Fake Colorized Image Detection with Channel-wise Convolution based Deep-learning Framework

Long Zhuo*, Shunquan Tan* and Jishen Zeng†, Bin Li†
* College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
† College of Information Engineering, Shenzhen University, Shenzhen, China
*† Shenzhen Key Laboratory of Media Security,
Guangdong Key Laboratory of Intelligent Information Processing,
National Engineering Laboratory for Big Data System Computing Technology,
Shenzhen 518060, China.

***Abstract*—Colorization is one remarkable emerging image manipulating technique, which maybe potentially used for illegal purpose. In this paper, we introduce WISERNet (Wider Separate-then-reunion Network), a recently proposed deep-learning based data-driven color image steganalyzer in the field of fake colorized image detection. We believe that statistical inconsistencies introduced by different automatic colorization methods can be captured by advanced deep-learning based data-driven color-image steganalyzers such as WISERNet. Experimental evidences reported in this paper supports our claims: the detection performance of our proposed detector obviously outperforms FCID-HIST and FCID-FE, two state-of-the-art hand-crafted features specific to fake colorized image detection. Please note that in our approach we have never explicitly utilized information from the specific channels other than the ordinary *red*, *green*, and *blue* color channel, which is completely different from prior works in this field.**

I. Introduction

With the wide availability of powerful media editing tools, it becomes much easier to manipulate digital images. Therefore there is an increasing concern about the trustworthiness of digital images. Effective forensic techniques are desperately needed to verify the authenticity, originality, and integrity of digital images.

In this arms race, colorization is one remarkable emerging image manipulating technique, which can generate visually indistinguishable fake colorized images from their grayscale counterparts. There are three state-of-the-art fully automatic colorization methods which require no professional supervision, which makes even ordinary people can produce fake colorized images with high quality: In [1] (referred as method #1), Larsson et al. utilized low-level and semantic representations to colorize the grayscale images in Hue-Chroma-Lightness color space. In [2] (referred as method #2), Zhang et al. trained a CNN (Convolutional Neural Network) to map from a grayscale input to a distribution over quantized color value outputs, and then proposed a classification-style colorization approach. In [3] (referred as method #3), Iizuka et al. proposed the third automatic colorization method by jointly utilizing the local and global priors with an end-to-end deep-learning network. As demonstrated in Fig. 1, the three above mentioned automatic colorization methods can all generate high-quality visually indistinguishable fake colorized images.

Until recently, no forensic technique has been proposed to detect fake colorized images. In [4], Guo et al. conducted the first successful attempt. They pointed out that there are statistical inconsistencies in the hue, saturation, dark and bright channels. Based on their observations, they proposed two detection methods for fake colorized images: histogram-based (FCID-HIST) and feature encoding-based (FCID-FE). Experimental results showed that their proposed methods exhibit a decent performance against the three automatic colorization methods mentioned above [1], [2], [3].

In the last decade, the confrontation between steganography and steganalysis, one adjacent research field besides multimedia forensics remains intense [5], [6], [7]. Since image steganalysis is similar to forged image detection, important steganalytic algorithms have been applied to digital image forensics and have achieved good performance. For instance, in [8], Qiu et al. applied different universal steganalytic hand-crafted features, including the famed spatial-domain rich model (SRM) [6] to varying image forensics tasks and evaluated their performance. According to their report, some advance steganalytic features, e.g. SRM, outperformed the specific forensic methods on hand at that time significantly.

What proposed in [4] is a traditional hand-crafted features based forensic technique. In this paper, we introduce WISERNet (Wider Separate-then-reunion Network), a new deep-learning based data-driven color image steganalyzer [9] in the field of fake colorized image detection. We recently proposed WISERNet to attack true-color image steganography. Experimental results reported in [9] showed that it clearly outperformed other state-of-the-art true-color image steganalyzers, no matter hand-crafted or deep-learning based. We set forth the motivation behind the introduction of WISERNet in fake colorized image detection, and provide experimental evidences to demonstrate its effectiveness compared to hand-crafted FCID-HIST and FCID-FE.

