Hindawi
Security and Communication Networks
Volume 2021, Article ID 5511435, 10 pages
https://doi.org/10.1155/2021/5511435

Research Article

Detection of GAN-Synthesized Image Based on Discrete Wavelet Transform

Guihua Tang, Lei Sun, Xiuqing Mao, Song Guo, Hongmeng Zhang, and Xiaoqin Wang

Information Engineering University, Zhengzhou 450001, China

Correspondence should be addressed to Lei Sun; sl0221@sina.com

Received 25 January 2021; Revised 13 April 2021; Accepted 3 June 2021; Published 16 June 2021

Academic Editor: Beijing Chen

Recently, generative adversarial networks (GANs) and its variants have shown impressive ability in image synthesis. The synthesized fake images spread widely on the Internet, and it is challenging for Internet users to identify the authenticity, which poses huge security risk to the society. However, compared with the powerful image synthesis technology, the detection of GAN-synthesized images is still in its infancy and face a variety of challenges. In this study, a method named fake images discriminator (FID) is proposed, which detects that GAN-synthesized fake images use the strong spectral correlation in the imaging process of natural color images. The proposed method first converts the color image into three color components of R, G, and B. Discrete wavelet transform (DWT) is then applied to RGB components separately. Finally, the correlation coefficient between the subband images is used as a feature vector for authenticity classification. Experimental results show that the proposed FID method achieves impressive effectiveness on the StyleGAN2-synthesized faces and multitype fake images synthesized with the state-of-the-art GANs. Also, the FID method exhibits good robustness against the four common perturbation attacks.

1. Introduction

With the remarkable development of artificial intelligence (AI) and progress of high-performance computing hardware, image synthesis technology has evolved dramatically. The Internet users share a large number of multimedia contents on social media every day. It is challenging to identify authenticity of these contents, posing huge security risk to social. In particular, the generative adversarial networks (GANs) proposed in 2014 [1] have spawned a new type of image synthesis method. The images synthesized by four typical GANs are shown in Figure 1, which are really hard for humans to distinguish at the first glance. Besides, GAN's powerful image synthesis and editing capabilities bring new industrial value. For example, it can be used to create virtual characters, perform video rendering and sound simulation in film production, and create a new way of communication. However, security and privacy concerns are also raised. If these fake contents are disseminated as news materials, they will damage the reputation of news organizations and the public's confidence in the media and even mislead the public opinion and disturb the social order. The increasingly open network environment creates an ideal space for the spread of fake information. In the countries such as Britain and France, there have been cases of using deep-learning forgery technology to produce fake images, deceive the public to even conduct espionage. The hazard and impact of synthesized images has spread throughout the world, resulting in ethical, legal, and security problems. It is extremely urgent to find effective techniques for detection of fake images.

GAN-synthesized images show impressively high quality. Accordingly, the detection of GAN-synthesized images has become a hot research field. Various detection methods for GAN-synthesized images have been proposed successively [2–5] and achieved good results. However, with the increasing variety and quality of GAN-synthesized images, as well as the various perturbation attacks, these methods begin to expose their limitations.

To overcome the limitation in existing methods for detecting GAN-synthesized images, a method named fake images discriminator (FID) is proposed in this study. The FID method relies on both the discrete wavelet transform (DWT) and the standard correlation coefficient to extract the spectral correlation

Figure 1: Fake images synthesized with various GANs.

of natural color images. Besides, the support vector machine (SVM) was used for classification. Experimental results show that the FID method outperforms prior works of AutoGAN [6] and FakeSpotter [7] on StyleGAN2-synthesized faces and maintains robustness in tackling four common perturbation attacks. An additional experiment is conducted on images forged by other state-of-the-art (SOTA) GANs, and the FID method also achieves good effectiveness on multiple types of fake images.

The main contributions of this study are as follows:

(1) FID method: the fake images discriminator (FID) method employs the DWT and the standard correlation coefficient to detect fake images. Through the analysis of the imaging process of natural color images, it is found that the spectral correlation between RGBs can be utilized to distinguish GAN-synthesized images, which is also robust against the four common perturbation attacks at various intensities.

(2) The first comprehensive evaluation is on typical GAN-synthesized images. Experiments are conducted on high-quality fake images synthesized with SOTA GANs. These fake images include faces, buildings, animals, natural scenes, and so on. Experimental results indicate good effectiveness and robustness of the proposed FID method.

(3) Extensibility: the FID method is based on the imaging process of natural color images and the analysis on the difference between real and GAN-synthesized images. This difference may be widespread in fake images, and it could be extended to other AI-synthesized images and DeepFake.

The rest of the study is organized as follows. Section 2 reviews the related literature of GAN-synthesized images and detection methods. Section 3 describes the imaging process of digital images, followed by the presentation of the proposed FID method in Section 4. The experimental results and analysis are illustrated in Section 5. Section 6 concludes the study.

2. Related Work

Digital image forensics is a technology that distinguishes the authenticity, completeness, and source of image content. It mainly includes active forensics technology and passive (blind) forensics technology [8]. Active forensics is suitable for an image authentication scenario where digital signatures, digital watermarks, or digital fingerprinting have been embedded in digital images in advance. But in the actual environment, most images do not have embedded prior information, which limits the application of active forensics technology. Passive forensics does not require any prior information, and the images are identified based on the changes of image characteristics caused by the forgery operation. Currently, most of the detection methods for GAN-synthesized images conforms to the passive forensics. In the following sections, the latest developments in GAN-synthesized images and image forgery detection methods will be discussed.

