Journal of Cyber Security
DOI:10.32604/jcs.2020.014310
Review

Tech Science Press

An Overview of Face Manipulation Detection

Xingwang Ju[*]

School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[*]Corresponding Author: Xingwang Ju. Email: XwangJu@163.com

Received: 12 September 2020; Accepted: 19 October 2020

Abstract: Due to the power of editing tools, new types of fake faces are being created and synthesized, which has attracted great attention on social media. It is reasonable to acknowledge that one human cannot distinguish whether the face is manipulated from the real faces. Therefore, the detection of face manipulation becomes a critical issue in digital media forensics. This paper provides an overview of recent deep learning detection models for face manipulation. Some public dataset used for face manipulation detection is introduced. On this basis, the challenges for the research and the potential future directions are analyzed and discussed.

Keywords: Fake face; deep learning; faces manipulation detection

1 Introduction

Recently, tens of thousands of images and videos are updated on the Internet every day, but nearly half of them are manipulated for some benign reasons or malicious intent [1]. In fact, face manipulation as one of the most common parts is a very serious problem, because the face is not only an important interactive role but also used widely in many biological identification and authenticity devices [2]. So, the reasonable manipulation of face samples will seriously undermine trust in digital communications and security applications. However, the powerful image or video editing software such as FaceSwap (FS), FaceApp, etc., even an ordinary person can modify a facial image. Therefore, the spread of manipulated face has attracted widespread attention.

Nowadays, with the advent of various applications, there are mainly four face manipulations types: face swap manipulation, face synthesis manipulation, facial attributes manipulation, and facial expression manipulation [3]. Specifically speaking, face swap replaces the face of one person with another face. There are two different approaches: The classical computer graphics-based techniques such as FS [4], and the deep learning-based techniques known as Deep-Fakes [5]. Face synthesis creates entire nonexistent faces through the powerful GANs [6], e.g., StyleGAN [7]. Facial attributes carried out through GANs to modify some attributes of the face, such as the color of the hair or the skin, the age, the cover of hat or glasses, etc. Facial expression manipulation transfers facial expression from one face to another, which carries on the local modification to achieve the realistic effect, such as the Face2Face (F2F) [8], etc. As shown in Fig. 1, the samples for the four manipulations are presented.

In order to detect the face whether is real or manipulated, lots of methods have been proposed. The paper reviewed recent existing advanced detection methods from face swap manipulation, face synthesis manipulation, facial attributes manipulation and facial expression manipulation, respectively. The remainder of the paper is organized as: the public recognized datasets for these four types of detection are described in Section 2. The models for detection of face manipulations are reviewed in Section 3. Section 4 discusses the challenges and the future researches. Section 5 concludes the paper.

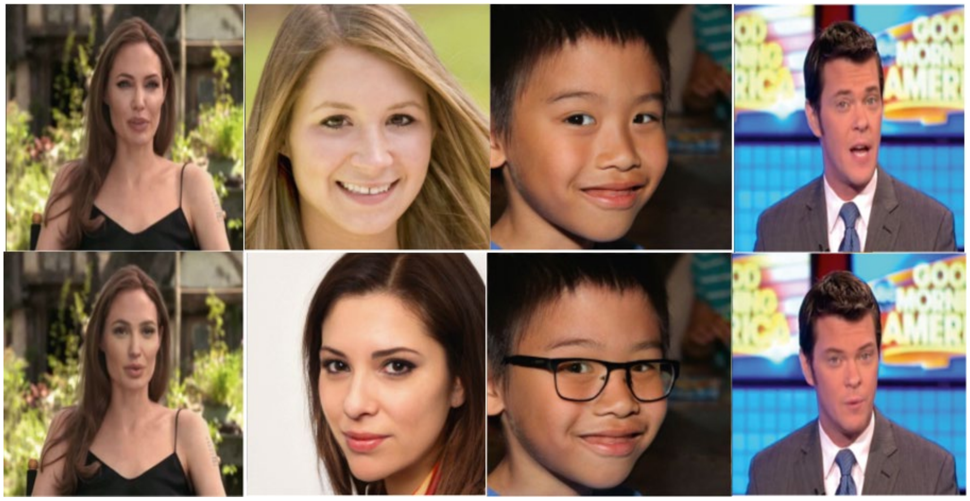

Figure 1: The samples for the four manipulations, respectively. The first line is real face images, the second line is the face manipulation images

2 Public Databases

For the four kinds of face manipulation detection, there are many datasets with different sizes for them. Here, we summarize some highly recognized and public datasets.