The rest of the paper is organized as follows. In Sect. II

This work was supported in part by the NSFC (61772349, U1636202, 61702340), Shenzhen R&D Program (JCYJ20160328144421330). This work was also supported by Alibaba Group through Alibaba Innovative Research (AIR) Program. *(Corresponding author: Shunquan Tan)*

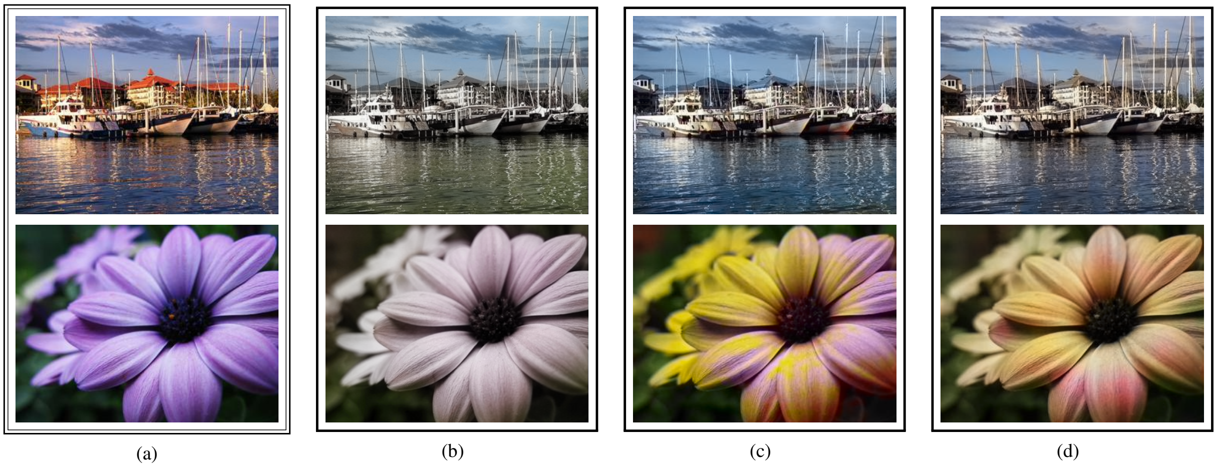

Fig. 1. (a) Ground-truth true-color images. (b) Fake colorized images generated by method #1. (c) Fake colorized images generated by method #2. (d) Fake colorized images generated by method #3.

we firstly provided the motivation behind the introduction of WISERNet, and then describe its detailed structure. Experimental results are presented in Sect. III. Finally, we make a conclusion in Sect. IV.

II. Our proposed deep-learning approach

A. *Motivation*

For the sake of simplicity, in this paper we only consider RGB true-color model. Given a true-color image, it comprises three channels, namely the *red*, the *green*, and the *blue* channel. Researchers have acknowledged that there are many inherent statistical peroperties within natural images, especially those true-color ones. For instance, given a true-color image, the intensity values in the same location of the three channels exhibit strong inherent relationship which is hard to be modelled even with state-of-the-art advanced digital image models. Therefore when fully automatic colorization methods reconstruct the *red*, the *green*, and the *blue* channel from a signle grayscale template, it is inevitable that artifacts are introduced in the inherent statistical peroperties among three color channels. Since true-color image steganographic algorithms add stego noises with feeble energy to three color channels and hence break the inherent statistical peroperties among the channels, the task of true-color image steganalysis is also to expose the artifacts hidden among three color channels of the maliciously manipulated images. From this perspective, the task of detecting fake colorized images is similar to the task of true-color image steganalysis and it is reasonable that we apply state-of-the-art true-color image steganalyzer to detect fake colorized images generated by fully automatic colorization methods.