2.1. GAN-Based Images Synthesis Methods. Generally, the GAN contains a generator and a discriminator. The generator synthesizes images and the discriminator differentiates between the fake and real images. The generator and discriminator play game mutually and finally achieve a dynamic balance. Since it is first proposed in 2014, the GAN has shown an impressive ability in image synthesis, the most studied area of GAN applications.

Entire face synthesis means that a facial image can be wholly synthesized with GANs, and the synthesized faces do not exist in the world. In entire face synthesis, the progressive growing of GANs (PGGAN) [9] and style-based generator architecture for GANs (StyleGAN) [10, 11], released by NVIDIA, produce an unprecedented high-quality and high-resolution entire synthesis face. As one of the models that can generate images with highest quality, StyleGAN has a new generator architecture proposed by NVIDIA. Without affecting other layers, the input of each layer is modified separately to control the visual features represented by each layer. CycleGAN [12] has achieved remarkable success in image-to-image conversion in two domains. Since each pair of image domains requires independent modeling, the scalability and robustness of CycleGAN are limited for processing of more than two domains. STGAN [13] and StarGAN [14] focus on face editing through manipulating the attributes and expressions of humans' faces, such as changing the color of hair, facial decorations, and expressions. StarGAN designed a generator of star structure to perform image-to-image conversion for multiple domains. The unified model architecture of StarGAN allows training datasets from multiple domains simultaneously in a single network. STGAN aims to improve the accuracy and quality of attribute manipulation. FaceApp, ZAO, and FaceSwap employ GANs to produce DeepFake which involves the swap of person's face [15, 16].

GANs can be applied in numerous aspects of image synthesis and swapping personal identities. In many cases, the fake images synthesized with SOTA GANs are nearly indistinguishable to humans. We cannot believe our eyes anymore in the media.

2.2. Detection of GAN-Synthesized Images. Traditional forensics-based techniques [17–19] usually analyze the traces inducted in image synthesis and inspect the pixel-level

disparities in real and fake images. Compared with traditional fake images, GAN-synthesized images have better quality, and no traces are inducted in image mosaic. Therefore, the effectiveness of these detection methods is greatly reduced. Also, these methods are sensitive to perturbation attacks like blur that is common in media images.

Nataraj et al. [3] built a pixel-level image detection model based on the deep neural network (DNN) and detected GAN-synthesized images by extract co-occurrence matrices on three color channels in the pixel domain. McCloskey et al. [2] found that the frequency of saturated pixels in GAN-synthesized images is limited due to the normalization operation in the generator. Also, the statistical relationship of color component of GAN-synthesized images is different from natural images. Though corresponding detection strategies are designed using these two clues, it is vulnerable to noise and adversarial examples attacks.

Another way to detect GAN-synthesized images is to learn the difference between real and fake images with DNN. Stehouwer et al. [20] introduced an attention mechanism to improve facial forgery detection and manipulated region localization. Wang et al. [21] used ResNet-50 to design a binary classifier to detect images synthesized by the convolutional neural network (CNN). Zhang et al. [6] explored the fingerprint of GAN [22] and proposed a classifier model named AutoGAN based on the input of frequency spectrum. AutoGAN identifies the artifacts inducted in the upsampling component of the GAN so as to realize the detection of GAN-synthesized images. The DNN-based methods [21, 23, 24] achieve better performance than the methods based on traditional image forensics and pixel-level differences. Other work explores various special features to study the disparities between real and synthesized facial images. For example, the uncoordinated facial features of the fake faces is exposed through the facial landmarks [4]. Lyu et al. [5] used the difference in head pose as the classification characteristic. However, GAN technology progresses rapidly, making the GAN features extracted by the above detection methods hard to keep good durability and universality. Besides, these works are vulnerable to common perturbation attacks, and robustness is essential for detecting fake images in the wild. The FakeSpotter proposed by Wang et al. [7] depends on monitoring neuron behaviors to spot AI-synthesized fake faces. This approach exhibited effectiveness on SOTA GANs and robustness against perturbation attacks.

3. Study on Spectral Correlation of Digital Imaging

Spectral correlation means the correlation existing between the three color components in finite neighboring pixels of color images. In the color imaging system, most consumer-grade digital cameras use one CCD or CMOS, and the imaging process of natural color image is shown in Figure 2.

The single-sensor camera obtains the color information of the image through a color filter array (CFA). The Bayer CFA is the most widely used array, using an alternate sampling mode, the RGB components are shown in Figure 3. The number of sampling in the G channel is twice of that in

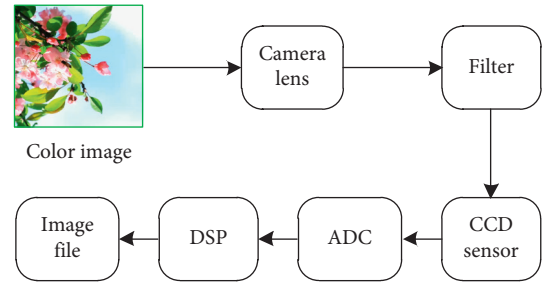

Figure 2: The single-CCD imaging process.

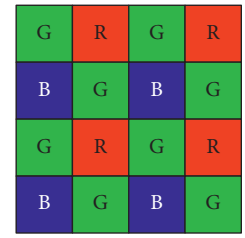

Figure 3: The Bayer CFA.

the R and B channels, which conforms to the spatial sensitivity of the human visual system to different spectral wavelengths. Since only one color component is captured per pixel, the CFA interpolation algorithm is needed to calculate the missing two color values at the pixel.

