Journal of Information Hiding and Privacy Protection
DOI: 10.32604/jihpp.2022.031707
Article

Tech Science Press

A Survey of Anti-forensic for Face Image Forgery

Haitao Zhang*

Engineering Research Center of Digital Forensics of Ministry of Education, School of Computer,
Nanjing University of Information Science & Technology, Nanjing, 210044, China
*Corresponding Author: Haitao Zhang. Email: yzzhanghaitao@163.com

Received: 25 April 2022; Accepted: 28 May 2022

Abstract: Deep learning related technologies, especially generative adversarial network, are widely used in the fields of face image tampering and forgery. Forensics researchers have proposed a variety of passive forensic and related anti-forensic methods for image tampering and forgery, especially face images, but there is still a lack of overview of anti-forensic methods at this stage. Therefore, this paper will systematically discuss the anti-forensic methods for face image tampering and forgery. Firstly, this paper expounds the relevant background, including the relevant tampering and forgery methods and forensic schemes of face images. The former mainly includes four aspects: conventional processing, fake face generation, face editing and face swapping; The latter is mainly the relevant forensic means based on spatial domain and frequency domain using deep learning technology. Then, this paper divides the existing anti-forensic works into three categories according to their method characteristics, namely hiding operation traces, forgery reconstruction and adversarial attack. Finally, this paper summarizes the limitations and prospects of the existing anti-forensic technologies.

Keywords: Anti-forensics; face tempering and forgery; forensics; generative adversarial network

1 Introduction

Face images are widely used in daily life and work. However, with the rapid development of image processing technologies and the success of deep learning related technologies, people can not only edit and tamper with the existing face images, but also swap faces and reenact faces, and even generate fake faces that do not exist at all. The boundary between these forged and tampered face images and real face images becomes blurred, which is difficult to distinguish by human eyes. In addition, the software package provided free of charge on the Internet allows any individual to easily process or generate very realistic fake face images even without special skills. On the one hand, it opens the door to a series of amazing applications in different fields such as advertising and film production. On the other hand, in today's world, once fake face images are maliciously used, they will pose a huge security threat and seriously affect social stability. For example, using fake face images to register accounts or swap faces will lead to privacy or security problems.

The challenges posed by fake face images have promoted the development of forensics technology [1,2], and researchers have proposed a large number of forensic solutions to address the above problems. The source of the tampered images is traced through active forensics, and the authenticity of the image is checked through passive forensics. Furthermore, an initial defense [3] is proposed to prevent forgers from successfully tampering with real faces. At present, passive forensics schemes have achieved very impressive performance in detecting image authenticity and are trusted by researchers. However, these forensic solutions may still have some kind of loopholes, and these loopholes may be utilized purposefully to make the tampered faces escape detection.

The research on anti-forensics [4,5] has emerged to reveal the vulnerabilities of existing forensic detectors. The so-called anti-forensics means that the tampered and forged image cannot be detected by image forensics technology after some processing, which reduces the performance of the detectors. Studying anti-forensics techniques in turn motivates researchers to come up with more reliable and robust forensic schemes, which are crucial in security-related forensics tasks. At present, there is still a lack of relevant research reviews on anti-forensic technologies. This article will briefly introduce relevant face tempering and forgery technology, forensics methods and systematically discuss the existing anti-forensic methods of images, especially face images.

2 Face Forgery and its Forensics

In this section, the relevant background of anti-forensic technology, including the main means of forgery and forensics schemes are discussed based on face images.

2.1 Face Forgery

In recent years, with the great success of deep learning technology, especially convolutional neural network (CNN) and generative adversarial network (GAN) [6], fake faces have been more and more difficult to be identified by human eyes. In this paper, according to the different ways of face tampering or generation, as shown in Fig. 1, face forgery technologies can be divided into four categories: conventional processing, face generation, face editing and face swapping.

