2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Bad Teacher or Unruly Student: Can Deep Learning Say Something in Image Forensics Analysis?

Paolo Rota*, Enver Sangineto†, Valentina Conotter‡ and Christopher Pramerdorfer*
*CVL, Vienna University of Technology (Austria)
†DISI, University of Trento (Italy)
‡SocialIT, Trento (Italy)

***Abstract*—The pervasive availability of the Internet, coupled with the development of increasingly powerful technologies, has led digital images to be the primary source of visual information in nowadays society. However, their reliability as a true representation of reality cannot be taken for granted, due to the affordable powerful graphics editing softwares that can easily alter the original content, leaving no visual trace of any modification on the image making them potentially dangerous. This motivates developing technological solutions able to detect media manipulations without a prior knowledge or extra information regarding the given image. At the same time, the huge amount of available data has also led to tremendous advances of data-hungry learning models, which have already demonstrated in last few years to be successful in image classification. In this work we propose a deep learning approach for tampered image classification. To our best knowledge, this the first attempt to use the deep learning paradigm in an image forensic scenario. In particular, we propose a new blind deep learning approach based on Convolutional Neural Networks (CNN) able to learn invisible discriminative artifacts from manipulated images that can be exploited to automatically discriminate between forged and authentic images. The proposed approach not only detects forged images but it can be extended to localize the tampered regions within the image. This method outperforms the state-of-the-art in terms of accuracy on CASIA TIDE v2.0 dataset. The capability of automatically crafting discriminant features can lead to surprising results. For instance, detecting image compression filters used to create the dataset. This argument is also discussed within this paper.**

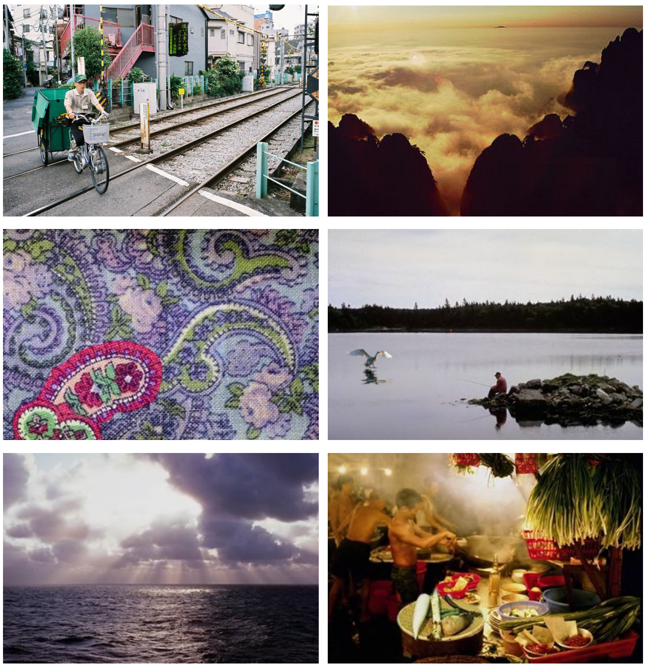


Fig. 1. Examples of tampered images available in CASIA TIDE v2.0 dataset. The tampering operation is often very well performed and the subjects are heterogeneous (e.g. open air and indoor pictures, textures, etc.)

I. Introduction

Ever since images have been digitally browsed on our screens and widely spread through the Internet, serious security issues have arisen regarding the originality of the images themselves. The urgent need of coping with such security issues motivates the computer science community to focus its attention in developing technological solutions able to detect media manipulations without requiring any prior knowledge or extra information regarding the given image. Nowadays skilled operators are able to forge photo-realistic pictures which are hardly obtrusive from the original. Under this perspective Image Forensics community has grown in last 15 years, analyzing images in order to practically solve problems such as source identification [1], splicing detection [2], [3], copy-move detection [4], etc. In this paper we focus our attention on the detection of images affected by splicing operations, where a selected region from an image is pasted into another image with the aim to change its content. In the recent literature most of the methods which have been proposed so far are relying on handcrafted features e.g based on geometric approaches [5] [6], on image residual analysis [7] or Color Filter Array (CFA) analysis [8].

