Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2020.09839
Review

A Survey of GAN-Generated Fake Faces Detection Method Based on Deep Learning

Xin Liu* and Xiao Chen

Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Xin Liu. Email: lx@nuist.edu.cn
Received: 21 May 2020; Accepted: 28 June 2020

Abstract: In recent years, with the rapid growth of generative adversarial networks (GANs), a photo-realistic face can be easily generated from a random vector. Moreover, the faces generated by advanced GANs are very realistic. It is reasonable to acknowledge that even a well-trained viewer has difficulties to distinguish artificial from real faces. Therefore, detecting the face generated by GANs is a necessary work. This paper mainly introduces some methods to detect GAN-generated fake faces, and analyzes the advantages and disadvantages of these models based on the network structure and evaluation indexes, and the results obtained in the respective data sets. On this basis, the challenges faced in this field and future research directions are discussed.

Keywords: Generative adversarial networks; fake faces detection; deep learning

1 Introduction

With the rapid development of digital images, people can not only perform dermabrasion, manipulation and other operations on existing faces, but also generate a completely fake face. In particular, some generative adversarial networks (GANs), such as PGGAN [1] and StyleGAN [2] can generate very realistic faces. The fake faces are difficult to distinguish in a short period of time, as shown in Fig. 1. Nowadays, the media network is so developed and human faces are widely used in biometric-based face recognition and authentication. Once fake faces and videos are widely spread on the Internet, some moral problems, social problems and security problems (e.g., fake news, fraud) may arise. Therefore, the authenticity of images is increasingly valued by people.

The traditional fake faces were mainly caused by tampering on the original image. Any images have been tampered with leave traces, leading to changes in statistical data. So, some researchers detect the fake faces based on these statistics. For example, the paper [3] detected tampered faces through Binarized Statistical Image Feature (BSIF), Local Binary Patterns (LBP), Histogram of Gradients (HOG) and Scale Invariant Feature Transform (SIFT). The accuracy of this method exceeds 83.8%. Next, the deep learning is applied to detect the falsified image. For example, using the classic AlexNet convolutional neural network to learn features from the training data to detect tampering images. The paper [4] proposed a constrained convolution layer, which could better learn features and improve the accuracy. MesoNet [5] proposed that the features are extracted through the original convolutional layer not enough to classify the images. So, the author replaced the first two layers of ordinary convolution with the improved Inception module [6] to extract more features. Xception [7] was a further improvement based on Inception-v3, which proposed a deep separable convolution. There is a good detection result. The paper [8] proposed a method for detecting fake faces in video using RNN. These methods can detect tampering images with a high accuracy. But, in recent years, with the rapid development of GANs, many fake faces are generated by GANs. Since GANs generate faces from a random vector, the generated faces do not carry any tampering information. As a result, the methods above have failed to detect the fake faces. Therefore,

researchers have begun to use the deep learning to learn the main differences between the real faces and the fake faces directly. This paper wants to review and summarize these deep learning-based methods.

In order to better understand the following detection methods, the principle and development of GANs will be explained in Section 2. Then, some methods for detecting fake faces generated by GANs will be described in Section 3. Section 4 describes the possible challenges and possible future research directions. Section 5 concludes the paper.

Figure 1: Faces generated by PGGAN

2 Principle and Development of GANs

The generative adversarial networks (GANs) was first proposed in 2014 by Goodfellow [9]. The basic model is shown in Fig. 2. One simple idea is using two models to fight each other. One of model is generator G and the other is discriminator D. Then, it needs to input a random noise z obeying the prior probability to the G, and then the G outputs the data as G(z). Finally, the G(z) and the real data P(data x) are added to the discriminator. After the training, the D network judges whether the input data is real data or generated by the generator. The D improves its discriminating ability through continuous learning, and the G makes the data generated by itself more realistic through continuous learning. So as to deceive the discriminator, G and D fight each other, making their ability stronger and stronger, and finally form a relatively stable state. At this point, the discriminator can't recognize the data generated by the generator, and achieve the purpose of falsification.