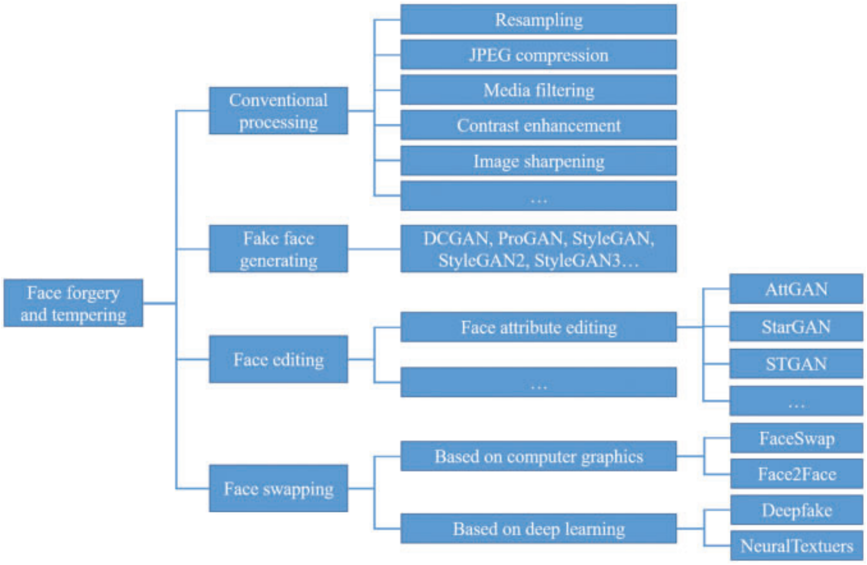

Figure 1: The technology categories of face forgery and tempering

Conventional processing. It mainly refers to the conventional image processing operations, mainly including resampling, JPEG compression, media filtering, contrast enhancement and image sharpening. These operations are very common and can be easily realized by image processing software such as Photoshop, GIMP and ACD See in real life.

Fake face generating. This forgery method mainly depends on the development of GAN technology. Through the adversarial training between the generator and the discriminator, the fake face images generated by the generator are closer and closer to the natural images in the process of weight alternating optimization.

The original GAN can only generate low resolution pictures, and the effect is relatively poor. Then, researchers carry out a series of research on the quality of face image generation. Radford et al. [7] combine convolution structure with GAN called DCGAN, which makes GAN easier to deal with many problems in the image field. Karras et al. [8] propose ProGAN based on step-by-step amplification training method to solve the problem of model collapse when generating high-resolution pictures. Then, inspired by style transfer, they decouple the input vectors to realize style control and propose StyleGAN [9], so that the generator can generate high quality and diverse face images with specified features. For the water droplet artifacts of faces generated by StyleGAN, they also propose StyleGAN2 [10] to further improve the quality of the generated face. Lately, Karras et al. [11] improve the generator network structure of StyleGAN2 to make it have high-quality equivariance. Fig. 2 shows some examples of the above GAN generated face and it can be found that it is difficult for human eyes to detect the difference between them and natural faces in addition to DCGAN.

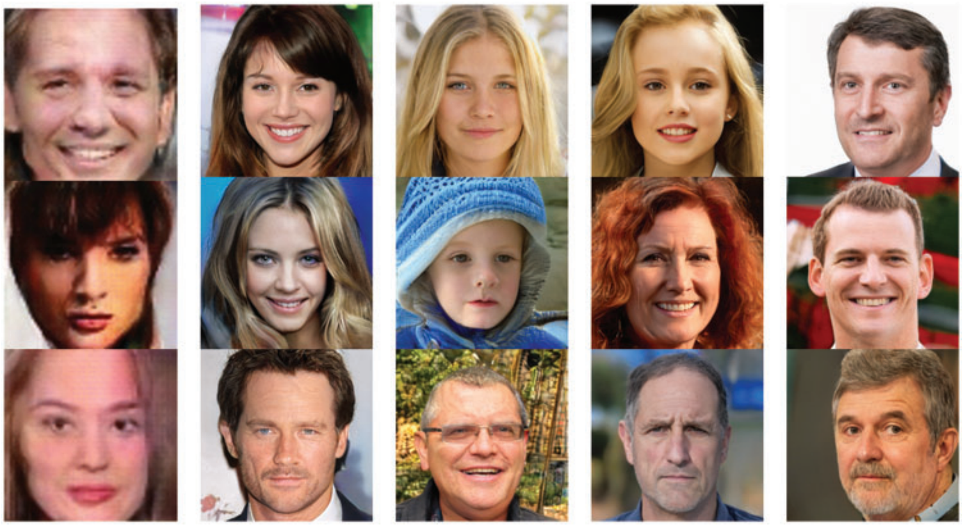

Figure 2: Some examples of GAN generated faces, including DCGAN, ProGAN, StyleGAN, StyleGAN2, StyleGAN3 (left to right columns)

