A SURVEY ON DEEPFAKES CREATION AND DETECTION

Rajnandini Santosh Bhingare , Sunil Hirekhan

Department of Electronics and Telecommunication, Government College of Engineering, Aurangabad.

Abstract: In the evolutionary world of technology, Deep learning has been skillfully employed to resolve different kinds of high ranging problems which human cannot resolve. However Deeplearning have also been applied to generate fake content which can challenge national democracy, privacy, security threats etc. One of those deep learning impowered approach is "DeepFake". DeepFake technique can generate fake images and videos that humans cannot make difference between forged and real media. This may lead to threatening to world security as well as privacy. The malicious use of this technique exceeding than positive use day by day. To prevent and control this threat various researches worked to detect it for resolving the problem.

In this survey paper, we are going to see manipulation techniques, types of DeepFake creation and detection with reference to previous work on DeepFakes.

I. INTRODUCTION

By using digital manipulation technique Fake images and videos which include facial data bring about specifically by DeepFake techniques. The word DeepFake refers to Deep Learning and Fake. Deepfake is a method in which the fake images or videos can be created by switching the faces of a person to the face of another person. Deepfake images can be created by using easy to handle and imposing tools like GAN (Generative Adversial Network). To handle the public belief many unrealistic things like fake news, celebrity videos swapped faces images can be created by using DeepFake method [1],[2].

In 2017, the first deepfake video was created by switching the face of a well-known person to the porn actor. For misreporting purpose, videos of famous leader speeches were created and this was frightening to the world reliability. Some DeepFakes are not very laborious to detect as they are created for the entertaining purpose. Nevertheless, finding the digital cohesion is tough if the Deepfake image or video includes ordinary or a common person.

Research area is enormously increasing, for the detection of DeepFaake images and videos. Searching out the facts in digital field, it is more and more condemnatory. This fake detection is carried out by some international projects like DARPA (Defense Advanced Research Project Agency) which supported MediFor (Media Forensics). Also, NIST (National Institute of Standards and Technology) started Media Forensics Challenge (MFC18). For the Deepfake ease, Facebook has been operating on

detecting models and their attempts are accelerating the detection and verification. Lately, Facebook undertaken the DeepFake Detection Challenge (DFDC) COLLABORATED with Microsoft and high geared AI model in challenge could detect artificial videos with 82.56% accuracy. [3]

II. DeepFake Manipulation Techniques

Depending upon the level of the manipulation, the facial manipulations can be classified into four categories: namely,

1] Entire Face Synthesis 2] Identity Swap 3] Attribute Manipulation 4] Expression Swap

1. Entire Face Synthesis:

By using a significant Deep learning technique i.e., GAN; this technique by T. Karras et al. generates completely unreal face images. Recently StyleGAN generated an excellent face image with advanced reality.[7]

2. Identity Swap:

The perspective of this identity swap is carried out with the help of two different methods: i) FaceSwap ii) DeepFake. In this type of manipulation, the face of a source person in video is swapped with the face of target person.

3. Attribute Manipulation:

For this manipulation StarGAN is used by Y. Choi et al [9]. This technique is carried out by E. Gonzalez et al [8]. for changing the features of face like hair or skin color, gender of a person, adding or removing glasses, etc., This method of manipulation is also known as face modifying technique.

4. Expression Swap: Generally, this technique is focuses on Face2Face and Neural Textures by J. Thies et al [10]. In this type of manipulation expressions are modified, i.e., expression of a source face swapped with the expression of target face. [11]