2.1 Face Swap

Regarding the dataset for the face swap, six public datasets are introduced. As show in Table 1, There are DeepFakeTIMIT (DFTIMIT) [9], UADFV [10], FaceForensics++ (FF++) [11], DeepFakeDetection [12], Celeb-DF [13], and DFDC Preview [14].

The DFTIMIT's database contains 620 fake videos from 32 topics. The fake video is created by the algorithm CycleGAN [15] with loading the weights of FaceNet [16]. It generates two different qualities of image: 64 × 64 (low-quality images), and 128 × 128 (high-quality images). UADFV [10] contains 49 fake videos created by the FakeApp mobile application with 49 real videos from the web of Youtube. For each video, the resolution is 294 × 500 pixels. FaceForensics++ [11], as the most widely used dataset, was created in 2019 and is an extension of the original FaceForensics database [17], which only focuses on facial expression manipulation. FF++ utilized the computer graphics and the learning method (DeepFake, (DF)) to create 1000 face swapping/identity exchange fake videos with the real videos from Youtube. Later, with the support of Google [12], a new data set named DeepFakeDetection was added to the FF++ dataset. This dataset contains a total of 3,068 fake videos. The videos are shot by 28 paid actors in 16 different scenes. FF++ and DeepFakeDetection databases contain different quality videos: Original quality, High quality (HQ) and Low quality (LQ). Due to the low visual quality of the previous database, and many visible artifacts. The Celeb-DF database [13] provides fake videos with better visual quality. Celeb-DF dataset contains 795 fake videos, which improved the low-resolution and the inconsistencies of the synthesized faces. Finally, The Facebook company collaborated with other companies and academic institutions to launch a challenge contest called Deep Detection Challenge (DFDC) [14]. They first released a preview dataset generated by two different unknown methods, which contained 4119 fake videos from 66 actors.

Table 1: Public datasets for face swap

Database	Fake videos
DFTIMIT (2018) [9]	620 (FS-GAN)
UADFV (2018) [10]	49 (FakeApp)
FF++ (2019) [11]	1000 (FS);1000 (DF)
DeepFakeDetection (2019) [12]	3068 (DF)
Celeb-DF (2019) [13]	795 (DF)
DFDC Preview (2019)[14]	4119 (Variety)

2.2 Face Synthesis

As presented in Tab. 2, the main publicly available databases for the entire face synthesis are presented. The real public face datasets were considered for their creation, such as CelebA [18], FFHQ [7], CASIA-WebFace [19], and others.

The 100K-Generated-Images database [7] contains 100,000 synthetic face images, which was generated by their proposed StyleGAN architecture. StyleGAN is an improved version of their previous ProGAN [20], which was trained with the FFHQ dataset. The 100K-Faces [21] database contains 100,000 composite images generated by StyleGAN. In this database, the difference is that the StyleGAN network was trained with 29,000 photos in more controlled scenes. The FSRemoval dataset [22] includes a total of 150,000 synthetic face images created by StyleGAN. In this database, the GAN fingerprint generated by StyleGAN is deleted from the original synthetic forged image by autoencoder, while maintaining the visual quality of the image, which provides a higher level of manipulation for the detection research. Finally, the new database of Diverse Fake Face Dataset (DFFD) [23] contains 100k and 200k fake images created by the pre-trained ProGAN [20] and StyleGAN [7]) models, respectively, which has not yet been made public.