In this work we apply a deep learning algorithm in order to perform a binary classification between tampered and authentic images using patch based processing. This type of architecture has the ability to learn discriminant features directly from data without any a priori knowledge or feature extraction process. In this work we adopt Convolutional Neural Networks (CNN) which in last few years has led to very good results in many computer vision applications such as object recognition [9], [10], image segmentation [11], [12], head pose estimation [13] or face recognition [14], [15] to name a few. The training operation of this model requires a huge number of parameters to be learned, therefore also a huge amount of data is needed to successfully complete the process. However, common training sets for tampered images only have at most a few thousand of images, which are not sufficient to train (or 'fine-tune') big-size nets such as VGG-16 [16]. In order to deal with this problem, we propose a patch-based classification strategy, i.e, rather than classifying the whole image, we

978-1-5090-4846-5/16/$31.00

focus on classifying patches instead. This allows us to mainly address two issues: firstly the lack of data, by using patches we boosted the number of samples used to feed the network. Secondly it allows us to fix the input dimension to the size of the patch without introducing deformation, and therefore more artifacts, into the data. Moreover, since CASIA (and similar available smaller datasets) is not provided with segmentation-level annotation, we propose a straightforward but effective method to automatically compute this ground-truth tampering mask given the information contained in CASIA.

In this paper we also propose an extension of the classical image detection task to a more challenging tampered region localization task using a soft ground truth generated by an automatic method from the CASIA TIDE dataset. In this particular experiment we use a state of art edge detector to search candidate borders, adapting the methodology used for the traditional tampered image detection task.

Our main contributions are:

- A deep learning approach to efficiently learn and classify discriminant features directly from the data, proposing further experiments to better understand the issues related to the application of such technique to image forensics.
- State of art results on CASIA TIDE dataset which is, to the best of our knowledge, the biggest available for forensics applications.
- An automatic procedure to extract tampered border patches from a weakly annotated dataset.
- A patch-based classification approach which focuses on detecting patches lying on the border of the tampered regions in the image.

In section II we briefly overview the related work, focusing on the differences and the importance of our contribution with respect to the state of the art, while in section III we describe the proposed methodology and the architecture of the used CNN. Section IV shows the promising achieved results, while conclusions are drawn in section V.

II. Related Works

Deep learning has been recently applied in audio forensics applications. In [17] the authors use a Restricted Boltzman Machine to detect bootleg unofficially recorded and redistributed by fans after a concert. Luo et al. [18] use a 4 layer neural network to detect AMR double compressed audio signals (e.g. songs).

However in image forensics the use of Neural Networks have been limited to the classification stage while the features have been previously extracted using different techniques. In [19] the authors use a shallow artificial neural network to classify autoregressive features previously collected from training images with the purpose of detecting copy-move forgeries. A similar shallow network has been used by Huang et al. [20] to classify handcrafted features in order to authenticate digital photos detecting demosaicking correlation. An interesting approach has been proposed by Fan et al. [21] where a generalized neural network has been used to simulate computational rules in demosaicking adjusting bias and weights of the network. In [22] the authors designed a two-layer neural network in order to classify features based DCT coefficients in order to estimate primary quantization in a double compressed JPEG image. The use of neural networks as a bare classifier can be observed also in splicing detection works as in [23] where the authors use a shallow RBF network to classify high order statistical features. In [24], Zhang et al. extract moment based features to classify tampered and authentic images.

Using deep networks in the feature extraction process, however, may be a difficult task, in [25] the authors show how deep networks can be literally fooled, returning nonsensical results in an object classification scenario. In our work we are interested in fine grained artifacts which are neither semantically meaningful nor even directly observable, and for this reason they propose a challenge that, to the best of our knowledge, has never been undertaken. The goal of this work is to fill this shortcoming, setting the path for further works in this direction.

III. Framework Design

In order to learn discriminant features directly from data a consistent set of labeled images is needed. The unavailability of such dataset forced us to augment the data ourselves. Moreover, in order to avoid introducing unwanted artifacts, the image should be neither resized nor distorted. For these reasons we opted for a patch based approach. Depending on the application, the patches have been extracted in an exhaustive or by using a selective way (e.g. edge patches). In the first case a sliding window is passed through the image with a stride equal to the half of the patch size $s = \frac{p_{size}}{2}$ (cf. Section IV-A). The latter case, patches are extracted following the edge extraction (cf. Section IV-B) with a 50% non overlapping policy. The patches then have been fed to the network as they are without performing any kind of processing with exception of a linear normalization of the data between 0 and 1.