The main task of the CFA interpolation algorithm is the reconstruction of RGB images, specifically, to estimate the missing two color values from the neighborhood pixels. There are many CFA interpolation algorithms, such as the nearest neighbor, bilinear, bicubic, and convolution interpolation algorithms. These algorithms perform interpolation mainly in the neighborhood of a one-color channel. Taking bilinear algorithm as an example, each color component of R_1 is estimated as follows:

$$R_1 = R_1, \tag{1}$$

$$G(R_1) = \frac{G_1 + G_2 + G_3 + G_4}{4}, \tag{2}$$

$$B(R_1) = \frac{B_1 + B_2 + B_3 + B_4}{4}. \tag{3}$$

This example illustrates that the estimated color component is directly related to the value of the color pixels captured in the neighborhood, so there must be a strong spectral correlation between all RGB pixels of a real image. No matter which CFA interpolation algorithm is used to reconstruct the digital color image, all involve the neighborhood sampling values of 3 color components when estimating the missing color component, which leads to a strong spectral correlation existing in the R, G, and B channels.

Unlike the generation process of natural color images, the GAN trains the network with a large amount of data to synthesize images, which inevitably lead to the differences in some features, especially the spectral correlation between

RGB components of color images. To further prove the differences between GAN-synthesized images and real images, four types of GAN-synthesized images and real images, respectively, performed DWT in RGB channel, and the kernel density curve of transformed RGB components is shown in Figure 4.

Each figure includes three curves, representing the kernel density curve of the R, G, and B. The first row shows the RGB component distribution of the GAN-synthesized images; the RGB component of the second row is from the real images. It can be seen that the real image has similar kernel density curve on the three color channels, and the peaks and valleys appearing areas are highly coincident. The RGB components of the GAN-synthesized images are relatively independent, and the correlation cannot be clearly seen.

In conclusion, strong spectral correlation between RGB is caused by the interpolation operation in the color imaging process, while GAN-synthesized images do not have this characteristic. Therefore, GAN-synthesized images can be recognized based on this difference.

4. Our Method

The imaging process of natural color image causes high spectral correlation. In contrast, synthesizing fake images with the GAN can weaken or even eliminate this correlation. Consequently, the proposed method for detecting GAN-synthesized image employs wavelet multiscale decomposition to extract the correlation characteristics between the spectra of RGB channels. The FID method includes two stages of feature extraction and classification. The block diagram of this method is shown in Figure 5.

4.1. Features Extraction. DWT can decompose an image into subband coefficients that represent different direction information in same scale. Decomposing the two-dimensional image $f(x, y)$ with DWT, it can obtain

$$f(x, y) = W_j^A + \sum_{k \geq j} \left(W_k^H + W_k^V + W_k^D \right), \tag{4}$$

where W^A is the low-frequency approximation under scale j, and W_k^i, $i = \{H, V, D\}$, $k \geq j$ is the detailed component in the horizontal, vertical, and diagonal directions under different scales of the image. The multiresolution decomposition capability of pyramid wavelet transform can decompose the image information layer by layer, so it is widely used to extract image features, especially the statistical features in the spatial domain.

DWT is utilized to construct the correlation between the frequency spectrums of images in the three color spaces. Also, the correlation coefficient is used to measure the constructed correlation. The specific feature extraction process is described as follows:

(1) RGB channels separation: since a stronger statistical correlation of the three color components exists in the RGB color space. The color image is first converted into the three independent color components of R, G, and B.

(2) DWT: each color component is decomposed by level-1 DWT and divided into four subband images (plus the low-frequency approximation itself). Therefore, 12 subband images can be obtained from a color image.

(3) Calculate the correlation coefficient matrix F_{NCC}. The co-correlation coefficient is a basic measure of correlation. The standard correlation function is used to measure the correlation between the subband images of the three color components. The detailed calculation process is shown in Figure 6.

The correlation coefficient $NCC(I_1, I_2)$ corresponds subband image of two color components, and its calculation is shown in equation (2). After calculating all wavelet subband images, 3 correlation coefficient matrix F_{NCC} can be obtained.

$$NCC(I_1, I_2) = \frac{\sum_{I_1} \sum_{I_2} (I_1 - E(I_1))(I_2 - E(I_2))}{\sqrt{\sum_{I_1} (I_1 - E(I_1))^2} \sqrt{\sum_{I_2} (I_2 - E(I_2))^2}}. \tag{5}$$

$E(I_1)$ and $E(I_2)$ in equation (2) are the means of gray images I_1 and I_2, respectively. The calculation is shown in equation (3). $M \times N$ is the image size.

$$E(I) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} I(i, j). \tag{6}$$

(4) Extracting matrix feature: by calculating the four matrix features (kurtosis, mean, skewness, and standard deviation) of real and GAN-synthesized images separately, it is found that the real and GAN-synthesized images have the largest difference in kurtosis feature, which can better distinguish the real and GAN-synthesized images. The experimental results are shown in Figure 7.

The experimental results show that the difference between real and GAN-synthesized images in kurtosis is the largest. Therefore, the kurtosis ku of F_{NCC} is chosen as the final measurement for spectrum correlation of the color image, and its calculation is shown as follows:

$$ku = \frac{E\left[\left(f(i, j) - (1/M \times N) \sum_{i=1}^{M} \sum_{j=1}^{N} f(i, j) \right)^4 \right]}{\left(E[(f(i, j) - \mu)2] \right)^2}, \tag{7}$$

where $f(i, j)$ represents the element of F_{NCC}, and the size of F_{NCC} is $M \times N$. Three kurtosis values can be obtained by calculating the kurtosis of the correlation matrix F_{NCC} (RG), F_{NCC} (RB), and F_{NCC} (GB), respectively.