In term of how difficult the task is, true-color image steganalysis is much tougher than detecting fake colorized images. Firstly, automatic colorization methods construct color channels from nonexistence, while steganography merely add feeble energy to existing real color channels. Therefore undoubtedly the artifacts introduced by colorization methods are much severer than those introduced by true-color image steganography. Secondly, the artifacts introduced by existing automatic colorization methods seem to spread ove the whole scene, even the untruthfulness of the forged true-color scene is sometimes obvious. However state-of-the-art true-color image steganographic algorithms are all content adaptive, which means that they only embed secret data in highly-textured regions, namely only introduce artifacts in the regions hard to be modelled with the technologies nowadays. Therefore it is reasonable to expect that one state-of-the-art true-color image steganalytic algorithm can detect fake colorized images with supreme performance. WISERNet is a new deep-learning based data-driven color image steganalyzer recently proposed by us [9]. Since it is dominant over all other true-color image steganalyzers, we introduced it in the field of fake colorized image detection and expected its excellent performance, as finally has been demonstrated in the experiments.

B. *The channel-wise convolution based fake colorized image detector*

We introduce WISERNet in the field of fake colorized images detection. We originally proposed WISERNet, the wider separate-then-reunion network in [9] to attack color-image steganography. As illustrated in Fig. 2, it takes a true-color image as input and applies channel-wise convolution to the red, green, and blue channel of the input image, respectively. In the channel-wise convolutional layer, the weights of the kernels are initialized with the thirty 5×5 filters used in SRM [6]. The bottom channel-wise convolutional layer corresponds to the "separate" stage of our proposed network. The three separate groups of output channels are then concatenated together to form a ninety-channel input of the second convolutional layer. Started from the second convolutional layer, the upper structure of the network are re-unioned. It is a united wide

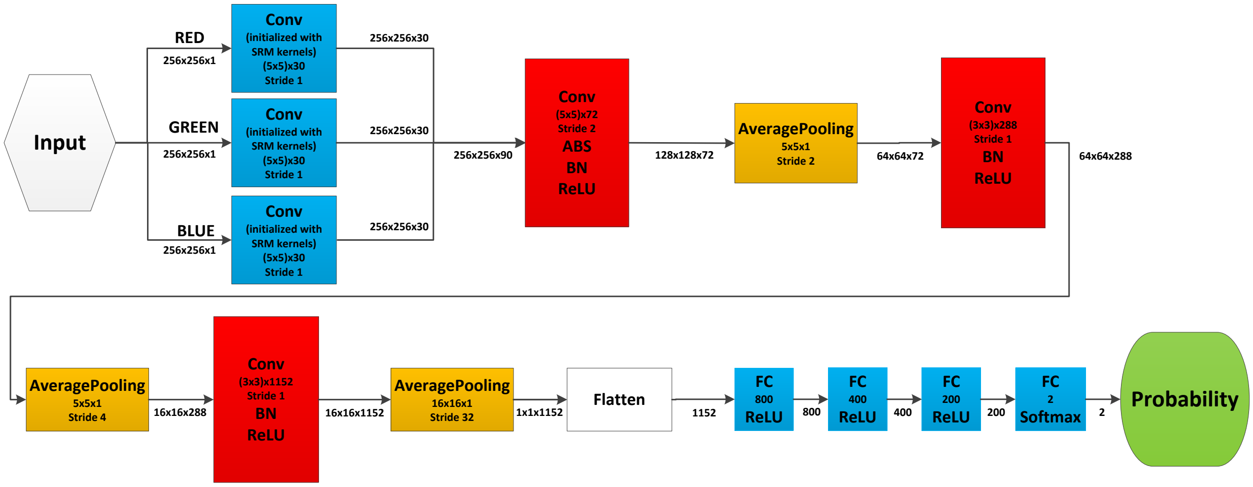

Fig. 2. Conceptual architecture of our proposed channel-wise convolution based fake colorized image detector.

and relatively shallow convolutional neural network, in which the three cascaded convolutional layers contain 72, 288, and 1152 3×3 convolution kernels, and generate output feature maps of size 256×256, 128×128, and 32×32, respectively. Ahead of the top-most normal convolutional layer, the output feature maps are pooled with a large stride with step=32 and then flatten to a 1152 dimensional feature vector. The 1152 dimensional feature vector acts as the input of the top fully-connected network containing 800, 400, 200, and 2 neurons, respectively. The final layer contains two neurons which denote "fake" prediction and "natural" prediction. Softmax function is used to output predicted probabilities.