A. Deep Neural Network

Recently the availability of huge amounts of data and the development of GPU computation has led computer scientists to raise their interests in data-hungry learning techniques. One of the most promising is certainly deep learning which has achieved important results in many image related applications (see Section I). The strength of such approaches is that the discriminant features are learned automatically from data and the image itself is often the only required input to the algorithm. On the other hand a deep network requires a huge amount of data to learn a model since the number of parameters to be learned is also very large (e.g tens of thousands to millions). Because of lack of such datasets for forensics applications we adopt a patch based solution. Indeed, as we show in Section IV, given a tampered image we can extract a variable number of patches on the manipulation border (positive class), while non-tampered patches (negative class) can be extracted randomly from the rest of the images. This allows us to collect a sufficiently large dataset for training a big-size network.

The proposed model is a VGG-like CNN architecture [16] and it is summarized in Fig. 2. It takes as input a 40×40 fixed size patches and it is composed by two Convolutional blocks and two fully connected layers. Each convolutional block is

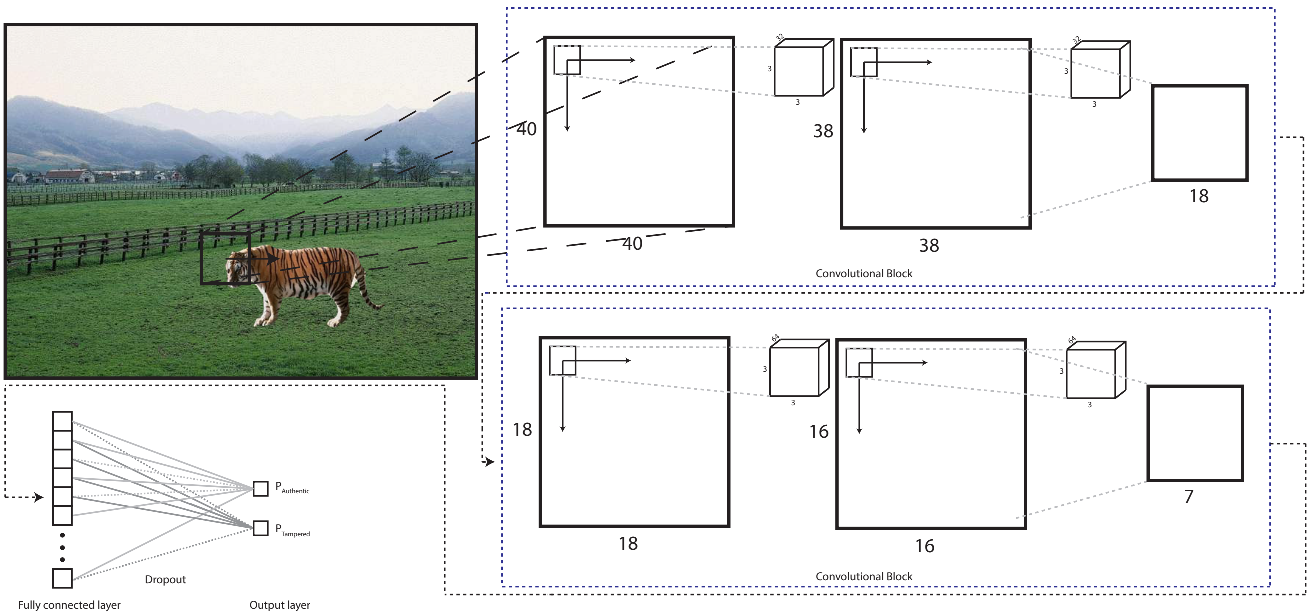


Fig. 2. A draft of the proposed CNN architecture. Activations have been omitted for sake of clarity.

composed by two convolutional layers with ReLU activation [9] followed by a pooling layer. All convolutional layers use a kernel size of 3×3 while pooling layer's kernel size is 2×2. Between different blocks we use Dropout layers [26] in order to address the problem of overfitting. A normalization has been performed on the image before feeding it to the network so as to bring the input numbers in a range between 0 and 1. The overall number of parameters to be learned during the training phase in the proposed network is 869,154.