Face editing. Editing is often based on face generation, which can provide free attribute transformation of face, such as changing hair color, face skin color and so on. He et al. [12] propose a GAN structure called AttGAN, which applies attribute classification constraints to the generated image to ensure the correct changes of the required attributes. Choi et al. [13] use only one generative network called StarGAN to deal with the problem of generating images between multiple domains, and realizes further face attribute migration. Liu et al. [14] utilize differential attributes to further realize the editing of high-precision attributes of faces and propose STGAN.

Face swapping. This is to generate an image that retains the original face attributes, such as expression, illumination, angle, etc., and possess the appearance of the target face at the same time. There are Face2Face and FaceSwap based on computer graphics method, DeepFakes and NeuralTextuers [15] based on deep learning method.

2.2 Fake Face Forensics

Fake face image forensics aims to identify whether the measured image is real or fake, which can be regarded as a binary classification problem. The detection technology mainly utilizes the distinguishable features of natural faces and fake faces extracted by artificial design or neural network learning and the features are input into the classifier to realize classification.

According to different feature extraction methods, detection technology can be divided into two categories: artificial features based and deep learning based. The former is mainly based on the specific traces left by some forgery methods, such as capturing color component difference [16] or abnormal tooth details [17]. It is lack of universality and is vulnerable to the changes of forgery methods and data distribution. The latter has gradually become the mainstream method because it can deal with more complex forgery methods. According to the difference of feature extraction and perspective, the scheme based on deep learning can be simply divided into: spatial domain-based methods and frequency domain-based methods.

Spatial domain-based methods. Some classical network structures, such as DenseNet [18], ResNet [19] and EfficientNet [20], have been directly applied to fake face detection and achieved good detection results. They directly extract feature information from the spatial domain of the image. In order to further improve the learning ability of neural network, researchers have introduced preprocessing, network module design, attention mechanism and other means. Both Nataraj et al. [21] and Goebel et al. [22] extract the co-occurrence matrix of three channels of images and utilize such features for classification. Guo et al. [23] suppress the interference of image content information and enhance tampering traces through adaptive extraction of high-frequency residuals. Liu et al. [24] introduce the gram-block to add the global texture information to different feature levels to improve the generalization and robust ability. Chen et al. [25] enhance the detection ability through the ensemble of luminance and chrominance component. Zhao et al. [26] utilize multiple attention mechanisms for deepfake detection. In addition, data enhancement strategy [27] can also work well.

Frequency domain-based methods. The forgery traces of fake face images also show abnormal characteristics in the frequency domain. Frank et al. [28] find that different up-sampling processes will lead to abnormal high-frequency components in the frequency domain and propose a detection method based on DCT transform. Similarly, Agarwal et al. [29] utilize color channel spectrum and capsule network. Qian et al. [30] propose F3-Net and carefully design the frequency domain information extraction module to enlarge the artifacts caused by the forgery process in the frequency domain. Chen et al. [31] propose MPSM method and improve frequency domain feature extraction of F3-Net. Liu et al. [32] propose a spatial shallow learning method SPSL based on phase spectrum information extraction.

Recently, the fake face detection models mentioned above have achieved considerable performance on many benchmark datasets, and even the detection accuracy has reached 100%.

3 Anti-forensic Methods

Face forgery anti-forensics, as the name suggests, is the confrontation form of forensic technology. Specifically, through some anti-forensic algorithm, the forged face cannot be detected by the forensics detector, i.e., the methods in Subsection 2.2 is invalid to detect the fake face in Subsection 2.1 after anti-forensic operations. The anti-forensic researches can be used to reveal the vulnerability of existing forensic methods and further promote the optimization and development of forensic algorithms. With the development of GANs and the emergency of adversarial attack, anti-forensic technology has ushered in new development opportunities and is easier to realize.

According to the different needs and means of anti-forensics, anti-forensic algorithms are classified as hiding operation traces, forgery reconstruction and adversarial attack.