III. Experiments

All experiments in this paper were conducted on the experimental database of [4] provided in the personal website of Dr. Y. Guo [10]. In the experimental database, the main dataset is *D1*, which contains 10000 natural true-color images randomly selected from ImageNet [11] and their corresponding fake colorized images generated by method #1, #2, and #3. Some demo images from *D1* has been shown in Fig. 1. Besides *D1*, six relatively small datasets, namely *D2*, *D3*, *D4*, *D5*, *D6*, and *D7* are provided to evaluate the performance of our proposed fake colorized image detector under the cover-source mismatching scenario and the colorization method mismatching scenario. Among them, *D2*, *D3*, and *D4* contains 2000 natural true-color images selected from ImageNet and their corresponding fake colorized images generated by method #1, #2, and #3, respectively. The natural images in *D2*, *D3*, and *D4* are guaranteed to be not overlapping with each other. *D5*, *D6*, and *D7* contains 2000 natural true-color images selected from Oxford building dataset [12] and their corresponding fake colorized images generated by method #1, #2, and #3, respectively.

We use the Testing Error Rate (TER) on the corresponding testing dataset to evaluate the performance of our proposed fake colorized image detector. Please note that in the experimental database, the number of the positive samples (fake colorized images) and the number of the negative samples (corresponding natural true-color images) are the same. Therefore TER, the performance metric used in our paper is indeed the same as the Half Total Error Rate (HTER) used in [4].

The implementation of our proposed WISERNet based fake colorized image detector was based on TensorFlow [13]. The detector was trained using mini-batch stochastic gradient descent with "inv" learning rate starting from 0.001 (power: 0.75; gamma: 0.0001; weight_decay: 0.0005) and a momentum fixed to 0.9. The batch size in the training procedure was 32 and the maximum number of epochs was set to 10. When training our proposed detector in *D1* dataset, 7000 natural-fake pairs were randomly selected for training. The remaining 3000 natural-fake pairs were for testing. 1000 natural-fake pairs were further randomly picked out from the training set for validation. For *D1*-*D7*, one model was trained for each one of them. In the training procedure, 1000 natural-fake pairs were for training while the rest were for validating for each dataset.

In Fig. 3, we show how the testing error rates changed with successive training epochs in the experiments which were conducted on dataset *D1*. The tests were performed on standalone testing dataset every 1 training epoch and the models were trained for 10 epochs in total. From Fig. 3 it can be seen when used to attack all of the three automatic colorization methods WISERNet exhibited good convergence and stability after less than 5 epochs. Our proposed fake colorized image detector could achieve as high as less than 5% testing error rate on dataset *D1*. No reports of corresponding

TABLE I
Detection results for the cross colorization method tests, and with different training vs testing sets. If possible, we list the testing error rates of our proposed fake colorized image detector, FCID-HIST, and FCID-FE separated by slashes. If the corresponding results for FCID-HIST and FCID-FE cannot be found in [4], we mark them with "∼".

Train \ Test	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
d2	0.65/22.50/22.30	12.75/28/23.65	5.65/33.95/31.70	1.85/22.85/51.40	14.3/∼	9.7/∼
d3	6.2/26.95/25.10	1.6/24.45/22.85	1.9/41.85/34.25	8.7/∼	6.15/21.50/22.70	6/∼
d4	5.2/38.15/38.50	3.7/43.55/36.15	1/22.35/17.30	5.9/∼	7.6/∼	5.45/30.95/20.20
d5	9.45/43.45/49.80	12.75/∼	9.4/∼	1/∼	4.15/∼	2.5/∼
d6	22.55/30.75/30.25	11.8/∼	12.15/∼	5.4/∼	0.95/∼	1.6/∼
d7	16.6/∼	11.35/∼	9.35/36.60/23.15	1.9/∼	1.55/∼	1.1/∼