The free parameters and the architecture of the network are tailored on the task, considering the input size and the number of available patches in the training set.

B. Weak Labeling Procedure

In order to describe the automatic segmentation-level procedure for region localization we firstly have to introduce the dataset we used for the experiments.

1) CASIA dataset: In section III-A we mentioned the importance of the dataset in deep learning applications, in computer vision a considerable step forward has been done after the release of ImageNet dataset [27] which contains millions of images divided in a thousand of object related classes. Unfortunately there is not any dataset for splicing detection that is even close to those numbers. However, CASIA TIDE v2.0 dataset [28] is composed by 7491 authentic and 5123 tampered images. The authentic pictures are taken from the Corel dataset[1] while tampered have been created using Adobe Photoshop CS3 on a Windows XP machine. The dataset includes images taken in different environments depicting different subjects in different situations. The tampering operations are encoded in the name of the file itself and include resizing, deforming, rotating or simply a copy and paste operation. The tampered region can be taken from the same image or from a different one and it can vary in sizes from image to image.

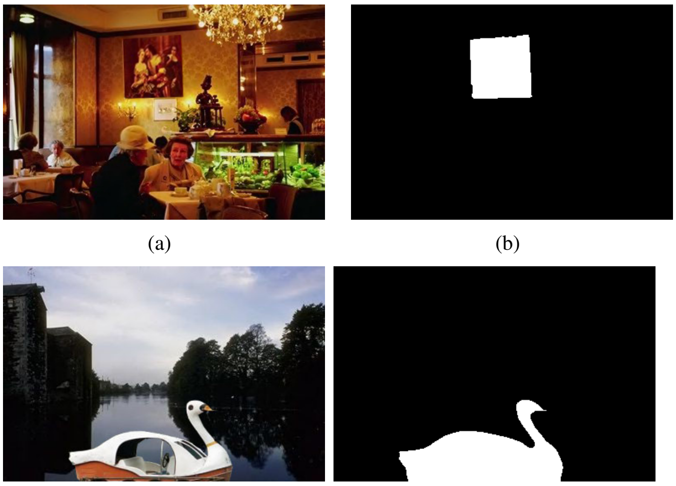
[1]Available at this website: https://archive.ics.uci.edu/ml/datasets/Corel+Image+Features

However, as reported in other works [29], [30] the dataset suffers of several issues which may accidentally ease the classification task for certain approaches. According to [30] the main issues are related to the uniform processing performed by the authors while saving the tampered images; this is inevitably introducing certain artifacts which are not completely related to the tampering itself but concerning other issues (e.g multiple compression). Nevertheless, to the best of our knowledge CASIA TIDE v2.0 is the biggest dataset collecting handcrafted spliced images available so far.

2) Automatic Segmentation-Level Procedure: The region localization task requires the tampering mask which is unfortunately not provided by the authors. However, the combination of authentic images used to compose the tampered image, has been encoded into the name along with some basic information about the generation process (i.e. size of the tampered region; basic operation as rotation, cropping, deformation of the patch; blurring of spliced contours, etc.). This allow us to propose a fairly reliable method to produce such a mask, starting from the encoded information. Let I_D be the absolute value of the image difference between the tampered (I_T) and the first authentic image (I_i)(cf. Eq.(1)).

$$I_D = \|I_T - I_b\| \tag{1}$$

The subscript b refers to the background image which is, among the original images composing I_T, the one who shares the major overlapping similarity with it. I_D is then binarized for a threshold $t = 20$ (cf. Eq (2)). The value of t has been decided in order to reduce the generation of *ghosts* leading to best visual results).



(a) (b) (c) (d)

Fig. 3. An example of automatic ground truth extraction from the CASIA dataset. In this figure (b) and (d) are the automatically generated ground truth for tampered figures (a) and (b) respectively.

$$I_{th} = \begin{cases} 1 & if \quad I_D > t \\ 0 & otherwise \end{cases} \tag{2}$$

The resulting image still requires a further cleaning process mainly for two reasons: firstly some images present spotted noise given by the heuristic thresholding approach, in order to overcome this problem we apply the closing morphological filters. The second reason stems from the fact that some images have irregular tampered regions. This problem is solved by filling completely these areas (e.g. in Fig. 3 (c-d), the inner part of the swan-boat has been completely filled even though it belongs to the background image).