4.2. Classification. SVM is commonly used for pattern recognition, classification, and regression analysis. LibSVM [25] is a tool library for SVM developed by Professor Chih-Jen Lin in 2001, which can be used for data classification or

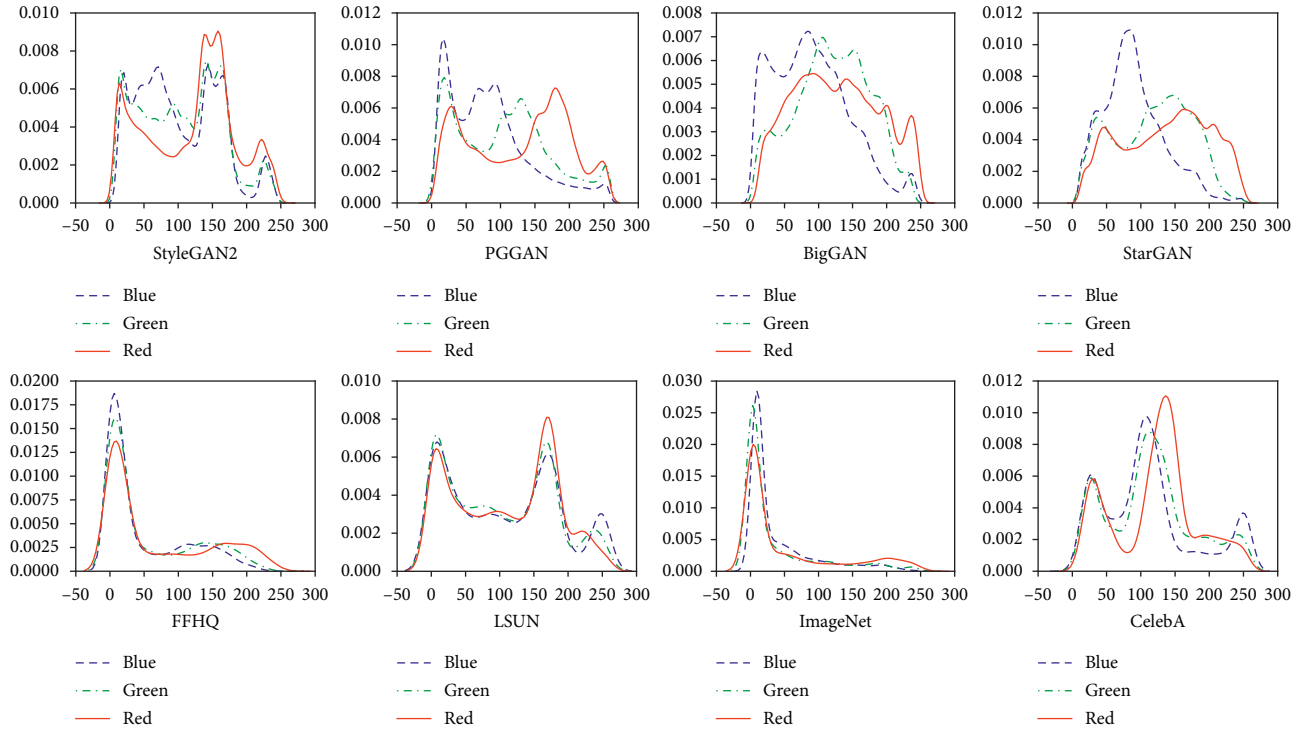

Figure 4: The kernel density curve for different color components.

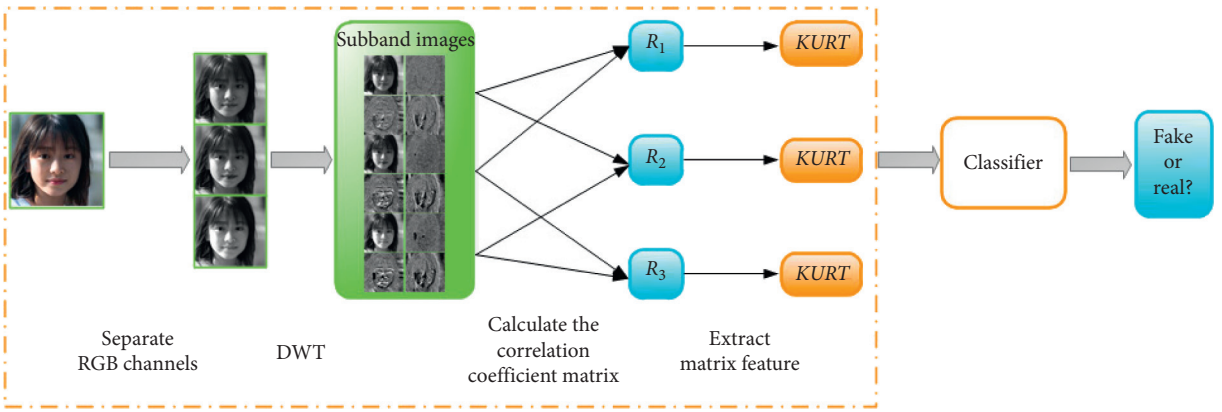

Figure 5: Block diagram of the proposed method.