Table 2: Public datsets for face synthesis

Database	Fake Images
100K-Faces (2019) [21]	100,000 (StyleGAN)
100K-Generated-Images (2019) [7]	100,000 (StyleGAN)
FSRemovalDB (2019) [22]	150,000 (StyleGAN)
DFFD (2019) [23]	100,000 (StyleGAN); 200,000 (ProGAN)

2.3 Facial Attributes

Facial attributes mainly include eyes, mouth, nose, eyebrows and other facial specific attributes. Human can modify these attributes to make the image more in line with people's aesthetic through some software applications. Regarding the public dataset, there is no publicly available dataset for facial attributes manipulations, but the research can obtain this type dataset easily. With the GAN-based model, the researchers can generate their own datasets as they like. The advanced GANs such as IcGANs [24], StarGAN [25], attGAN [26], and so on. The DFFD dataset mentioned above is also considered the facial attributes manipulations. But it is main for the face synthesis. So, the public datasets for Facial attributes manipulations is needed to be created for the research of detection.

2.4 Facial Expression

Facial expression manipulation is one of the facial manipulations, which as one of the most powerful and universal signals can convey emotional states and intentions. Facial unconscious expression analysis has important practical significance in many other human-computer interaction scenarios. There are also not many public datasets for facial expression manipulation detection. The only public dataset is the FF++ [11] mentioned above. However, many other existed methods now allow modification of facial expressions. For example, FaceApp, an app based on GAN architectures, can change the expression of one person from happier to angrier easily; Choi et al. [25] proposed the potential StarGAN, which can change the face to different levels such as happy, surprised, sad, and so on.

3. Manipulation Detection

Recently, different facial manipulation detection technologies have been proposed to detect whether the image is fake or real. In the following, the paper analyzes the performance and experiments of face manipulation detection models over the years based on the four manipulations, respectively.

3.1 Face Swap Manipulation Detection

The detection of face swap is one of the most research facial manipulation, which is evolving constantly. Regarding the detection, the comparison of classifiers, dataset and performance is presented in Tab. 3. Zhou et al. [27] considered two streams, one of which is the GoogLeNet [28], another stream is a path triplet stream with images patches as the input. Then, the SVM is used for classification. Afchar et al. [29] proposed the MesoInception-4 model to detect DeepFake with a private database, which utilized a variant of the Inception module [28] to optimize the structure of the model. They proved the robust performance with the FF++ dataset and achieved the performance of 0.984. Yang et al. [30] considered the face warping artifacts as evidence. They analyzed that current generation models require warping to match the original faces, which limits the resolution of the image created. According to this evidence, the authors detect whether such artifacts exists in the detected face regions and the surrounding areas. The different backbone networks are considered to extract the feature: VGG16, ResNet50, ResNet101, and ResNet152. They obtained outperformance results than other advanced methods at the UADFV and DFTIMIT databases.

To analyze the difference factors of extracted feature, Rossler et al. [11] utilized a variety of feature information to verify the effect with the FF++ database, There are: 1) Handcrafted steganalysis features [31] extracted by a CNN model is utilized. 2) A CNN model is designed to suppress the high frequency information of the image. 3) Four statistics features (mean, variance, maximum, and minimum) [32] were computed in the global pooling layer is considered. 4) The MesoInception detection model described in [29] is considered. 5) The Xception [33] model re-trained for the face swap task with loading the pre-trained weight on ImageNet database [34]. As for the results, the detection based on Xception achieved the best results in both DeepFakes and FaceSwap. Also, to adapt to the real scenario, the model was evaluated with the different video quality. It found that the accuracy decreases with the lower quality. In [35], Nguyen et al. utilized the multi-task learning to locate the manipulated regions and detect whether is manipulated. With the three tasks: classification, segmentation, and reconstruction, they shared the valuable information of them with autoencoder to improve the performance of the model. Based on FF++ database [17], they achieved the ERR with 0.015. The result is not good and seems not to have a generalization ability for other datasets. Stehouwer et al. [23] designed an attention mechanism through a convolution layer to deal with the feature maps of the classifier model, which can insert into any backbone networks easily. In the experiments, they provided state-of-the-art results among the DFFD dataset, FF++ dataset, and their own collected data from internet. The performance is AUC of 0.9943 and EER of 0.031.

