A universal detector of CNN-generated images using properties of checkerboard artifacts in the frequency domain

1st Miki Tanaka
Tokyo Metropolitan University
Tokyo, Japan
tanaka-miki@ed.tmu.ac.jp

2nd Sayaka Shiota
Tokyo Metropolitan University
Tokyo, Japan
sayaka@tmu.ac.jp

3rd Hitoshi Kiya
Tokyo Metropolitan University
Tokyo, Japan
kiya@tmu.ac.jp

Abstract—We propose a novel universal detector for detecting images generated by using CNNs. In this paper, properties of checkerboard artifacts in CNN-generated images are considered, and the spectrum of images is enhanced in accordance with the properties. Next, a classifier is trained by using the enhanced spectrums to judge a query image to be a CNN-generated ones or not. In addition, an ensemble of the proposed detector with emphasized spectrums and a conventional detector is proposed to improve the performance of these methods. In an experiment, the proposed ensemble is demonstrated to outperform a state-of-the-art method under some conditions.

Index Terms—GAN, CNN, checkerboard artifact, fake image,

I. INTRODUCTION

Convolutional neural networks (CNNs) have led to major breakthroughs in a wide range of applications. In contrast, they have generated new concerns and problems. Recent rapid advances in deep image synthesis techniques, such as generative adversarial networks (GANs) have easily generated fake images, so detecting manipulated images has become an urgent issue [1], [2].

To overcome this issues, we propose a universal detector of images generated by using CNNs in this paper. Recently, CNN-generated images were investigated to include a trace of checkerboard artifacts [3]–[6], although the trace is weak in general. We focus on a trace of checkerboard artifacts in the frequency domain to detect CNN-generated images.

In the proposed method, the spectrum of images is enhanced in accordance with properties of checkerboard artifacts, and a classifier is trained by using the enhanced spectrums. In addition, an ensemble of the trained detector and a state-of-the-art detector [7] is proposed for improving the performance of the detectors. In an experiment, the proposed ensemble is demonstrated to outperform the state-of-the-art one under the use of 11 models.

II. RELATED WORK

A. Generator model

A variety of generator models have been proposed for image generation and image-to-image translation. Models based on variational autoencoder (VAE) or GAN are typical image generator models [8], [9]. Autoencoders including VAE translate an image into latent variable $z$ by using an encoder, and generate an image from $z$ by using a decoder. Since $z$ has the standard normal distribution, VAE can generate images from a noise having the standard normal distribution by the decoder. Deepfakes [10] with VAE become a major threat to the international community.

GAN models estimate a generative model and a discriminative model via an adversarial process. PGGAN [11], BigGAN [12], StyleGAN [13] and StyleGAN2 [14] generate high resolution images from a random noise vector. In addition, CycleGAN [15] and StarGAN [16] translate an image from a source domain to a target domain, e.g. changing apples to oranges. GauGAN [17] were proposed to generate an image from an input semantic layout.

B. Detecting CNN-generated images

The first approach for detecting CNN-generated images was inspired by photo-response non-uniformity noise (PRNU) that was used for discriminating camera devices [18], [19]. This approach enables us to discriminate GAN-generated images from fingerprints caused by GAN, and assume that the same GAN models are used for training and testing images.

To universally detect CNN-generated images even when images are not ones generated from a model used for training the detector, a universal detector was proposed, where it was trained by using images generated only by AutoGAN [20]. In this work [20], the use of the frequency domain was demonstrated to improve the performance of the detector. In contrast, a universal detector by training only one specific GAN (PGGAN) was proposed [7], where RGB images are directly used for training a detector. This method was shown to detect not only PGGAN but also other generator models that were not used for training the detector.

III. PROPOSED DETECTOR

A. Detector with enhanced spectrum

Figure [1] shows an overview of the proposed detector with enhanced spectrums. For training a classifier, a novel en-
hancement method for clearly showing checkerboard artifacts included in CNN-generated images is applied to training image \( I_i \). Enhanced spectrum \( E_i \) calculated from \( I_i \) are used for training the classifier. For testing, a query image \( I_q \) is enhanced as well as for training, the spectrum \( F_Q \) calculated from \( I_q \) is inputted to the trained classifier to judge it to be a CNN-generated one or not in accordance with an outputted probability score \( r_F \in [0, 1] \).

Fig. 1. Overview of proposed detector (single)

Enhanced spectrum \( F_i \) is calculated as below.

1) A median filter with a size of \( 5 \times 5 \) is applied to \( I_i \), and the difference between \( I_i \) and \( I'_i \) is calculated, where \( I'_i \) is an output image from the median filter.

2) Cropping \( I'_i \) into \( L \) rectangles with a size of \( N \times N \) at random positions to generate \( L \) images with a size of \( N \times N \), \( I_{1i}, \ldots, I_{Li} \).

3) Applying \( N \times N \)-DFT to \( I_{1i}, \ldots, I_{Li} \) to obtain their spectrums \( F_{1i}, \ldots, F_{Li} \).

4) Computing enhanced spectrum \( E_i \) as.

\[
E_i = \sum_{n=1}^{L} \log_{10} |F^n_i| \tag{1}
\]

Similarly, enhanced spectrum \( E_q \) is calculated by using query \( I_q \) (see Fig 2).

Fig. 2. Example of enhanced spectrum

From properties of checkerboard artifacts, CNN-generated images are confirmed to have the influence of checkerboard artifacts at the same positions in the frequency domain for all cropped images. Accordingly, averaging spectrums as shown in the procedure can enhance the features that CNN-generated images have.