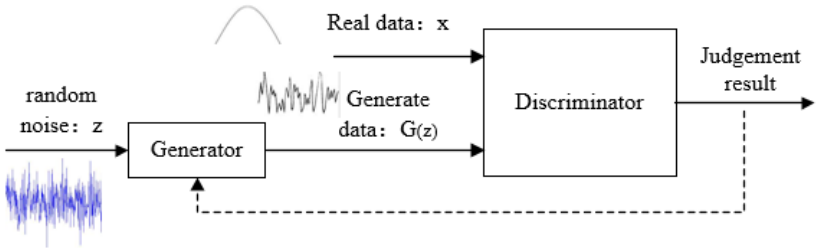

Figure 2: Basic model of GANs

The initial GANs has some flaws. Firstly, the training is unstable. It is difficult to guarantee the synergy between G and D. Secondly, the fake images lack diversity. Finally, it does not have a uniform effective criterion for the quality of the generated image. Therefore, several improvements to GANs have been proposed recently. For example, Radford et al. [10] proposed DCGAN (deep convolutional GAN). DCGAN mainly improves the original GANs in the network structure. It replaces generator and discriminator with two convolutional neural networks, which improved the stability of the network, but it did not solve the problem fundamentally. Mao et al. [11] proposed LSGAN, which mainly changed the objective function of GANs from cross entropy loss to least squares loss. Arjovsky proposed WGAN (Wasserstein GAN) [12], which replaced the JS distance with Wasserstein distance (EM distance). WGAN can fundamentally solve the problem of instability. Gulrajani et al. [13] proposed that WGAN-GP

which can improve WGAN. EBGAN [14] and BEGAN [15] can also generate very realistic face pictures. TeroKarra et al. [1] proposed PGGAN (Pro-GAN), PGGAN and StyleGAN [2] used a step-by-step method to generate high-quality, high-resolution images. Other GANs such as CycleGAN [16], StarGAN [17] can perform image-to-image translation and style altering.

3 Deep Learning-Based Methods for Detecting GAN-Generated Fake Faces

Due to most discriminators and generators used in GANs are mainly based on CNN, it is reasonable to use CNN-based methods to detect the fake faces.

3.1 Methods for PGGAN-Generated Fake Faces

PGGAN can generate the fake faces with almost the highest quality and resolution among the existing GAN models. The generated fake faces are very difficult to identify for people. Therefore, many researchers are committed to design the detection method for PGGAN-generated fake faces.

Nhu et al. [18] employed VGG-Net structure for detection. The net-work is shown in Fig. 3. This structure consists of five modules. Each module has a convolutional layer and a max-pooling layer. They are used for feature extraction. Then, the feature maps input to fully-connected layer. Finally, the softmax layer is used to output the possibility of being a true image.

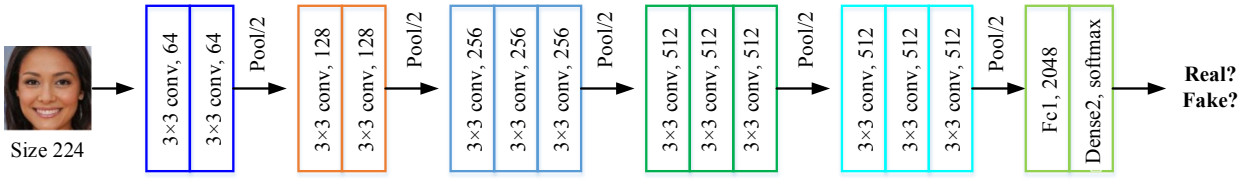

Figure 3: Network structure proposed by Nhu et al.

Mo et al. [19] proposed that the main differences between the real faces and the fake faces were reflected in the residual field. So, firstly, they images are processed with high-pass filter. The resulting residuals are input to three-layer groups. Each group includes a convolutional layer and a max-pooling layer. The out-feature maps of the last group are aggregated and input to two fully-connected. Finally, the softmax layer is used to output the possibility of being a true image.

The results of the above two detection methods are shown in Tab. 1. The performance of the second method is better obviously.

Table 1: Test results of two methods

Method	Training data set	Testing data set	Accuracy
Nhu et al.	CelebA + PGGAN	PGGAN	80%
Mo et al.	CelebA + PGGAN	PGGAN	96.3%

3.2 General Methods for GAN-Generated Fake Faces

In most cases, people don't know the source of fake images from which GAN model. So, it is desirable to design a general method for detecting fake faces generated by most of GAN models.