Finally, to detect fake videos, Sabir et al. [36] utilized the temporal domain information as the evidences. The idea is to use the difference between frames. Therefore, instead of using a pre-trained model, they considered an end-to-end training with RNN. The detection model proposed was tested on the FF++ database, and the AUC results for the DeepFake and FaceSwap were 0.969 and 0.963, respectively.

Table 3: The comparison of the existing advanced detection methods for face swap manipulation. AUC, ACC and EER are the area under the curve, the accuracy, and the equal error rate

Papers	Classifiers	Datasets	Results
Zhou et al. [27]	CNN + SVM	UADFV	AUC = 0.851
		DFTIMIT (LQ, HQ)	AUC = 0.835, AUC = 0.735
		FF++ / DFD	AUC = 0.701
		Celeb-DF	AUC = 0.557
Afchar et al. [29]	CNN	UADFV	AUC = 0.843
		DFTIMIT (LQ, HQ)	AUC = 0.878, AUC = 0.627

		FF++ (DF, LQ, HQ)
		FF++ (FS, LQ, HQ)
			ACC = 0.90, ACC = 0.94
			ACC = 0.83, ACC = 0.93

		Celeb-DF	AUC = 0.536
Yang et al. [30]	SVM	UADFV	AUC = 0.89
		DFTIMIT (LQ, HQ)	AUC = 0.551, AUC = 0.532
		FF++ / DFD	AUC = 0.473
		Celeb-DF	AUC = 0.548

Rössler et al. [11]	CNN	FF++ (DF, LQ, HQ)
		FF++ (FS, LQ, HQ)
			ACC = 0.94, ACC = 0.98
			ACC = 0.93, ACC = 0.97

Nguyen et al. [34]	Autoencoder	UADFV
		DFTIMIT (LQ, HQ)
		FF++ / DFD
		FF++ (FS, HQ)
			AUC = 0.658
			AUC = 0.622, AUC = 0.553
			AUC = 0.763
			EER = 0.0151

Stehouwer et al. [23]	CNN + Attention	DFFD	AUC = 0.994, EER = 0.031
Sabir et al. [35]	CNN + RNN	FF++ (DF, LQ, FS, LQ)	AUC = 0.969, AUC = 0.963

3.2 Face Synthesis Manipulation Detection

Due to the appearance of various GAN networks, face synthesis becomes more and more realistic. The study of Face synthesis manipulations detection has come into notice. Regarding the detection models, the comparison of classifiers, dataset and performance is shown in Tab. 4. In [20], McCloskey et al. analyzed the different artifacts between GAN generated fake images and real images. Based on the color difference, they utilized the linear Support Vector Machine (SVM) to classify, and obtained a final 0.70 of AUC on the NIST MFC dataset [37].

Then, considering the specific fingerprints of GANs, Yu et al. [38] proposed a learning-based mechanism with a network to map the corresponding fingerprint of the input image. For each GANs, a model fingerprint is obtained, then the correlation between the image fingerprint and each model fingerprint can be used as the basis of classification. Based on the CelebaA dataset [18], and different GANs generated datasets such as ProGAN [20], SNGAN[39], and so on, they achieved the best performance of 0.995 accuracy. However, the model does not have good robust performance. In [23] mentioned above, they achieved impressive results with the attention mechanisms for the face synthesis manipulation on the fake images created through ProGAN [20] and StyleGAN [7] approaches. Inspired by steganalysis and natural image statistics, Nataraj et al. [40] proposed a detection model based on the combination of pixel co-occurrence matrix and CNN. For testing, they used a database of various objects and scenes created by CycleGAN [15]. Besides, in order to evaluate the robustness of the proposed method, the authors analyzed and achieved a good generalization effect on the fake images created by different GAN (CycleGAN and StarGAN). Considering the images from the 100K-Faces database [22], the best performance of 0.072 EER was obtained.