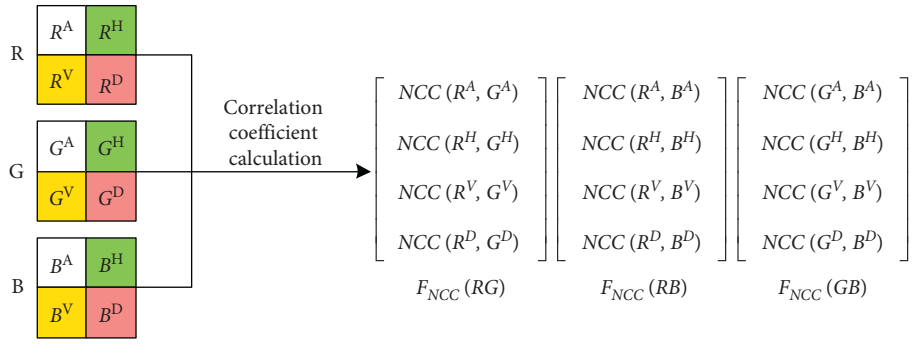

Figure 6: Calculation of the correlation coefficient matrix.

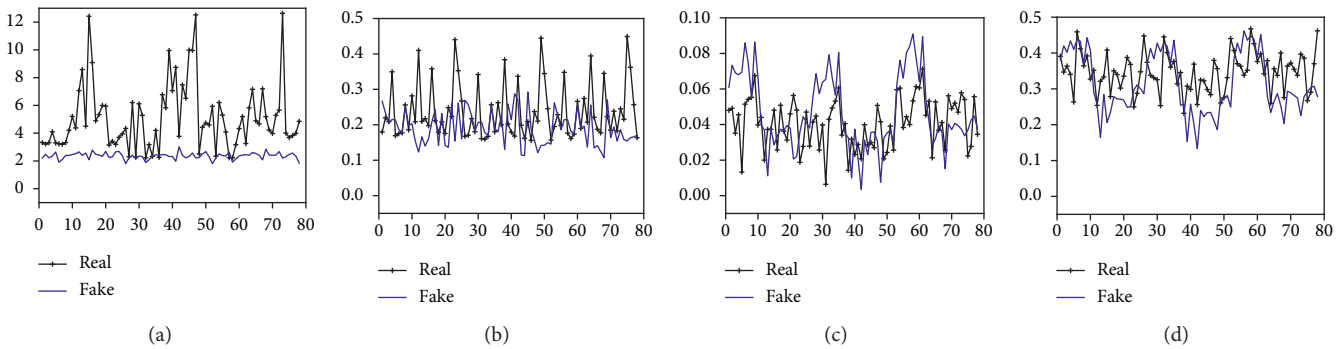

Figure 7: Matrix features of real and GAN-synthesized images. (a) Kurtosis. (b) Mean. (c) Skewness. (d) Standard deviation.

regression conveniently. Since the focus of this study is to employ DWT for feature extraction, there is no special requirement for the classification, and the final feature used for classification is a set of three-dimensional vectors in a simple form. Therefore, LibSVM is used in this study to implement a simple binary classifier, and the radial basis function (RBF) kernel is used to train the SVM for classification.

$$\kappa\left(x_i, x_j\right) = e^{-g\left\|x_i - x_j\right\|^2}, \quad g > 0, \tag{8}$$

where x_i, x_j is a vector; g is the only hyperparameter of RBF; $\|x_i - x_j\|$ indicates the vector norm. The grid-search method is used to optimize the parameters.

5. Result and Analysis

In this section, experiments are conducted to evaluate the effectiveness of the proposed FID method in detecting GAN-synthesized images and its robustness against the four common perturbation attacks. First, experiments are conducted on StyleGAN2-synthesized faces, and the results are compared with that of recently published work, i.e., AutoGAN and FakeSpotter.

5.1. Experimental Setup

5.1.1. Data Collection. For the experiment, real faces are collected from CelebFaces Attributes Dataset (CelebA) [26] due to its good diversity. StyleGAN2 is used to synthesize fake faces. To ensure the diversity and high-quality of the fake image dataset, the various images produced by other newest GANs (e.g., StarGAN and PGGAN) are used. Table 1 presents statistics of the collected fake image dataset from [21]. The first column shows the data type, where variety means that there are more than ten different types of fake images (e.g., building, animals, airplane, and so on). The second column denotes the source of real faces for synthesizing fake images. The last column indicates the source of synthesized fake images, released by official, collected from online, or synthesized by ourselves.

5.1.2. Implementation Details. Binary classifier is implemented by LibSVM for detecting fake images, and the kernel function is RBF. The training dataset includes 5,000 real and 5,000 StyleGAN2-synthesized faces and 1,000 real and 1,000 StyleGAN2-synthesized faces for test. The training dataset and the test dataset are employed for evaluating the effectiveness and robustness of the FID method. Four common perturbation attacks are selected to evaluate the robustness, namely, compression, blur, resizing, and adding noise.

Table 1: Dataset description.

Fake images	GAN type	Collection	Total
Entire synthesis faces	StyleGAN2	Officially released	6 k
Bedroom Cat Car	StyleGAN	[21]	12.0 k
Variety Apple Horse Orange Summer Winter Zebra	BigGAN [27]	[21]	4.0 k
	CycleGAN	[21]	2.6 k
Variety	GauGAN [28]	[21]	10.0 k
Variety	PGGAN	[21]	8.0 k
Variety	PixelRNN	Self-synthesis	3 k
Edited faces	StarGAN	[21]	4.0 k
Edited faces	DiscoGAN	Self-synthesis	1.2 k

5.1.3. Evaluation Metrics. In detecting StyleGAN2-synthesized faces, eight popular metrics are adopted to obtain a comprehensive performance evaluation of the FID method. Also, the performance is compared with prior works, i.e., AutoGAN and FakeSpotter. Specifically, the precision, recall, F1-score, accuracy, AP (average precision), AUC (area under curve of receiver operating characteristics), FPR (false-positive rate), and FNR (false-negative rate) are reported. The AUC is also used as a metric to evaluate the performance of the FID method in tackling the four perturbation attacks and detecting other GANs-synthesized images.