Hsu et al. [20] proposed the deep forgery discriminator (Deep FD) structure, as shown in Fig. 4. It can detect images generated by various GANs, and is no longer limited to images generated by a

particular GANs. The paired data is input into the feature extraction network. The network updates the parameters of the network through the contrast-loss [21]. In this way, the extracted features will be more perfect. In the feature extraction network, joint discriminative feature learning method is employed (Jointly Discriminative Feature Learning). After the features are extracted, these feature maps are fed into the discriminator to detect the authenticity of the image.

During training, the author used CelebA as the data set of the real face and used the false face generated by DCGAN, LSGAN, WGAN, WGAN-GP and PGGAN as the fake face data set. The test results were shown in Tab. 2.

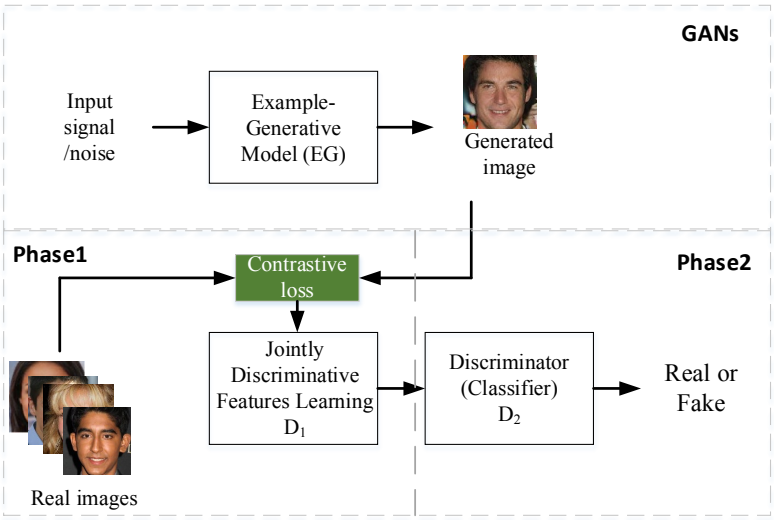

Figure 4: Framework proposed in [20]

Zhuang et al. [22] proposed a coupled network with two-step pairwise learning. As shown in Fig. 5, the idea is similar to the idea of the above article. The author of this article believes that the contrast-loss is not stable, so the author used triplet-loss [23] to update the parameters of the network. The author also proposed that since the fake faces generated by different GANs may have different characteristics, the original one-stream CNN cannot extract all the features. To solve this problem, the paper proposed to use the coupled deep neural network (CDNN). The network includes a 3 × 3 convolution kernel and a 5 × 5 size convolution kernel. The 3 × 3 convolution kernel is used to extract local false face features, and the 5 × 5 size convolution kernel is used to capture the global fake face features. Finally, these feature maps are input into a classifier that includes two fully connected layers to obtain the final prediction result.

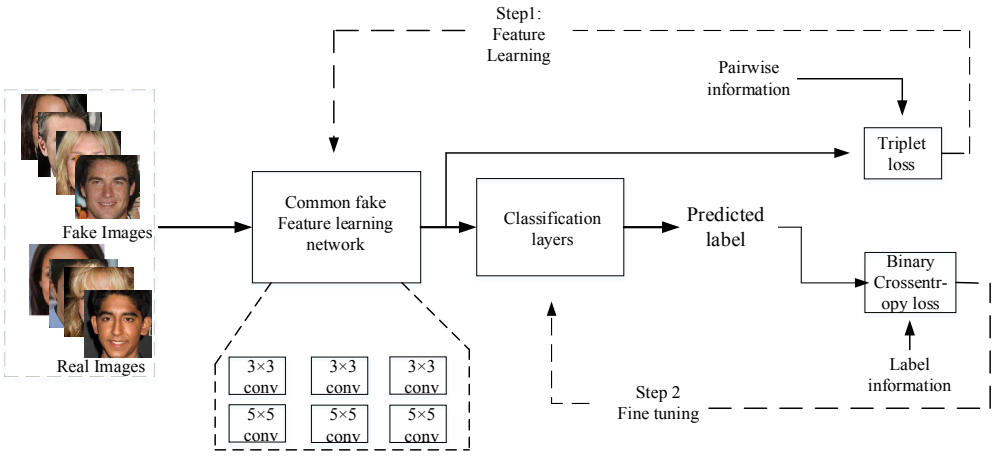

Figure 5: Framework proposed in [22]

The results of the above two methods are shown in Tab. 2. Precision is the proportion of the correct sample predicted in the set of samples that are predicted to be positive. Recall is the proportion of the correct sample predicted in the actual positive sample. It can be seen from the results that the two detection methods not only can detect images generated by different GANs but also have better performance.