3.1 Hiding Operation Traces

This kind of anti-forensic research has a long history and most of such anti-forensic methods are aimed at the operations in the conventional processing mentioned in the Subsection 2.1. Of course, there are also few relevant studies on the generated fake face.

For the JPEG compression, the anti-forensics has attracted extensive attention. The anti-forensics mainly focused on how to hide the quantization trace and block trace in the process of JPEG compression, which is realized by adding median filter and Gaussian white noise [33] or redistributing DCT coefficients by adding noise [34]. For the contrast enhancement, the aim of anti-forensics is to make the image have no spike or gap effect on its histogram. The main methods include adjusting the image histogram by resampling or noise [35] and offsetting the original traces by forging traces [36]. For the media filtering, it mainly attempts to use the optimization method to eliminate the left traces [37] and change the image pixel distribution by adding noise [38] to achieve the anti-forensic effect. For other forgery and tempering methods, Kirchner et al. [39] propose a method to hide image resampling traces. Lu et al. [40] mask the edge overshoot effect and end mutation trace caused by sharpening via adding jitter noise. In addition, for GAN-generated fake faces, Yu et al. [41] analyze the existence of “GAN fingerprints”. Therefore, trying to remove the “GAN fingerprint” [42] left by the generator of GAN can play a certain anti-forensics effect.

3.2 Forgery Reconstruction

Such anti-forensic methods mainly utilize the GAN technology and the study has been largely converted to image conversion problem and achieved good results. Specifically, by reconstructing the image to remove the relevant tampering and forgery features, the effect of fooling the detector can be achieved. This kind of anti-forensic means can be further divided into single forgery processing and multiple forgery processing. Specifically, the former mainly aims at a single way of tampering or forgery, the latter attempts to remove the traces of multiple ways of forgery at the same time and needs to ensure the anti-forensic effect under a single forgery way. Therefore, how to deal with the traces left by various tampering or forgery methods is still a great challenge. The following describes these two types of related articles in detail.

Single forgery processing. First, the related anti-forensic schemes for JPEG compression are introduced. Luo et al. [43] have achieved good results by using discriminator to capture statistical differences between images and uncompressed images. Wu et al. [44] use the loss function based on the high-frequency DCT coefficient to reconstruct the high-frequency component of the image. Wu et al. [45] propose JPEG restoration and anti-forensics GAN (JRA-GAN) and improve the loss function of high-frequency DCT coefficient in work [44].

Then, the anti-forensic algorithms for other conventional operations will be described as follows. Shen et al. [46] realize the similar processing of USM sharpening algorithm through pix2pix [47], so that the generated image has the characteristics of sharpening, but it cannot be simply regarded as a sharpened image, because the traditional sharpening operation is not used in the process. As a result, it can deceive CNN sharpening detector with high accuracy and achieve a good anti forensics effect. Kim et al. [48] reconstruct the image before median filtering, so that the reconstructed image can follow the statistical characteristics of the original image. Xie et al. [49] propose a dual domain anti-forensic GAN network, which discriminates in the forensics feature domain and spatial domain respectively through two discriminators, and achieve good anti-forensic effect.

Finally, for the computer-generated faces, researchers have achieved good anti-forensic effect by increasing the naturalness of the images. Concretely, Nguyen et al. [50] propose a generator called H-Net to transform the distribution feature of generated faces into the latent feature of natural faces. Aiming at the problem of insufficient color of face image in work [50], Peng et al. [51] propose a new GAN structure called CGR-GAN to convert style of natural face into generated faces inspired by style transfer. For the face Swapping, Ding et al. [52] propose a new GAN structure with multiple generators and multiple discriminators to resist DeepFake detection

Multiple forgery processing. The relevant anti-forensic work is relatively few**.** Wu et al. [53] design an anti-forensic algorithm for multiple editing processing by reconstructing images according to WGAN-GP network architecture. Based on this, they [54] extend the training mode and integrated mode, and introduce integrated multiple operation anti-forensic strategies.

3.3 Adversarial Attack

Firstly, the related terms and definitions related to this method are introduced.

Adversarial attack. Szegedy et al. [55] first propose the concept of adversarial example, that is, deliberately adding some well-designed perturbations to the input sample that people can't detect, resulting in the model giving a wrong output with high confidence. The process is called adversarial attack.