A small amount of images have been discarded ($\sim$ 30) because miss-referenced or the algorithm could not give a satisfactory region approximation.

IV. Results

In this section a detailed description of the experiments have been drafted along with a brief discussion of the results.

A. Tampered Image Detection

In this first experiment we show how a deep neural network performs in discriminating authentic and tampered images. In order to carry out this evaluation the dataset has been divided in training and test sets (95 and 5% respectively). This operation is performed 10 times in order to evaluate the significance of the results. Tampered patches have been extracted from the borders of the tampered area (c.f. Section III-B) while authentic patches are randomly sampled from authentic images. As explained in Section III-A the deep network used in this paper requires a fixed size input. The optimal size of the patches is highly related to the architecture of the network, for our configuration best results are achieved by a patch size of 40×40. The patches are exhaustively extracted from the image with a stride of 20 pixels between each other. The pixel values on the images are linearly normalized between 0 and 1. The resulting number of trainable patches is quite high (1,642,766 patches) which is beneficial for this type of approach. The net is trained for 30 epochs using a batch size of 256 images. During the testing phase, the patches are extracted using the same methodology used for the training part while the final decision is taken image-wise using a patch majority voting policy.

TABLE I
Comparison of the proposed method against competitors on CASIA TIDE v2.0 dataset.

Method	Accuracy	Precision	Recall	F-Score	AUC
[30] (Best)	79.74	-	0.7243	-	0.87
[31] (Best)	96.8	-	-	-	-
[32] (Best)	95.6	-	-	-	-
[33] by [34]	90.1	-	-	-	-
Proposed Approach	**97.44**	0.9616	**0.9881**	0.9747	**0.9936**

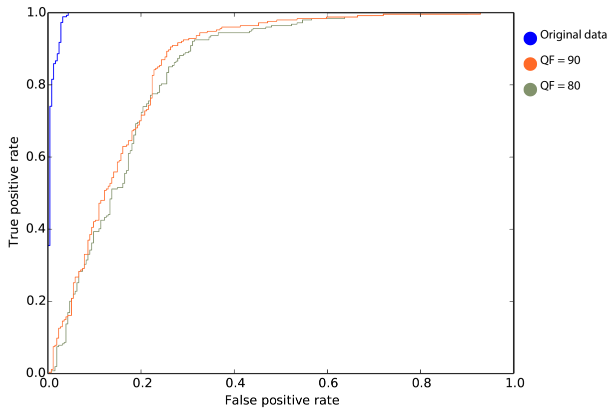


Fig. 4. A performance comparison between the experiments on the original data and those executed on images compressed to different quality factor.

Tab. I shows a comparison between our method and some recently developed approaches, we can see that it outperforms the competitors in this experiment. In Fig.4 the Receiver Operating Curve (ROC) is displayed for sake of comparison with other works in literature.

Considering the issues related to the used dataset, in Tab. II we outline an experiment where all the images are re-saved in jpeg format using different quality factor[2]. These results show a considerable impact of the image compression on the classification performances, where drop of about 30% in terms of accuracy can be observed. This confirms the claims of [30] about the dataset, it also shows that the learning process of a deep network can easily drift from the expected task if the training data is somehow biased. The network learns how to solve the given problem focusing on the most discriminant feature which is, in this case, the artifact introduced by the tampering software leading to state of art performances.

B. Tampered Region Localization

In order to validate the good results reached in the tampered image detection we propose a test that aims at localizing the

[2]We used the batch processing feature of Nomacs (www.nomacs.org) to perform the generation of the compressed images.

Fig. 5. An example of tampering region localization using our deep learning approach. Tampered tiles are red dyed in the top and top right images. In the bottom images the weak labeled mask, the tiled best result and the predicted mask are respectively displayed.

TABLE II
COMPARISON OF THE PROPOSED METHOD ON CASIA TIDE V2.0 DATASET AFTER RE-SAVING THE IMAGES WITH DIFFERENT COMPRESSION FACTOR.