Table 2: The test results of the above two methods were tested on LSGAN, DCGAN, WGAN, WGAN-GP, PGGAN respectively

Method	Hsu et al.		Zhuang et al.	
	Precision	Recall	Precision	Recall
LSGAN	0.947	0.922	0.981	0.956
DCGAN	0.871	0.844	0.986	0.986
WGAN	0.838	0.847	0.895	0.881
WGAN-GP	0.818	0.835	0876	0.881
PGGAN	0.926	0.918	0.951	0.936

By studying the structure of GANs, References [24–25] found some invisible differences between real faces and GAN-generated fake faces. The differences of the color spaces were analyzed, namely, RGB, HSV and YCbCr. The chi-square distance was used to evaluate the difference between the image statistics generated by the GANs and the real image statistics. The larger the chi-square distance, the more obvious the difference. So, researchers began try to extract features from the color spaces.

In [26], the features were extracted from the color spaces and then input into the network for detection. The framework is shown in Fig. 6. This structure used the co-occurrence matrix [27] widely used in image texture analysis as feature descriptors. For a given image, the features of the color components were first calculated and then connected into a feature vector. because high-frequency filtering can be better captured by high-pass filtering. So, during the feature extraction process, the image was processed by high-pass filtering. Finally, a classifier was trained to detect whether the input image is real or generated by the GANs. In the selection of the training data set, the real face selected CelebA and the fake face selected the fake face generated by DCGAN, W-GAN, and PGGAN. After training, the final test results were ideal, and higher accuracy can be obtained, which can reach more than 95%.

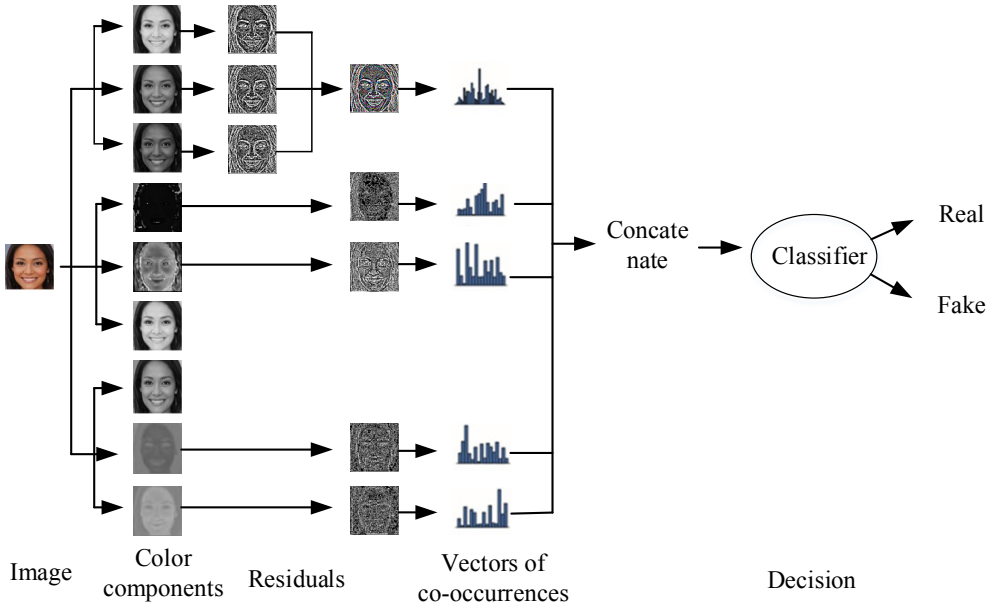

Figure 6: Framework proposed in [26]

The above methods do not consider the robustness of the model. In real life, most of the pictures transmitted on the network are post-processed pictures (such as JPEG compression, Gaussian noise, etc.). This increases the difficulty of detection.

He et al. [28] proposed that common post-processing attacks will make the abnormal traces in the RGB space unreliable, while the statistical characteristics of chrominance information in other color spaces may be more distinguishable and robust. Therefore, the author exploited a well-designed shallow CNN to extract the features of the chrominance component, and then Random Forest (RF) [29] for classification. The flowchart of this method is shown in Fig. 7. Finally, the author did experiments on six attack methods such as JPEG compression, Gaussian noise, and bilateral filtering, and achieved very good results, proving that the model has good robustness.