5.2. Detection Performance. In the section, the influence of DWT levels for detecting StyleGAN2-synthesized face is first explored. In the feature extraction stage, 1000 real and 1000 StyleGAN2-synthesized faces are subjected to multilevel

DWT, and the AUC score is adopted to evaluate the performance. The experimental results are shown in Figure 8. The overall value of AUC fluctuates with the increase of the DWT level. The AUC score is the highest when the DWT level equals 1, so the level-1 DWT is selected to extract the spectral correlation.

The performance of the three methods, i.e., the FID, AutoGAN, and FakeSpotter, in detecting StyleGAN2-synthesized faces is measured, and the result is given in Table 2. AutoGAN is an open-source work published in 2019 that exploits the artifacts in GAN-synthesized images and detects the fake images with a classifier based on the deep neural network. FakeSpotter spots AI-synthesized fake faces through monitoring the neuron behaviors. Experimental results demonstrate that the FID method outperforms AutoGAN and FakeSpotter for all eight metrics, achieving competitive performance with a high detection rate and low false alarm rate in detecting the StyleGAN2-synthesized faces.

To illustrate the performance of the FID method in balancing the precision and recall, the precision and recall curves are presented in Figure 9 as well. The proposed method achieves a good balance between precision and recall on StyleGAN2-synthesized faces.

5.3. Robustness Analysis. Since image transformations are common, especially in the social media, the objective of robustness analysis is to evaluate the capabilities of the FID method against perturbation attacks. Four different perturbation attacks (compression, blur, resizing, and adding noise) under different intensities are used for evaluation, and the AUC is taken as a metric for the performance evaluation.

As for the four perturbation attacks, the compression quality measures the intensity of compression. 0 and 100, respectively, are the maximum and minimum values. Blur indicates that the Gaussian blur is employed to faces. The value of Gaussian kernel standard deviation is adjusted to control the intensity of blur, and the Gaussian kernel size is (3, 3). In resizing, the scale factor is applied to control the size of an image in horizontal and vertical axes. The Gaussian additive noise is added to produce noisy images, and the variance is used for controlling the intensity of the noise.

The experimental results of the FID method against the four common perturbation attacks are shown in Figure 10. As the intensity of perturbation attacks increases, the AUC score of the FID method fluctuates within a small range. Due to the interpolation and quantization operations in the resizing and compression, the pixel relationship in finite neighborhood changes, making a relatively obvious variation. The FID method achieves an AUC score of about 80% and more than 85% for tackling the compression and resizing attacks, respectively. Besides, the AUC score of the FID method is more than 95% for tackling the blur and noise attacks under different intensities.

Similarly, the proposed FID method is evaluated on other GANs-synthesized image datasets, which contain rich image types, and the results are compared with AutoGAN; the training datasets and the test datasets were divided in 5 to 1.

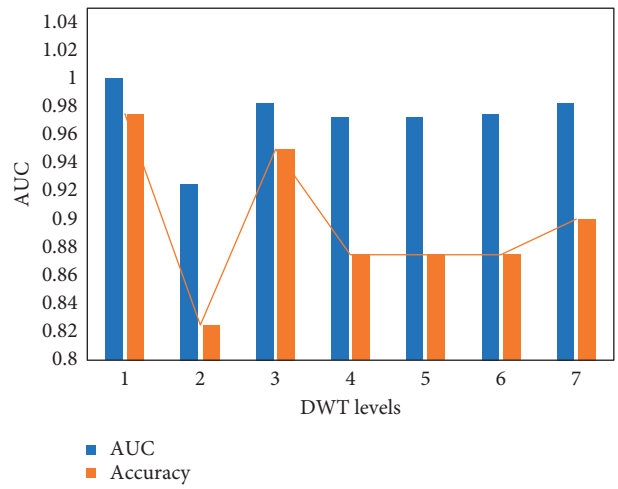

Figure 8: AUC score under multilevel DWT.

The AP score is also taken as a metric for performance evaluation, and the experimental results are given in Table 3. It can be seen that the FID method always maintains a good performance for different types of images synthesized with SOTA GANs. Because the pretrained model was trained on CycleGAN and StarGAN, AutoGAN obtained 100% AP on CycleGAN and StarGAN. DiscoGAN and CycleGAN have a similar architecture, so AutoGAN also achieved a good performance on DiscoGAN. While on other GANs, except BigGAN, FID has achieved better performance compared to AutoGAN. The performance of the FID method in detecting images synthesized by BigGAN, PGGAN, and StyleGAN are not as high as other types of fake images. The reason for the inferior performance could be that the fake images synthesized by BigGAN and PGGAN involve more image types and more complicated image content; thus, the feature vector for classification is more scattered in the hyperplane. FID got a relatively low AP on StyleGAN, because StyleGAN-synthesized image has high quality and contains three types, more difficult to detect. Although GauGAN also contains a variety of images, the quality of the images is not good, and AP arrives at 91.22%. The AP of detecting other types of fake images is also above 90%. According to the experimental results, the detection of fake images with complex types is still challenging.