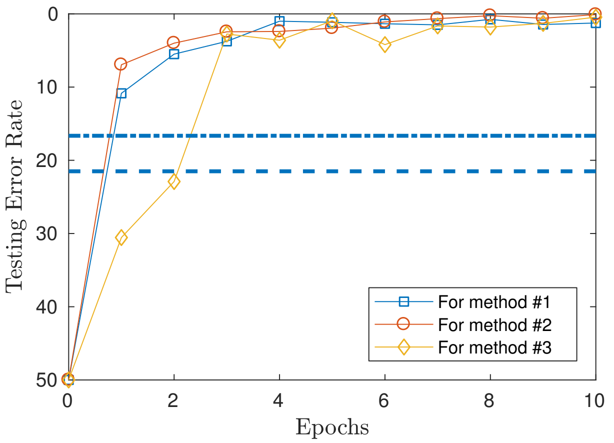

Fig. 3. Testing error rates versus training epochs for our proposed WISERNet based fake colorized image detector. The experiments were conducted on dataset *D1*. The dash-dotted and the dashed reference lines denote the best testing error rate (a.k.a HTER) of FCID-FE and FCID-HIST in different subsets of *D1* respectively, as reported in [4].

performance of FCID-FE and FCID-HIST can be found in [4]. However, the best testing error rates (a.k.a HTER) of FCID-FE and FCID-HIST in different subsets of *D1* (with 1000 natural-fake pairs) was reported and we can use them as a direct, though not very fair comparison. We can see from Fig. 3 that our proposed WISERNet based fake colorized image detector outperformed both FCID-FE and FCID-HIST by a clear margin. The reduction of the testing error rates can be expected to be as large as 15%.

For the sake of completeness, in Tab. I, we give a full comparison of the detection results for the cross colorization method tests, and with different training vs testing sets. From Tab. I we can see no matter under what scenarios, our proposed WISERNet based fake colorized image detector offered a tremendous advantage compared with FCID-FE and FCID-HIST, the two specific hand-crafted features proposed in [4]. Our proposed model experienced the worst performance in the case that trained with *D6* training dataset while tested with *D2* validation dataset. But it is still much better than what reported for FCID-FE and FCID-HIST in [4].

IV. Concluding Remarks

In this paper, we propose a channel-wise convolution based deep-learning framework based fake colorized image detector. The deep-learning framework we adopt is WISERNet, a dominant true-color image steganalyzer. Our experimental results show that such a deep-learning based data-driven framework is well-suited for fake colorized image detection. When used to detect fake colorized images generated by three state-of-the-art automatic colorization methods, it can achieve excellent performance. The performance boost on top of specific hand-crafted features, namely FCID-FE and FCID-HIST is obvious.

Our future work will focus on the incorporation of our proposed fake colorized image detector into the generative adversarial network to propose a forensics-aware deep-learning based automatic colorization method.

References

[1] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 577–593.

[2] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 649–666.

[3] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification", *ACM Transactions on Graphics*, vol. 35, no. 4, p. 110, 2016.

[4] Y. Guo, X. Cao, W. Zhang, and R. Wang, "Fake colorized image detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 1932–1944, 2018.

[5] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *Proc. IEEE 2014 International Conference on Image Processing, (ICIP'2014)*, 2014, pp. 4206–4210.

[6] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[7] J. Zeng, S. Tan, B. Li, and J. Huang, "Large-scale JPEG steganalysis using hybrid deep-learning framework," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1242–1257, 2018.

[8] X. Qiu, H. Li, W. Luo, and J. Huang, "A universal image forensic strategy based on steganalytic model," in *Proc. 2nd ACM Information Hiding and Multimedia Security Workshop (IH&MMSec' 2014)*, 2014, pp. 165–170.

[9] J. Zeng, S. Tan, G. Liu, B. Li, J. Huang, "WISERNet: Wider Separate-then-reunion Network for Steganalysis of Color Images," *arXiv:1803.04805*, 2015. [Online]. Available: http://arxiv.org/abs/1803.04805

[10] Y. Guo, "Yuanfang Guo's personal website," https://eeandyguo.github.io/, [Online].

[11] "ImageNet," http://image-net.org/, [Online].

[12] "The Oxford Buildings Dataset," http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/, [Online].

[13] "TensorFlow™," https://www.tensorflow.org/, [Online].