To evaluate this type of facial manipulation comprehensively, Neves et al. [22] conducted a series of experiments with the latest detection models under different experimental conditions (i.e., controlled and field scenes). They mainly considered two different fake databases: a database of 150,000 unique faces collected online, and the 100K faces. Under controlled conditions, they achieved similar results to the best previous studies (EER = 0.008). But, in scenarios where the images come from different sources, the performance will be highly degraded. In addition, they also proposed a novel method that can delete GAN fingerprint information from fake images, thereby deceiving the detection model. The results obtained show that for invisible scenes, a more powerful detection model needs to be developed. Considering the invisible conditions, Marra et al. [41] carried out an interesting study to detect invisible forged data types. Specifically, they proposed an incremental learning method with multi-task goal to detect new types of images generated by GAN without degrading the performance of previous images. Five different GAN methods CycleGAN [15], ProGAN [20], Glow [42], StarGAN [25] and StyleGAN [7] were considered. they have achieved encouraging results with the Xception model, which can detect the image manipulations by the new GAN.

Table 4: The comparison of the existing advanced detection methods for face synthesis manipulation. AUC, ACC and EER are the area under the curve, the accuracy, and the equal error rate

Papers	Classifiers	Datasets	Results
McCloskey et al. [20]	SVM	NIST MFC2018	AUC = 0.70
Yu et al. [38]	CNN	Own Dataset	ACC = 0.995
Stehouwer et al. [22]	CNN + Attention	DFFD	AUC = 1.0, EER = 0.001
Neves et al. [21]	CNN	100K-Faces (StyleGAN)	EER = 0.008
		FSRemovalDB (StyleGAN)	EER = 0.206
Marra et al. [41]	Incremental Learning	Own Dataset	ACC = 0.99.3

3.3 Facial Attributes Manipulation Detection

When the appearance of the FaceApp applications, the research of face attribute manipulation detection has been aroused. About this type of the datasets, two different approaches are considered: The GAN-based method such as the ProGAN [20], and the manual method such as utilizing Adobe Photoshop CS6.

As shown in the Tab. 5, the comparison of classifiers, dataset and performance is presented. Tariq et al. [43] used classification models such as VGG16, ResNet, and XceptionNet to evaluated the performance. They found that a high degradation of the performance is observed among the GAN-based method and manual-based method with AUC from 0.999 to 0.749. Bharati et al. [44] proposed a deep learning approach to detect digital retouching of face images, which is based on the Restricted Boltzmann Machine (RBM). They considered local and global regions of the face to deal with, such as the mask of glasses, hair, and hats. They achieved good performance for face attribute manipulation detection. Wang et al. [45] proposed the FakeSpoter model to extract more subtle features by monitoring neuron behavior serve as an asset, which is important for facial manipulation detection. From three different deep face recognition systems, the FakeSpoter extracted features and then trained an SVM for the final classification. Based on the dataset of InterFaceGAN [46] and StyleGAN [7], they achieved a final 0.847 accuracy with the best performance model FaceNet. Stehouwer et al. [23] utilized attention mechanisms to weigh the feature maps, and analyzed different facial manipulation methods completely. They obtained very good results with their novel database DFFD, which close to 0.999 of AUC and 0.001 error.

To synthesize new face dataset, Wang et al. [47] collaborate with professional artists to manipulate 50 real photographs with the Adobe Photoshop. They carried out a set of artificial classification study. They asked participants to categorize each image into one class. The results show that human cannot classify the image basically with a final of 0.535 accuracy. So, the authors designed a detection model based on Deep Recurrent Networks achieving 0.998 and 0.974 performances for automatic and manual face manipulations. Considering the spectrum domain, Zhang et al. [48] deal with each RGB channel with 2D DFT to obtain frequency feature. Then, the AutoGAN model is proposed as the classifier. To evaluate the performance of the generalization capacity, they used unseen GAN models, in particular, StarGAN [25] and GauGAN [49] were considered in the test. They achieve the 1.0 accuracy with the frequency domain features at the dataset created by StarGAN. However, a high degradation of the performance, accuracy with 0.50, was obtained for the dataset created by GauGAN approach.