White-box setting. The specific information of the model can be obtained when adversarial attacking, including network structure, network parameters and training data.

Black-box setting. The specific information of the model cannot be obtained.

Transferability. It can be understood as generalization. The adversarial example generated by attacking a white-box model also has an attack effect on the unknown model. This characteristic is called the transferability of adversarial examples.

The anti-forensic process realized by adversarial attack is shown in the Fig. 3. By adding adversarial perturbations to the fake face images, the forensics detector fails to detect the fake face.

The common adversarial attack baseline methods applied to anti-forensic tasks are introduced first. Fast gradient sign method (FGSM) [56], the most commonly used baseline anti-forensic attack method, is proposed by Goodfellow. The idea of FGSM algorithm is that an adversarial example is generated by performing a one-step gradient update along the direction of the gradient sign at each pixel. Based on FGSM attack, a series of iterative attack algorithms are derived. In the field of anti-forensics, project gradient descent (PGD) [57] and MIM [58] algorithms are the most representative. The former carries out multi-step perturbation along the increasing direction of the gradient through iteration, and recalculates the gradient direction after each small step. The latter introduces momentum on the basis of iteration.

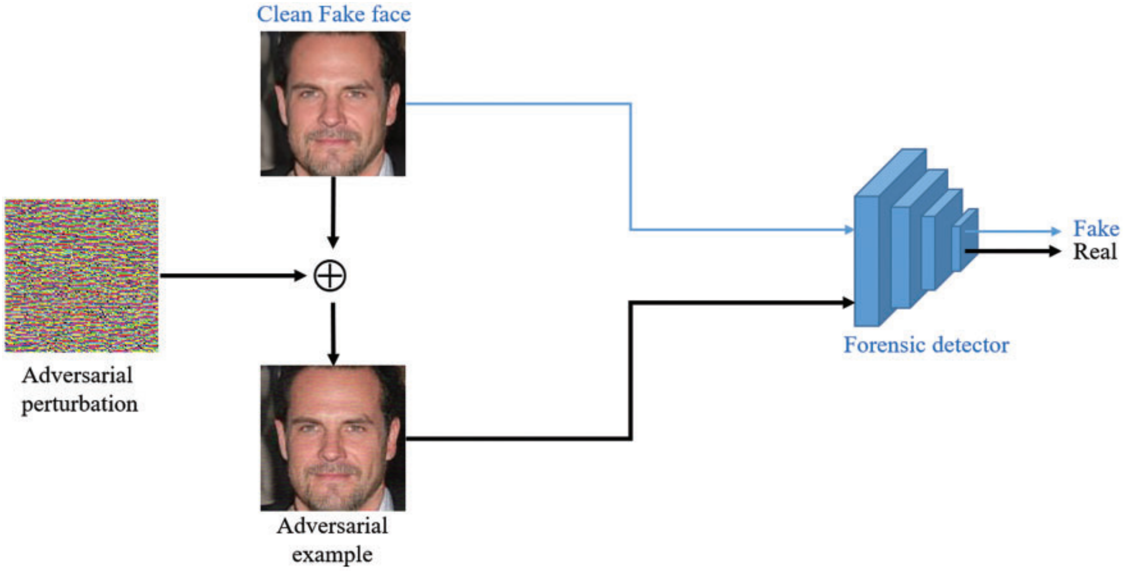

Figure 3: The anti-forensic process realized by adversarial attack

Goebel et al. [59] try to make the co-occurrence matrix of the generated anti-forensic image close to the co-occurrence matrix of the target real image and achieve excellent results for the forensic detector based on co-occurrence matrix but it has poor transferability for other detectors. Wang et al. [60] apply FGSM attack and MIM attack to GAN-generated face images and find that most of the perturbations are concentrated in the Y channel, which seriously affect the visual quality, so they try to add perturbation restrictions to allocate more perturbations Under the Cb and Cr channels. Zhao et al. [61] introduce an ensemble strategy to improve the transferability of anti-forensics and apply the algorithm to GAN-generated face scenes [62]. Li et al. [63] directly start from the source of the StyleGAN-generated facevy attacking the input latent vector and noise via PGD method to generate the anti-forensic faces. Jia et al. [64] propose a frequency domain adversarial attack against Deepfake detection instead of injecting adversarial perturbations into spatial domain and enhance the transferability. Hussain et al. [65] combine with PGD algorithm and some transformations, such as gaussian blur and translation, to realize the attack on DeepFake detectors and make adversarial perturbations robust.