B. Ensemble of RGB image and enhanced spectrum

A state-of-the-art method for detecting CNN-generated images [7] is directly carried out with RGB images. In this paper, we also propose an ensemble of this conventional detector and the detector with enhanced spectrums.

An overview of the ensemble is shown in Fig. 3. Final probability score \( r \) is calculated from \( r_I \) and \( r_F \) in accordance with Algorithm 1, where \( r_I \) and \( r_F \) are scores from the detector with enhanced spectrums and the detector with RGB images, respectively. Each detector has its own strengths and weaknesses, so this ensemble is expected to improve the performance of each detector.

Algorithm 1 the ensemble algorithm

1: if \((|r_I - 0.5| > |r_F - 0.5|) \) then
2: \( r = r_I \)
3: end if
4: if \((|r_I - 0.5| < |r_F - 0.5|) \) then
5: \( r = r_F \)
6: end if

Fig. 3. Proposed detector (ensemble)

IV. EXPERIMENT RESULTS

A. Experiment setup

In this experiment, the dataset prepared by Wang et al. [7] was used, where the training dataset consist of 720K images generated by using PGGAN [11], and test datasets consisted of 4K images in which 2K images are generated by using 11 models as CNN-generated images, and the others are images captured by cameras.

The performance of detectors was evaluated by using F-score and average precision (AP). F-score is given by,

\[
\text{F-score} = \frac{2RP}{R + P}, \tag{2}
\]

where \( P \) and \( R \) are the precision and recall at a selected threshold \( th \), and \( th = 0.5 \) was selected for this evaluation.
AP is also computed by summarizing a precision-recall curve as the weighted mean of precisions achieved at each threshold:

$$AP = \sum_j (R_j - R_{j-1}) P_j,$$  \hspace{1cm} (3)

where $P_j$ and $R_j$ are the precision and recall at the $j^{th}$ threshold, and an input image is judged to be a CNN-generated image, when the probability score $r = [0, 1]$ of the input image is higher than $th$.

**B. Experiment results**

Table I and II show AP and F-score values under the use of test images from each model. Table I shows the proposed method (single) outperformed Wang’s method under a number of models that had similar network structures to PGGAN. In addition, the proposed method (ensemble) outperformed Wang’s method for almost all models. Wang’s method and the proposed method (single) have their own strengths and weaknesses for detecting CNN-generated images, respectively. In contrast, the ensemble enables us to adopt only strong points of each method. Table III also shows mean values calculated from AP values of 11 models. From the table, the proposed detector with enhanced spectrums was confirmed to improve the accuracy of Wang’s method, although the models were trained only by using images from PGGAN. The difference between Table I and Table II was caused due to the difference in the selection of threshold values.

**V. Conclusion**

We proposed a universal detector with enhanced spectrums for detecting CNN-generated images. The proposed ensemble was confirmed to outperform the state-of-the-art under the use of 11 models, where the classifier for the detection was trained by using image generated only from one model, PGGAN.

**REFERENCES**

[1] L. Verdoliva, “Media forensics and deepfakes: An overview,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

[2] M. Tanaka and H. Kiya, “Fake-image detection with robust hashing,” in Proc. of IEEE 3rd Global Conference on Life Sciences and Technologies, 2021, pp. 40–43.

[3] Y. Sugawara, S. Shiota, and H. Kiya, “Super-resolution using convolutional neural networks without any checkerboard artifacts,” in Proc. of IEEE International Conference on Image Processing, 2018, pp. 66–70.

[4] Y. Kinoshita and H. Kiya, “Fixed smooth convolutional layer for avoiding checkerboard artifacts in cnns,” in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 3712–3716.

[5] Y. Sugawara, S. Shiota, and H. Kiya, “Checkerboard artifacts free convolutional neural networks,” APSIPA Transactions on Signal and Information Processing, vol. 8, p. e9, 2019.

[6] T. Osakabe, M. Tanaka, Y. Kinoshita, and H. Kiya, “CycleGAN without checkerboard artifacts for counter-forensics of fake-image detection,” in International Workshop on Advanced Imaging Technology (IWAIT) 2021, vol. 11766, International Society for Optics and Photonics. SPIE, 2021, pp. 51 – 55.

[7] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-generated images are surprisingly easy to spot... for now,” in Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” arXiv preprint arXiv:1406.2661, 2014. [Online]. Available: https://arxiv.org/abs/1406.2661.

[9] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114.

[10] Deepfakes github. [Online]. Available: https://github.com/deepfakes/faceswap.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” arXive preprint arXiv:1710.10196, 2018. [Online]. Available: https://arxiv.org/abs/1710.10196.

[12] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” arXive preprint arXiv:1809.11096, 2019. [Online]. Available: https://arxiv.org/abs/1809.11096.

[13] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2019.

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. of IEEE International Conference on Computer Vision, Oct 2017.

[16] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, June 2018.

[17] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2019.

[18] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do gains leave artificial fingerprints?” in Proc. of IEEE Conference on Multimedia Information Processing and Retrieval, 2019, pp. 506–511.
[19] N. Yu, L. Davis, and M. Fritz, “Attributing fake images to gans: Learning and analyzing gan fingerprints,” in Proc. of IEEE/CVF International Conference on Computer Vision, 2019, pp. 7555–7565.

[20] X. Zhang, S. Karaman, and S. Chang, “Detecting and simulating artifacts in gan fake images,” in Proc. of IEEE International Workshop on Information Forensics and Security, 2019, pp. 1–6.