Table 5: The comparison of the existing advanced detection methods for face attributes manipulation. AUC, ACC and EER are the area under the curve, the accuracy, and the equal error rate

Papers	Classifiers	Datasets	Results
Tariq et al. [43]	CNN	Own Dataset	AUC = 0.999
Bharati et al. [44]	RBM	Celebrity Retouching	ACC = 0.962
		ND-IIITD Retouching	ACC = 0.871
Wang et al. [45]	SVM	Own Dataset	ACC = 0.847
Stehouwer et al. [22]	CNN + Attention	DFFD	AUC = 0.999, EER = 0.01
Wang et al. [47]	DRN	Own Dataset	AP = 0.998
Zhang et al. [48]	GAN Discriminator	Own Dataset	ACC = 1.0

3.4 Facial Expression Manipulation Detection

Facial expressions as one of the most powerful, natural, and universal signals can convey emotional states and intentions. Facial unconscious expression analysis has important practical significance in many other human-computer interaction scenarios. In the appearance of a video with a person from [50], the facial expressed modification draws attention and motivated the research to detect.

As shown in the Table 6, the comparison of classifiers, dataset and performance is presented. In [29], the Meso-inception model mentioned above achieved good performance, especially for raw-quality videos. But, when the model was tested on NeuralTextures fake videos [51], which have lower performance than the F2F dataset. In [17], the Xception model has the best performance close to 1.0 at F2F and NeuralTextures datasets. In [52], the proposed model only obtain 0.866 AUC with the Face2Face manipulation in FF++ database. In [35], they achieved 0.071 EER with HQ videos for FF++ database. And for the NeuralTexture method, the EER is a final 0.078 EER. In addition, Amerini et al. [53] considered the dissimilarities between inter-frame by applying the optical flow fields. The optical flow can extract the apparent motion between the observers. From this perspective, the optical flow matrices are suited for the unusual movement of lips, eyes, etc, that existed in fake videos. The best performance with 0.816 is obtained with the VGG16 and ResNet50 networks.

4 Challenges and Future Research Directions

Although many researchers are committed to face manipulation detection and have achieved fine performance mention in Section 3. But there are still many challenges in this area. With the development of GANs, many manipulated images are not distinguishable from real images by using some simple models, and the residual manipulation trajectories of these manipulated images are optimized by the GAN model to approximate to the real image, so the work in this aspect still needs to be continuously overcome.

Table 6: The comparison of the existing advanced detection methods for face expression manipulation. AUC, ACC and EER are the area under the curve, the accuracy, and the equal error rate

Papers	Classifiers	Datasets	Results
Afchar et al. [29]	CNN	UADFV	AUC = 0.843
		DFTIMIT (LQ, HQ)	AUC = 0.878, AUC = 0.627
		FF++ (DF, LQ, HQ)	ACC = 0.900, ACC = 0.940
		FF++ (FS, LQ, HQ)	ACC = 0.830, ACC = 0.930
		Celeb-DF	AUC = 0.536
Matern et al. [52]	LR, MLP	FF++ (F2F)	AUC = 0.866
Nguyen et al. [34]	Autoencoder	UADFV	AUC = 0.658
		DFTIMIT (LQ, HQ)	AUC = 0.622, AUC = 0.553
		FF++ / DFD	AUC = 0.763
		FF++ (FS, HQ)	EER = 0.015
Amerini et al. [53]	CNN + Optical Flow	FF++ (F2F)	ACC = 0.816

4.1 Public Dataset for Facial Attributes and Facial Expression

To improve the detection effect, a large number of databases are necessary to support the training. But the creation of datasets is time-consuming and labor-intensive to produce, and there are no shortcuts. But it had to be overcome. The public, recognized datasets for the facial attributes and facial expressions manipulations are still lacking. Many researchers carried out experiments at the datasets that they created, but the conditions of the created datasets are different. As a result, when several models are compared with each proposed model, it is less convincing. Therefore, it is necessary to create public, recognized, high-quality datasets with various scenarios.

