carried out with satisfactory results observed. Future works will be dedicated to investigate the specific behavior of some SNs and to study the problem of multiple upload/downloads.

REFERENCES

[1] Z. Dias, S. Goldenstein, and A. Rocha, "Large-scale image phylogeny: Tracing image ancestral relationships," *IEEE MultiMedia*, vol. 20, no. 3, pp. 58–70, July 2013.

[2] F. de O. Costa, M. A. Oikawa, Z. Dias, S. Goldenstein, and A. R. de Rocha, "Image phylogeny forests reconstruction," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1533–1546, Oct 2014.

[3] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008.

[4] X. Lin and C. T. Li, "Large-scale image clustering based on camera fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 793–808, April 2017.

[5] M. Goljan and J. Fridrich, "Camera identification from cropped and scaled images - art. no. 68190e," 03 2008.

[6] M. Goljan, M. Chen, P. Comesana, and J. Fridrich, "Effect of compression on sensor-fingerprint based camera identification," vol. 2016, pp. 1–10, 02 2016.

[7] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "Vision: a video and image dataset for source identification," *EURASIP Journal on Information Security*, vol. 2017, no. 1, p. 15, Oct 2017. [Online]. Available: https://doi.org/10.1186/s13635-017-0067-2

[8] R. Caldelli, R. Becarelli, and I. Amerini, "Image origin classification based on social network provenance," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1299–1308, June 2017.

[9] O. Giudice, A. Paratore, M. Moltisanti, and S. Battiato, "A classification engine for image ballistics of social data," in *Image Analysis and Processing - ICIAP 2017*, S. Battiato, G. Gallo, R. Schettini, and F. Stanco, Eds. Cham: Springer International Publishing, 2017, pp. 625–636.

[10] M. Moltisanti, A. Paratore, S. Battiato, and L. Saravo, "Image manipulation on facebook for forensics evidence," in *Image Analysis and Processing — ICIAP 2015*, V. Murino and E. Puppo, Eds. Cham: Springer International Publishing, 2015, pp. 506–517.

[11] I. Amerini, T. Uricchio, and R. Caldelli, "Tracing images back to their social network of origin: A cnn-based approach," in *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, Dec 2017, pp. 1–6.

[12] Q. Wang and R. Zhang, "Double JPEG compression forensics based on a convolutional neural network," *EURASIP Journal on Information Security*, vol. 2016, no. 1, p. 23, 2016. [Online]. Available: http://dx.doi.org/10.1186/s13635-016-0047-y

[13] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro, "Aligned and non-aligned double JPEG detection using convolutional neural networks," *J. Vis. Comun. Image Represent.*, vol. 49, no. C, pp. 153–163, Nov. 2017. [Online]. Available: https://doi.org/10.1016/j.jvcir.2017.09.003

[14] I. Amerini, T. Uricchio, L. Ballan, and R. Caldelli, "Localization of JPEG double compression through multi-domain convolutional neural networks," *Proc. of IEEE CVPR Workshop on Media Forensics*, 2017.

[15] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec 2016, pp. 1–6.

[16] L. Bondi, L. Baroffio, D. Gera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 259–263, March 2017.

[17] M. Boroumand and J. Fridrich, "Deep learning for detecting processing history of images," in *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2018*, 2018.

[18] M. K. Mihcak, I. Kozintsev, and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in *Proc. of IEEE ICASSP*, Phoenix, USA, 1999.

[19] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[21] G. Schaefer and M. Stich, "UCID - an uncompressed colour image database," in *Proceedings of the Storage and Retrieval Methods and Applications for Multimedia*, 2004, pp. 472–480.

[22] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.