Method	**Accuracy**	**Precision**	**Recall**	**F-Score**	**AUC**
Original	97.44	0.9616	0.9881	0.9747	0.9936
QF = 90	68.11	0.7987	0.4842	0.6029	0.8503
QF = 80	69.29	0.7663	0.5551	0.6438	0.8355

tampered patches within an image. In order to do this, and taking into account the two different processes that affected the two classes (see section III-B1), in our test we only consider the tampered images in the CASIA dataset. Doing this we intend to prevent the neural network to learn those highly discriminative compression artifacts which are unintentionally introduced in the dataset.

The numerical results for this task are proposed on patch basis. We consider a patch as tampered if it lies on the border of the tampered region itself and not within it. Localizing a patch on the border of a tampered object allows to roughly segment the region. For this task we propose a variant that focuses on object insertion. In order to tackle this task we propose an approach where at first we extract the significant edges from the image using the algorithm proposed by Dollar et al. [35]. We use the extracted edges to cross-check each patch and if an edge is present then it is considered valid and it is used for training, nothing is done otherwise. The same process is performed in test phase.

In Tab. III numerical results are shown; accuracy is expressed in terms of percentage of correctly classified patches over the possible candidates. The border approach, as expected, has a smaller recall which means that sometimes some forgeries are not producing any border (e.g. deletions), but on the other hand the increasing of the precision is beneficial in terms of overall accuracy. As we can see in Fig. 5 the proposed method is not able to completely solve the problem, however

TABLE III
RESULTS OF PATCH BASED TAMPERING LOCALIZATION.

Tampering Localization	Accuracy	Precision	Recall	F-Score
Proposed Method	89.04	0.4306	0.6712	0.5247
Proposed Method + Borders Detection	91.61	0.5304	0.6006	0.5634

we observe some good results in different situations i.e. flat images (a) as well as very detailed images (c). The results are promising if we consider that the image is fed blindly to the network without performing any type of preprocessing or handcrafted theory-driven feature extraction.

V. CONCLUSION

In this work we proposed a patch based deep learning approach to detect tampered images. The results show that deep learning approaches can learn discriminant features which are invisible to human eye and beat competitor methods on CASIA TIDE v2.0 dataset. However in this particular case the network does not learn the structure of the pixels around the tampered region but an uniform compression artifact which is spread all over the image. For this reason the learned model can not be generalized for any tampered image nevertheless it leads to the best accuracy on this specific dataset. In the paper we then focused on the harder task of the localization of such regions in this highly biased dataset, proposing also a weak labeling method to generate segmented regions from tampered images out of a generic forged image given at most the background original image. The results show that in this case the network, which is forced to learn different features, is able to reach promising results. In conclusion using a data-hungry learning approach for tampering detection turns out to be a quite hard undertaking when there is no direct control on the source of the training data. As future work we intend to generate a dataset comprised of raw images and including ground-truth labels for tampered regions in order to validate the achieved results in a controlled scenario.

ACKNOWLEDGMENT

The authors would like to thank Nvidia Corporation for the kind donation of the Titan X GPU used in this work.

REFERENCES

[1] A. Singh and J. Malik, “A comprehensive study of passive digital image forensics techniques based on intrinsic fingerprints,” *International Journal of Computer Applications*, vol. 116, no. 19, 2015.

[2] R. M. Joseph and A. Chithra, “Literature survey on image manipulation detection,” 2015.

[3] G. K. Birajdar and V. H. Mankar, “Digital image forgery detection using passive techniques: A survey,” *Digital Investigation*, vol. 10, no. 3, pp. 226–245, 2013.

[4] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, “An evaluation of popular copy-move forgery detection approaches,” *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 6, pp. 1841–1854, 2012.

[5] M. Iuliani, G. Fabbri, and A. Piva, “Image splicing detection based on general perspective constraints,” in *WIFS*. IEEE, 2015.

[6] V. Conotter, G. Boato, and H. Farid, “Detecting photo manipulation on signs and billboards,” in *ICIP*. IEEE, 2010.