5.4. Discussion. The proposed FID method achieves impressive effectiveness in detecting SOTA GANs-synthesized images. Also, the method exhibits satisfactory robustness against the four common perturbation attacks. Since the compression attack changes the pixel relationship in the finite neighborhood and affects the spectral correlation of color images, the performance degradation of the FID method under compression attack is relatively obvious.

However, the FID method also has some limitations. For example, the performance of detecting fake images of multiple types is inferior than that of a single type. The content in fake images of multiple types is quite different, making the distribution of the extracted feature vectors in

Table 2: Performance of FakeSpotter, AutoGAN, and FID.

	Precision	Recall	F1	Accuracy	AP	AUC	FPR	FNR
FID	0.9845	0.9845	0.9845	0.9845	0.9889	0.9901	0.021	0.015
FakeSpotter	0.912	0.924	0.918	0.919	0.881	0.919	0.076	0.087
AutoGAN	0.757	0.663	0.707	0.725	0.67	0.725	0.033	0.213

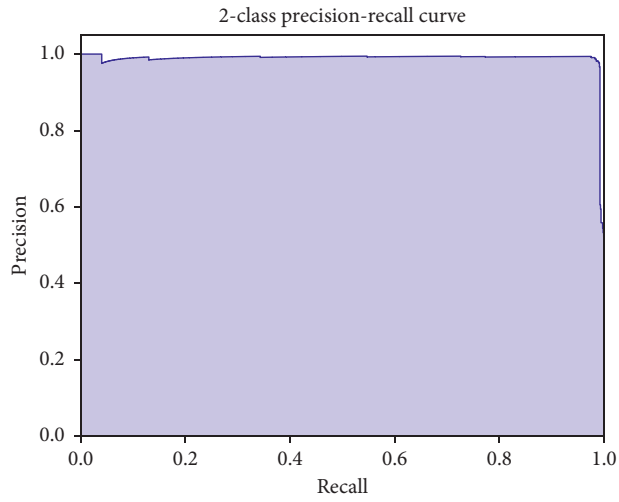

Figure 9: Precision-recall curves of StyleGAN2-synthesized faces.

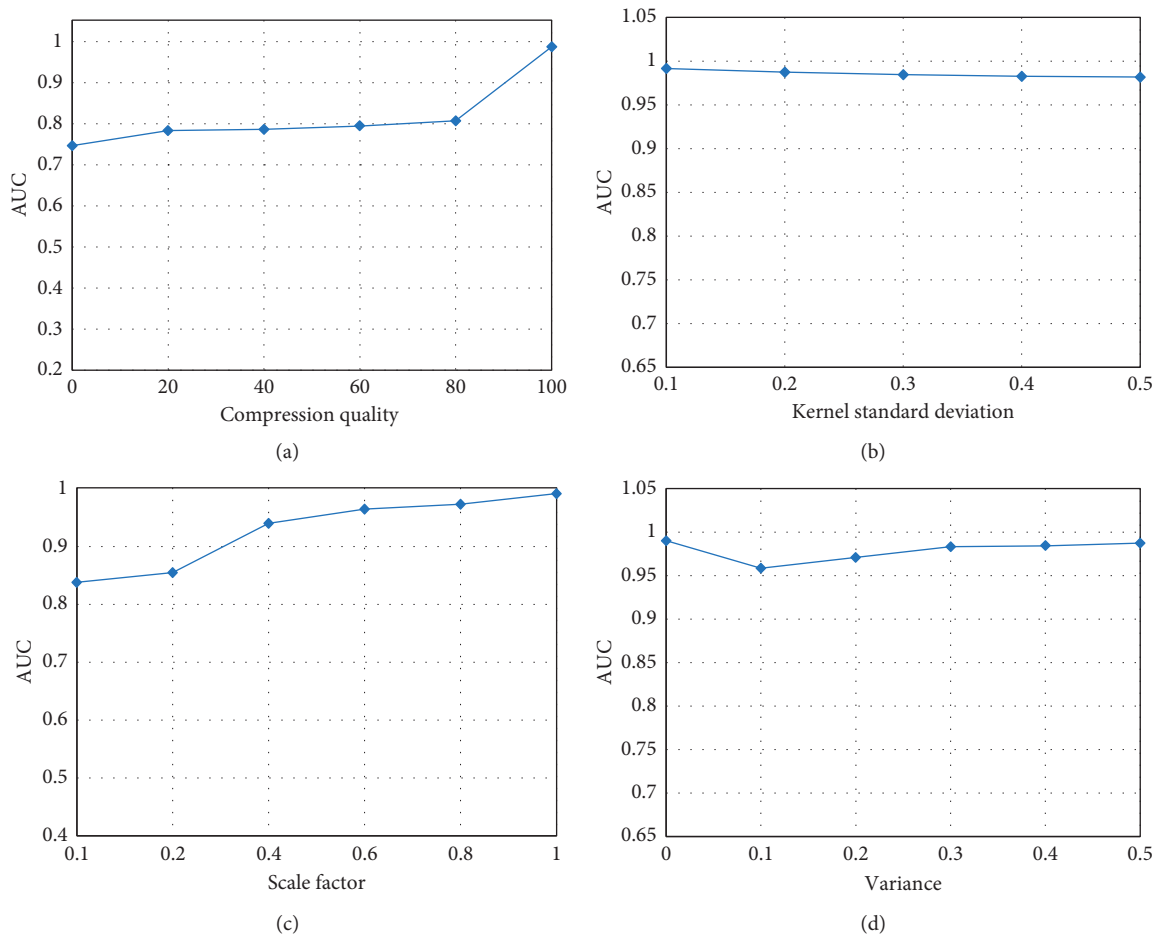

Figure 10: Four perturbation attacks under different intensities. (a) Compression. (b) Blur. (c) Resizing. (d) Noise.