4 Challenges and Prospects

At present, the forensics and anti-forensics related technologies of face tampering and forgery are still in their infancy, and this is a long-term and continuous process. The existing anti-forensic technologies still have some limitations.

The tradeoff between visual quality and anti-forensic effect. Due to the particularity of face image and the demand of anti-forensics, anti-forensic face needs to fool more detectors as many as possible without affecting the visual effect, that is, improve the transferability as much as possible. This balance is still a research focus and difficulty.

Insufficient interpretability. On the one hand, most of the existing advanced passive forensics technologies for face tampering and forgery are based on convolutional neural network, which leads to the lack of interpretability of the forensics detectors themselves. On the other hand, the anti-forensic method based on adversarial attack is more aimed at the loopholes of convolutional neural network, which lacks certain interpretability. In addition, the anti-forensic method based on forgery reconstruction also uses GAN technology based on deep learning, which also introduces new reconstruction traces.

Requirements of real scenes. The existing anti-forensic methods seldom consider various attacks in real scenes. For example, the anti-forensic means based on adversarial attack can easily make the adversarial perturbations that achieves the anti-forensic effect invalid when encountering operations such as blur and denoising. Therefore, the robustness of anti-forensic technology needs to be enhanced.

In the future, both forensic and anti-forensics methods need to make continuous progress to deal with the continuous development of face tampering and forgery technology. In view of the above limitations, the research on anti-forensics should dig more theoretical support and the trade-offs of generalization, robustness and visual effect on the basis of existing work.

5 Conclusion

This paper combs and summarizes the existing image anti-forensic technologies, especially the technologies related to face tampering and forgery. The objects, objectives, methods and limitations of existing anti-forensic schemes are systematically summarized. Aiming at face tampering and forgery, the existing anti-forensic methods reveal the relevant vulnerabilities of forensics detector by hiding operation traces, forgery reconstruction and adversarial attack.

In the future, face tampering and forgery anti-forensic technology will continue to tap the loopholes of existing forensic technology to develop more powerful forensic detectors to deal with the update iteration of tampering and forgery technology. Forensics and anti-forensic technologies have been continuously developed and improved in confrontation.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Wang, H. Guo, S. Hu, M. -C. Chang and S. Lyu, “Gan-generated faces detection: A survey and new perspectives,” arXiv preprint arXiv:2202.07145, 2022.

[2] S. H. Cao, X. H. Liu, X. Q. Mao and Q. Zhou, “A survey on face forgery and forgery-detection technologies,” *Journal of Image and Graphics*, vol. 27, no. 4, pp. 1023–1038, 2022.

[3] Q. Huang, J. Zhang, W. Zhou and N. Yu, “Initiative defense against facial manipulation,” arXiv preprint arXiv:2112.10098, 2021.

[4] W. Wang, F. Zeng, M. Tang, J. J. Chen and H. J. Li, “Survey on anti-forensics techniques of digital image,” *Journal of Image and Graphics*, vol. 21, no. 12, pp. 1563–1573, 2016.

[5] P. S. He, W. C. Li, J. Y. Zhang, H. X. Wang and X. H. Jiang, “Overview of passive forensics and anti-forensics techniques for GAN-generated image,” *Journal of Image and Graphics*, vol. 27, no. 1, pp. 88–110, 2022.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

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

[8] T. Karras, T. Aila, S. Laine and J. Lehtinen, “Progressive growing of gans for improved quality, stability, andvariation,” arXiv preprint arXiv:1710.10196, 2017.

[9] T. Karras, S. Laine and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 4401–4410, 2019.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen *et al.,* “Analyzing and improving the image quality of stylegan,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8110–8119, 2020.

[11] T. Karras, M. Aittala, S. Laine, E. Harkonen, J. Hellsten *et al.,* “Alias-free generative adversarial networks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 1–12, 2021.

[12] Z. He, W. Zuo, M. Kan, S. Shan and X. Chen, “Attgan: Facial attribute editing by only changing what you want,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[13] 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. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8789–8797, 2018.