[7] D. Cozzolino, G. Poggi, and L. Verdoliva, “Splicebuster: A new blind image splicing detector,” in *Information Forensics and Security (WIFS), 2015 IEEE International Workshop on*. IEEE, 2015, pp. 1–6.

[8] A. E. Dirik and N. D. Memon, “Image tamper detection based on demosaicing artifacts.” in *ICIP*, 2009, pp. 1497–1500.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee, “Improving object detection with deep convolutional networks via bayesian optimization and structured prediction,” in *CVPR*, 2015.

[11] M. Cimpoi, S. Maji, and A. Vedaldi, “Deep filter banks for texture recognition and segmentation,” in *CVPR*, 2015.

[12] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “Hypercolumns for object segmentation and fine-grained localization,” in *CVPR*, 2015.

[13] D. Conigliaro, P. Rota, F. Setti, C. Bassetti, N. Conci, N. Sebe, and M. Cristani, “The s-hock dataset: Analyzing crowds at the stadium,” in *CVPR*, 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *CVPR*, 2014.

[16] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[17] M. Buccoli, P. Bestagini, M. Zanoni, A. Sarti, and S. Tubaro, “Unsupervised feature learning for bootleg detection using deep learning architectures,” in *WIFS*. IEEE, 2014.

[18] D. Luo, R. Yang, and J. Huang, “Detecting double compressed amr audio using deep learning,” in *ICASSP*. IEEE, 2014.

[19] E. Gopi, N. Lakshmanan, T. Gokul, S. KumaraGanesh, and P. R. Shah, “Digital image forgery detection using artificial neural network and auto regressive coefficients,” in *Electrical and Computer Engineering, 2006. CCECE'06. Canadian Conference on*. IEEE, 2006, pp. 194–197.

[20] Y. Huang and Y. Long, “Demosaicking recognition with applications in digital photo authentication based on a quadratic pixel correlation model,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[21] N. Fan, C. Jin, and Y. Huang, “A pixel-based digital photo authentication framework via demosaicking inter-pixel correlation,” in *Proceedings of the 11th ACM workshop on Multimedia and security*. ACM, 2009, pp. 125–130.

[22] J. Lukáš and J. Fridrich, “Estimation of primary quantization matrix in double compressed jpeg images,” in *Proc. Digital Forensic Research Workshop*, 2003, pp. 5–8.

[23] W. Lu, W. Sun, J.-w. Huang, and H.-T. Lu, “Digital image forensics using statistical features and neural network classifier,” in *Machine Learning and Cybernetics, 2008 International Conference on*, vol. 5. IEEE, 2008, pp. 2831–2834.

[24] Z. Zhang, Y. Bian, and X. Ping, “Image blind forensics using artificial neural network,” in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 4. IEEE, 2008, pp. 847–850.

[25] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *CVPR*, 2015.

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv preprint arXiv:1207.0580*, 2012.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[28] CASIA Tampered Image Detection Evaluation Database, 2010, http://forensics.idealtest.org/casiav2/.

[29] G. Cattaneo, G. Roscigno, and U. F. Petrillo, “Experimental evaluation of an algorithm for the detection of tampered jpeg images,” in *Information and Communication Technology*. Springer, 2014, pp. 643–652.

[30] P. Sutthiwan, Y. Q. Shi, H. Zhao, T.-T. Ng, and W. Su, “Markovian rake transform for digital image tampering detection,” in *Transactions on data hiding and multimedia security VI*. Springer, 2011, pp. 1–17.

[31] W. Wang, J. Dong, and T. Tan, “Effective image splicing detection based on image chroma,” in *ICIP*. IEEE, 2009, pp. 1257–1260.

[32] ——, “Image tampering detection based on stationary distribution of markov chain,” in *ICIP*. IEEE, 2010.

[33] Z. Lin, J. He, X. Tang, and C.-K. Tang, “Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis,” *Pattern Recognition*, vol. 42, no. 11, pp. 2492–2501, 2009.

[34] G. Cattaneo and G. Roscigno, “A possible pitfall in the experimental analysis of tampering detection algorithms,” in *Network-Based Information Systems (NBiS), 2014 17th International Conference on*. IEEE, 2014, pp. 279–286.

[35] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *ECCV*. Springer, 2014, pp. 391–405.