4.2 Robustness and Generalization

Many methods have limitations and may only be applicable to a specific situation, but do not achieve robustness. When the data source is attacked, the existed solutions may fail. How to ensure good performance under different circumstances, that is to say, it is a very challenging thing to develop a more comprehensive detection system. Due to the high computing cost, many comprehensive models are not suitable for mobile application detection, so it is necessary to develop simple and portable small detection models.

In addition to robustness, the generalization ability is also considered as a manifestation of the performance of the model. Many methods have close to 100 percent effectiveness on in-library testing, but when test with cross-library, the results dropped significantly. With constant exploration, some researchers have found that fine-tuning can provide some relief that retraining the model with new datasets, but this will degrade the performance of the model against the original data. This is a very common 'disaster forgetting' phenomenon in models. So, how to design a sound detection system with good generalization performance is very challenging work.

5 Conclusion

The face manipulation detection method is concluded and analyzed in this paper. From the four main face manipulation types, firstly, this paper introduces the public recognized datasets for the face swap manipulation, face synthesis manipulation, facial attributes manipulation, and facial expression manipulation, respectively. Then, this paper reviewed the existing advanced detection methods, the structures and results of each model are analyzed. Finally, some existed problems and challenges are discussed in the paper, which requires future research.

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] V. Conotter, E. Bodnari, E. Boato and H. Farid, "Physiologically-based detection of computer-generated faces in video," in *Proc. IEEE Int. Conf. on Image Processing*, pp. 248–252, 2014.

[2] Z. Akhtar, D. Dasgupta and B. Banerjee, 'Face authenticity: An overview of face manipulation generation, detection and recognition," in *Proc. Int. Conf. on Communication and Information Processing*, 2019.

[3] R. Tolosana, R. Vera-Rodriguez and J. Fierrez, "Deepfakes and beyond: A survey of face manipulation and fake detection," arXiv preprint arXiv: 2001.00179, 2020.

[4] Faceswap-GAN, 2019. [Online]. Available: https://github.com/shaoanlu/faceswap-GAN

[5] Deepfakes_faceswap, 2020. [Online]. Available: https://github.com/deepfakes/faceswap

[6] T. Karras, S. Laine and T. Aila, "A Style-Based generator architecture for generative adversarial networks," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2019.

[7] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, 2018.

[8] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2016.

[9] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," arXiv preprint arXiv:1812.08685, 2018.

[10] Y. Li, M. Chang and S. Lyu, "In ICTU oculi: Exposing AI generated fake face videos by detecting eye blinking," *in Proc. Int. Workshop on Information Forensics and Security*, 2018.

[11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. Int. Conf. on Computer Vision*, 2019.

[12] Google AI, "Contributing data to deepfake detection research," 2019. [Online]. Available: https://ai.googleblog.com/2019/ 09/contributing-data-to-deepfake-detection.html

[13] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A new dataset for deepFake forensics," arXiv preprint arXiv: 1909.12962, 2019.

[14] B. Dolhansky, R. Howes, B. Pflaum, N. Baram and C. Ferrer, "The deepfake detection challenge (DFDC) Preview Dataset," arXiv preprint arXiv: 1910.08854, 2019.

[15] J. Zhu, T. Park, P. Isola and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. on Computer Vision*, 2017.

[16] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2015.

[17] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv: 1803.09179, 2018.

[18] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep learning face attributes in the wild," in *Proc. Int. Conf. on Computer Vision*, 2015.

[19] D. Yi, Z. Lei, S. Liao and S. Li, "Learning face representation from scratch," arXiv preprint arXiv: 1411.7923, 2014.

[20] 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*, 2018.

[21] 100,000 Faces Generated by AI, 2018. [Online]. Available: https: //generated.photos

[22] J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes and H. Proença, "Real or fake? Spoofing state-of-the-art face synthesis detection systems," arXiv preprint arXiv: 1911.05351, 2019.

[23] J. Stehouwer, H. Dang, F. Liu, X. Liu and A. Jain, "On the detection of digital face manipulation," arXiv preprint arXiv: 1910.01717, 2019.