[14] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding *et al.,* “Stgan: A unified selective transfer network for arbitrary image attribute editing,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3673–3682, 2019.

[15] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies *et al.,* “Faceforensics++: Learning to detect manipulated facial images,” in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 1–11, 2019.

[16] H. Li, B. Li, S. Tan and J. Huang, “Identification of deep network generated images using disparities in color components,” *Signal Processing*, vol. 174, pp. 107616, 2020.

[17] F. Matern, C. Riess and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, pp. 83–92, 2019.

[18] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell *et al.,* “Densenet: Implementing efficient convnet descriptor pyramids,” arXiv preprint arXiv:1404.1869, 2014.

[19] 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 the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8695–8704, 2020.

[20] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Int. Conf. on Machine Learning*, Long Beach, California, USA, pp. 6105–6114, 2019.

[21] L. Nataraj, T. M. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner *et al.,* “Detecting gan generated fake images using co-occurrence matrices,” *Electronic Imaging*, vol. 2019, no. 5, pp. 532–1, 2019.

[22] M. Goebel, L. Nataraj, T. Nanjundaswamy, T. M. Mohammed, S. Chandrasekaran *et al.,* “Detection, attribution and localization of gan generated images,” *Electronic Imaging*, vol. 2021, no. 4, pp. 276–1, 2021.

[23] Z. Guo, G. Yang, J. Chen and X. Sun, “Fake face detection via adaptive manipulation traces extraction network,” *Computer Vision and Image Understanding*, vol. 204, pp. 103170, 2021.

[24] Z. Liu, X. Qi and P. H. Torr, “Global texture enhancement for fake face detection in the wild,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8060–8069, 2020.

[25] B. Chen, X. Liu, Y. Zheng, G. Zhao and Y. Q. Shi, “A robust gan-generated face detection method based on dual-color spaces and an improved xception,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. http://dx.doi.org/10.1109/TCSVT.2021.3116679.

[26] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang *et al.,* “Multi-attentional deepfake detection,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 2185–2194, 2021.

[27] C. Wang and W. Deng, “Representative forgery mining for fake face detection,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 14923–14932, 2021.

[28] J. Frank, T. Eisenhofer, L. Schonherr, A. Fischer, D. Kolossa *et al.,* “Leveraging frequency analysis for deep fake image recognition,” in *Int. Conf. on Machine Learning*, Vienna, Austria, pp. 3247–3258, 2020.

[29] S. Agarwal, N. Girdhar and H. Raghav, "A novel neural model-based framework for detection of gan generated fake images," in *2021 11th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, pp. 46–51, 2021.

[30] Y. Qian, G. Yin, L. Sheng, Z. Chen and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European Conf. on Computer Vision*, Glasgow, UK, pp. 86–103, 2020.

[31] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li *et al.,* "Local relation learning for face forgery detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Palo Alto, California USA, vol. 35, no. 2, pp. 1081–1088, 2021.

[32] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He *et al.,* "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 772–781, 2021.

[33] M. C. Stamm, S. K. Tjoa, W. S. Lin and K. R. Liu, "Undetectable image tampering through jpeg compression anti-forensics," in *2010 IEEE Int. Conf. on Image Processing*, Hong Kong, China, IEEE, pp. 2109–2112, 2010.

[34] M. C. Stamm and K. R. Liu, "Anti-forensics of digital image compression," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1050–1065, 2011.

[35] M. Barni, M. Fontani and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. of the on Multimedia and Security*, New York, NY, United States, pp. 97–104, 2012.

[36] G. Cao, Y. Zhao, R. Ni, H. Tian and L. Yu, "Attacking contrast enhancement forensics in digital images," *Science China Information Sciences*, vol. 57, no. 5, pp. 1–13, 2014.

[37] W. Fan, K. Wang, F. Cayre and Z. Xiong, "Median filtered image quality enhancement and anti-forensics via variational deconvolution," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 1076–1091, 2015.

[38] D. T. Dang-Nguyen, I. D. Gebru, V. Conotter, G. Boato and F. G. De-Natale, "Counter-forensics of median filtering," in *2013 IEEE 15th Int. Workshop on Multimedia Signal Processing (MMSP)*, Pula, Italy, pp. 260–265, 2013.

[39] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 582–592, 2008.