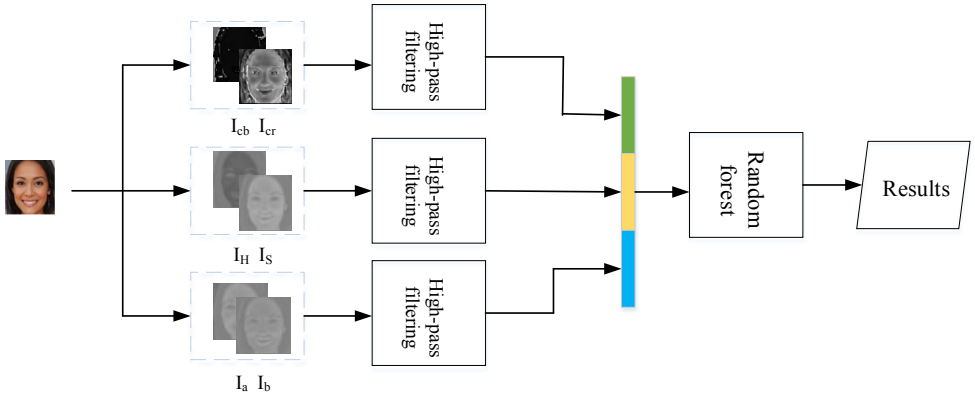

Figure 7: The flowchart of [28]

Through experimental and theoretical analysis, Liu et al. [30] got the conclusion that CNN mainly extracts regions with rich texture (such as skin) when performing fake face detection. Then, through further experiments they came to this conclusion that the global texture can effectively improve the robustness of the result. The main architecture of this method is shown in Fig. 8. The author uses the gram matrix to extract the global image texture feature and then detects it. Finally, the author tested the robustness and generalization of the model, and obtained very good results.

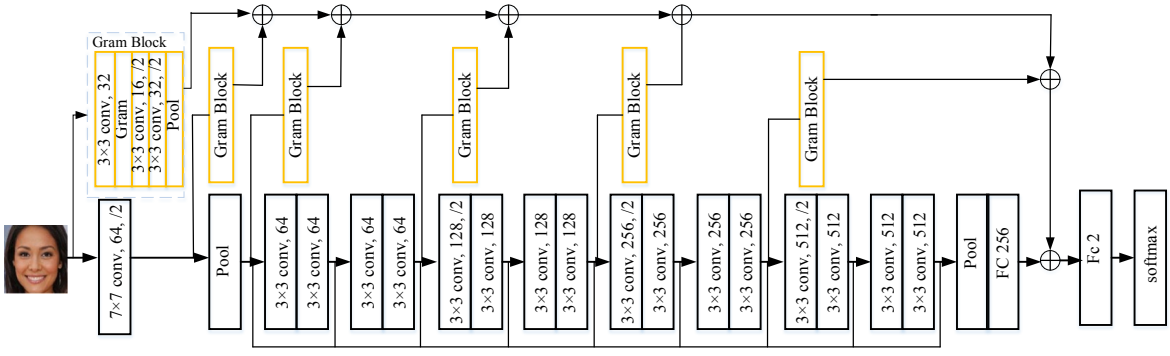

Figure 8: The flowchart of [30]

4 Challenges and Future Research Directions

The main challenges in the field of detecting fake face images generated by GANs and the problems to be solved are as follows:

(i) Generalization of detectors: The development of GANs is rapid and new GANs may appear in the future. Detectors should detect images generated by new GANs. So, researchers can consider improving the generalization of detectors.

(ii) Robustness of detectors: The fake faces may be compressed on online me-dia. So, researchers can consider improving the robustness against compression for the detectors.

(iii) Mobile device detection: Due to the large amount of computation, existing detectors are not suitable for mobile applications. Researchers can consider improving the computational complexity of the detectors.

(iv) Large data set: Currently, there are few large data sets available to the public.