[24] G. Perarnau, J. V. D. Weijer, B. Raducanu and J. A lvarez, "Invertible conditional GANs for image editing," in *Proc. Advances in Neural Information Processing Systems Workshops*, 2016.

[25] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain imageto- image translation," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2018.

[26] 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*, 2019.

[27] B. Dolhansky, R. Howes, B. Pflaum, N. Baram and C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," arXiv preprint arXiv:1910.08854, 2019.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2015.

[29] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: A compact facial video forgery detection network," *in Proc. Int. Workshop on Information Forensics and Security*, 2018.

[30] Y. Li and S. Lyu, "Exposing deepFake videos by detecting face warping artifacts," in *Proc. Conf. on Computer Vision and Pattern Recognition Workshops*, 2019.

[31] D. Cozzolino, G. Poggi and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural nNetworks: an application to image forgery detection," in *Proc. ACM Workshop on Information Hiding and Multimedia Security*, 2017.

[32] N. Rahmouni, V. Nozick, J. Yamagishi and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. Workshop on Information Forensics and Security*, 2017.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2017.

[34] J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2009.

[35] H. Nguyen, F. Fang, J. Yamagishi and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," arXiv preprint arXiv: 1906.06876, 2019.

[36] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *in Proc. Conf. on Computer Vision and Pattern Recognition Workshops,* 2019.

[37] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith and J. Fiscus, "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *Proc. IEEE Winter Applications of Computer Vision Workshops*, pp. 63−72, 2019.

[38] N. Yu, L. Davis and M. Fritz, "Attributing fake images to GANs: Analyzing fingerprints in generated images," in *Proc. Int. Conf. on Computer Vision*, 2019.

[39] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. on Learning Representations*, 2018.

[40] L. Nataraj, T. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J. Bappy and A. Roy-Chowdhury, "Detecting GAN generated fake images using co-occurrence matrices," arXiv preprint arXiv: 1903.06836, 2019.

[41] F. Marra, C. Saltori, G. Boato and L. Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," in *Proc. Int. Workshop on Information Forensics and Security*, 2019.

[42] D. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Advances in Neural Information Processing Systems*, 2018.

[43] S. Tariq, S. Lee, H. Kim, Y. Shin and S. Woo, "Detecting both machine and human created fake face images in the Wild," in *Proc. Int. Workshop on Multimedia Privacy and Security*, pp. 81−87, 2018.

[44] A. Bharati, R. Singh, M. Vatsa and K. Bowyer, "Detecting facial retouching using supervised deep learning," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1903−1913, 2016.

[45] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang and Y. Liu, "FakeSpotter: A simple baseline for spotting AI-synthesized fake faces," arXiv preprint arXiv: 1909.06122, 2019.

[46] Y. Shen, J. Gu, X. Tang and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," arXiv

preprint arXiv: 1907.10786, 2019.

[47] S. Wang, O. Wang, A. Owens, R. Zhang and A. Efros, “Detecting photoshopped faces by scripting photoshop,” arXiv preprint arXiv: 1906.05856, 2019.

[48] X. Zhang, S. Karaman and S. Chang, “Detecting and simulating artifacts in GAN fake images,” arXiv preprint arXiv: 1907.06515, 2019.

[49] P. Isola, J. Zhu, T. Zhou and A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2017.

[50] S. Suwajanakorn, “Fake videos of real people and how to spot them,” 2019. [Online]. Available: https://www.ted.com.

[51] J. Thies, M. Zollhfer and M. Niener, “Deferred neural rendering: Image synthesis using neural textures,” *ACM Transactions on Graphics*, vol. 38, no. 66, pp. 1−12, 2019.

[52] F. Matern, C. Riess and M. Stamminger, “Exploiting visual artifacts to expose deepFakes and face manipulations,” in *Proc. IEEE Winter Applications of Computer Vision Workshops*, 2019.

[53] I. Amerini, L. Galteri, R. Caldelli and A. Bimbo, “Deepfake video detection through optical flow based CNN,” in *Proc. Int. Conf. on Computer Vision*, 2019.