[40] L. J. Lu, G. B. Yang and M. Xia, "Anti-forensics for unsharp masking sharpening in digital images," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 5, no. 3, pp. 53–65, 2013.

[41] N. Yu, L. Davis and M. Fritz, "Attributing fake images to gans: Analyzing fingerprints in generated images," arXiv preprint arXiv:1811.08180, vol. 2, 2018.

[42] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca *et al.,* "Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1038–1048, 2020.

[43] Y. Luo, H. Zi, Q. Zhang and X. Kang, "Anti-forensics of jpeg compression using generative adversarial networks," in *2018 26th European Signal Processing Conf. (EUSIPCO)*, Rome, Italy, pp. 952–956, 2018.

[44] J. Wu, L. Liu, X. Kang and W. Sun, "A generative adversarial network framework for jpeg anti-forensics," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, Auckland, New Zealand, pp. 1442–1447, 2020.

[45] J. Wu, X. Kang, J. Yang and W. Sun, "A framework of generative adversarial networks with novel loss for jpeg restoration and anti-forensics," *Multimedia Systems*, vol. 27, no. 6, pp. 1075–1089, 2021.

[46] Z. Shen, F. Ding and Y. Shi, "Anti-forensics of image sharpening using generative adversarial network," in *Int. Workshop on Digital Watermarking*, Chengdu, China, pp. 150–157, 2019.

[47] P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1125–1134, 2017.

[48] D. Kim, H. U. Jang, S. M. Mun, S. Choi and H. K. Lee, "Median filtered image restoration and anti-forensics using adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 278–282, 2017.

[49] H. Xie, J. Ni and Y. Q. Shi, "Dual-domain generative adversarial network for digital image operation anti-forensics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1701–1706, 2021.

[50] H. H. Nguyen, N. D. T. Tieu, H. Q. Nguyen-Son, J. Yamagishi and I. Echizen, "Transformation on computer generated facial image to avoid detection by spoofing detector," in *2018 IEEE Int. Conf. on Multimedia and Expo (ICME)*, San Diego, CA, USA, pp. 1–6, 2018.

[51] F. Peng, L. P. Yin, L. B. Zhang and M. Long, "Cgr-gan: Cg facial image regeneration for anti-forensics based on generative adversarial network," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2511–2525, 2019.

[52] F. Ding, G. Zhu, Y. Li, X. Zhang, P. K. Atrey *et al.,* "Anti-forensics for face swapping videos via adversarial training," *IEEE Transactions on Multimedia*, 2021. http://dx.doi.org/10.1109/TMM.2021.3098422.

[53] J. Wu, Z. Wang, H. Zeng and X. Kang, "Multiple operation image anti-forensics with wgan-gp framework," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, Lanzhou, China, pp. 1303–1307, 2019.

[54] J. Wu and W. Sun, "Towards multi-operation image anti-forensics with generative adversarial networks," *Computers & Security*, vol. 100, pp. 102083, 2021.

[55] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan *et al.,* "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[56] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[57] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.

[58] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu *et al.,* "Boosting adversarial attacks with momentum," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 9185–9193, 2018.

[59] M. Goebel and B. Manjunath, "Adversarial attacks on co-occurrence features for gan detection," arXiv preprint arXiv:2009.07456, 2020.

[60] Y. Wang, X. Ding, Y. Yang, L. Ding, R. Ward *et al.,* "Perception matters: Exploring imperceptible and transferable anti-forensics for gan-generated fake face imagery detection," *Pattern Recognition Letters*, vol. 146, pp. 15–22, 2021.

[61] X. Zhao, C. Chen and M. C. Stamm, "A transferable anti-forensic attack on forensic cnns using a generative adversarial network," arXiv preprint arXiv:2101.09568, 2021.

[62] X. Zhao and M. C. Stamm, "Making gan-generated images difficult to spot: A new attack against synthetic image detectors," arXiv preprint arXiv:2104.12069, 2021.

[63] D. Li, W. Wang, H. Fan and J. Dong, "Exploring adversarial fake images on face manifold," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 5789–5798, 2021.

[64] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding *et al.,* "Exploring frequency adversarial attacks for face forgery detection," arXiv preprint arXiv:2203.15674, 2022.

[65] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 3348–3357, 2021.