Table 3: Performance of AutoGAN and FID.

	StyleGAN	PGGAN	BigGAN	CycleGAN	StarGAN	GauGAN	PixelRNN	DiscoGAN
FID	85.33	76.32	75.10	96.19	99.71	91.22	90.33	95.23
AutoGAN	68.60	75.60	84.90	100.00	100.00	61.00	71.01	95.10

the hyperplane more scattered. This brings challenges to the classification and inevitably leads to a declined detection effect. The detection of multitype fake images may be a trend in the future, which poses a challenge and calls for effective approaches.

6. Conclusion and Future Research Directions

The rapid development of AI technology makes it possible to produce fake content (e.g., fake audio, fake video, and fake image) that can deceive humans, posing potential challenges to the society and people. This study proposes a method for detecting GAN-synthesized fake images based on DWT and the standard correlation coefficient. Also, the RGB correlation introduced in the imaging process of natural color images are studied. Besides, an extensive evaluation of the FID method on detecting fake images synthesized by StyleGAN2 and several typical SOTA fake images is performed. Experimental results show that the proposed method achieves effectiveness in detecting GAN-synthesized fake images and exhibits robustness against common perturbation attacks. Furthermore, the analysis on the difference between real and fake images in the image imaging process could be extended to other AI-synthesized images.

The research on forgery and fake detection is fundamental, and it is necessary to establish a powerful defense mechanism to avoid AI risks. Currently, the face swap is common with DeepFake, and application of the FID method to DeepFake could be our future work.

Data Availability

The related images used to support the findings of the study are at https://github.com/NVlabs/stylegan2 and https://github.com/peterwang512/CNNDetection. The source codes will be uploaded to GitHub and are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2016YFB0501900).

References

[1] J. P.-A. Goodfellow and M. Mirza, "Generative adversarial nets," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, June 2014.

[2] S. McCloskey and M. Albright, "Detecting GAN-generated imagery using saturation cues," in *Proceedings of the 26th IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, China, September 2019.

[3] L. Nataraj, "Detecting GAN generated fake images using Co-occurrence matrices," *Journal of Electronic Imaging*, vol. 5, pp. 1–7, 2019.

[4] X. Yang, Y. Li, H. Qi et al., "Exposing GAN-synthesized faces using landmark locations," in *Proceedings of the 7th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, Paris, France, July 2019.

[5] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019.

[6] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *Proceedings of the 11th IEEE International Workshop on Information Forensics and Security (WIFS)*, Delft, The Netherlands, December 2019.

[7] R. Wang, L. Ma, F. Juefei-Xu et al., "Fakespotter: a simple baseline for spotting ai-synthesized fake faces," in *Proceedings of the 29th International Joint Conference on Artifical Intelligence (IJCAI)*, Yokohama, Japan, July 2020.

[8] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[9] T. Karras, T. Aila, S. Laine et al., "Progressive Growing of GANs for improved quality, stability, and variation," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, May 2018.

[10] T. Karras, S. Laine, T. Aila et al., "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Angeles, USA, June 2019.

[11] T. KarrasS. Laine et al., "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[12] J.-Y. Zhu, T. Park, P. Isola et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[13] M. Liu, Y. Ding, M. Xia et al., "STGAN: a unified selective transfer network for arbitrary image attribute editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Angeles, USA, June 2019.

[14] Y. Choi, M. Choi, M. Kim et al., "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 2018.

[15] S. Agarwal, H. Farid, Y. Gu et al., "Protecting world leaders against deep fakes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Los Angeles, USA, June 2019.

[16] S. Cole, "We are truly F—ed: everyone is making AI-generated fake porn now," 2018, https://www.vice.com/en us/article/bjye8a/reddit-fake-porn-app-daisy-ridley/.

[17] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 758–767, 2005.

[18] H. Wang and H. Wang, "Perceptual hashing-based image copy-move forgery detection," *Security and Communication Networks*, vol. 2018, pp. 1–11, Article ID 6853696, 2018.

[19] E. Gürbüz, G. Ulutas, and M. Ulutas, "Detection of free-form copy-move forgery on digital images," *Security and Communication Networks*, vol. 2019, pp. 1–14, Article ID 8124521, 2019.

[20] H. Dang, F. Liu, J. Stehouwer et al., "On the detection of digit manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[21] S.-Y. Wang, O. Wang, R. Zhang et al., "CNN-Generated images are surprisingly easy to spot for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[22] F. Marra, D. Gragnaniello, L. Verdoliva et al., "Do GANs leave artificial fingerprints?" in *Proceedings of the 2th IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, San Jose, USA, March 2019.

[23] H. Mo, B. Chen, and W. Luo, "Fake faces identification via convolutional neural network," in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, New York, USA, June 2018.

[24] T. Do Nhu, N. In Seop, H.-J. Yang et al., "Forensics face detection from GANs using convolutional neural network," in *Proceedings of the International Symposium on Information Technology Convergence (ISITC)*, Chonbuk National University, South Korea, October 2018.

[25] C.-C. Chang and C.-J. Lin, "Libsvm," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[26] Z. Liu, P. Luo, X. Wang et al., "Deep Learning face attributes in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, June 2015.

[27] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, USA, April 2019.

[28] T. Park, M. Y. Liu, T. C. Wang et al., "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Angeles, USA, June 2019.