5 Conclusion

This paper mainly introduces several detection methods for detecting fake faces generated by GANs. From the results, it can be concluded these detection methods have high accuracy. Although the GANs are developing faster and faster. Various GANs may emerge in the future. They may generate higher quality fake faces. But we can get inspiration from the above methods. The GANs cannot completely describe many intrinsic properties of real images, such as differences in color spaces. So, trying to find the differences between fake images and true images. Then, according to the differences develop more advanced detectors.

Funding Statement: This work is supported by National Natural Science Foundation of China (62072251).

Conflicts of Interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. Int. Conf. on Learning Representations*, Vancouver, BC, Canada, 2018.

[2] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 4401–4410, 2019.

[3] U. Scherhag, C. Rathgeb and C. Busch, "Towards detection of morphed face images in electronic travel documents," in *Proc. 2018 13th IAPR Int. Workshop on Document Analysis Systems*, Vienna, Austria, pp. 187–192, 2018.

[4] B. Bayar, Belhassen and C. S. Matthew, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, 2018.

[5] D. Afchar, V. Nozick and J. Yamagishi and E. Isao, "Mesonet: A compact facial video forgery detection network," in *Proc. 2018 IEEE Int. Workshop on Information Forensics and Security*, Hong Kong, CN, pp. 1–7, 2018.

[6] C. Szegedy, W. Liu, Y. Jia and S. Pierre, "Going deeper with convolutions," in *Proc. Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 1–9, 2015.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1251–1258, 2017.

[8] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 2018 15th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Auckland, New Zealand, pp. 1–6, 2018.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza and X. Bing, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems,* pp. 2672–2680, 2014.

[10] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[11] A. Brock, J. Donahue and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. on Learning Representations*, Vancouver, BC, Canada, 2018.

[12] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. on Machine Learning*, vol. 70, 2017.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, "Improved training of wasserstein

gans," in *Proc. Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

[14] Z. Junbo, M. Michael and Y. LeCun, "Energy-based generative adversarial networks," in *Proc. 5th Int. Conf. on Learning Representations*, Palais des Congrès Neptune, Toulon, Fr, 2017.

[15] D. Berthelot, T. Schumm and L. Metz, "Began: Boundary equilibrium generative adversarial networks," arXiv preprint arXiv:1703.10717, 2017.

[16] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2223–2232, 2017.

[17] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim *et al.,* "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake, UT, USA, pp. 8789–8797, 2018.

[18] T. D. Nhu, I. S. Na and S. H. Kim, "Forensics face detection from GANs using convolutional neural network," in *Proc. Int. Sym. on Information Technology Convergence*, 2018.

[19] H. Mo, B. Chen, W. Luo, "Fake faces identification via convolutional neural network," in *Proc. 6th ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck, Austria, pp. 43–47, 2018.

[20] C. C. Hsu, C. Y. Lee and Y. X. Zhuang, "Learning to detect fake face images in the wild," in *Proc. Int. Sym. on Computer, Consumer and Control*, Taiwan, pp. 388–391, 2018.

[21] E. Simo-Serra, E. Trulls, L. Ferraz and L. Kokkinos, "Discriminative learning of deep convolutional feature point descriptors," *in Proc. IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 118–126, 2015.

[22] Y. X. Zhuang and C. C. Hsu, "Detecting generated image based on a coupled network with two-step pairwise learning," in *Proc. IEEE Int. Conf. Image Processing*, Taiwan, pp. 3212–3216, 2019.

[23] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop on Similarity-Based Pattern Recognition*, Springer, Cham, pp. 84–92, 2015.

[24] H. Li, B. Li and S. Tan, "Detection of deep network generated images using disparities in color components," arXiv preprint arXiv:1808.07276, 2018.

[25] J. F. Lalonde and A. A. Efros, "Using color compatibility for assessing image realism," in *Proc. 2007 IEEE 11th Inte. Conf. on Computer Vision*, Venice, Italy, pp. 1–8, 2017.

[26] H. Li, B. Li, S. Tan and J. Huang, "Detection of deep network generated images using disparities in color components," arXiv preprint arXiv:1808.07276, 2018.

[27] R. M. Haralick, K. Shanmugam and Ih. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, pp. 610–621, 1973.

[28] P. He, H. Li and H. Wang, "Detection of fake images via the ensemble of deep representations from multi color spaces," in *Proc. IEEE Int. Conf. on Image Processing*, Taiwan, pp. 2299–2303, 2019.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] Z. Liu, X. Qi and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 8060–8069, 2020.