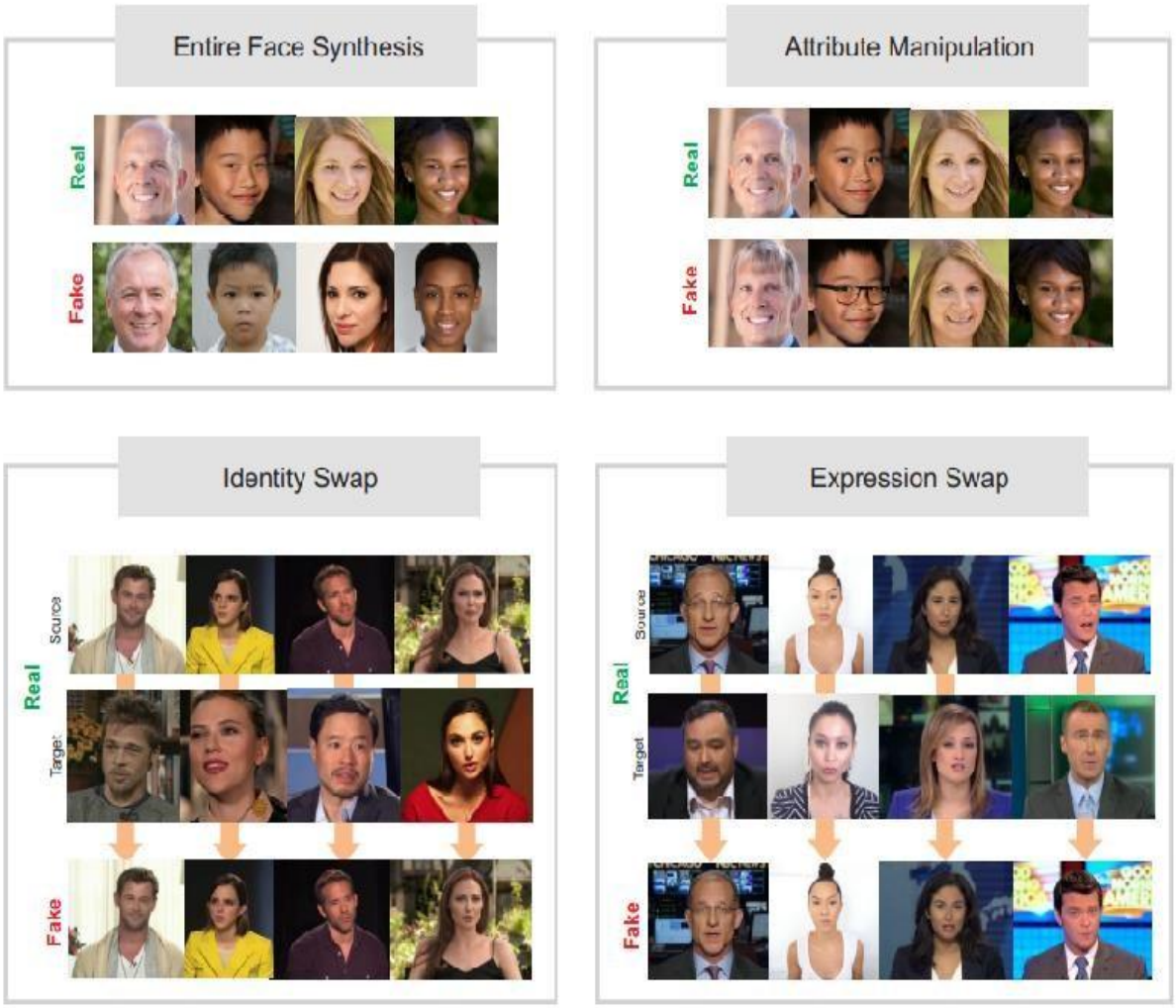

Fig1: example of Entire face synthesis from http://www.whichfaceisreal.com/ and fake images from https://thispersondoesnotexist.com., Attribute manipulationreal images are extracted from http://www.whichfaceisreal.com/ and fake images are generated using FaceApp, identity swap, face images are extracted from Celeb-DF database, Expression Swap images are extracted from FaceForensics++.

III. Generation of DeepFakes

Day by day, number of techniques are introduced for manipulate perceptible denotations. Most of the users are having different computer abilities and high quality of amended videos cause DeepFake have eventually favored. For generating fake content most usual procedures are appending, detaching or eliminating items from an image are widely in use.

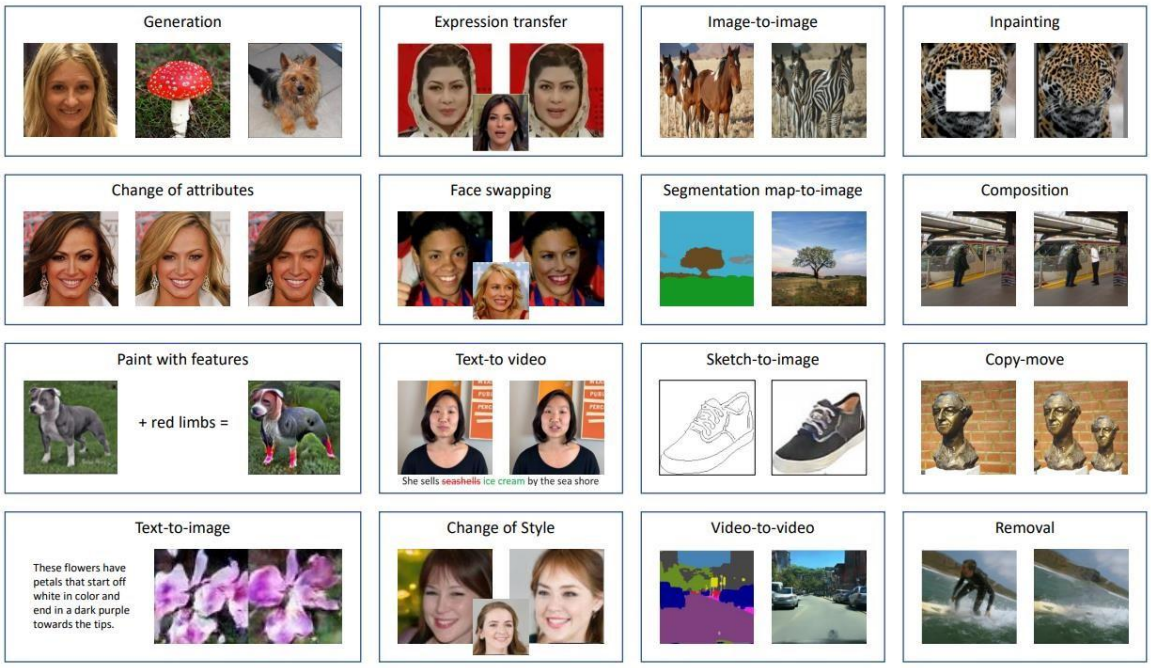

Fig2: Image and Video Manipulation. [1].

To increase the reasonable content in perceptible aspect Operations like, splicing, copy-move, inpainting [Splicing: putting a new object by reprinting it through another image. Copy-move:

Reprinting an object from same image. Inpainting: to complete an image missing data or information is filled.] can be done by various DeepFake tools which are extensively in use.

Deep learning-based approaches (like GAN) [12-14] are popular for its effectiveness of serving complicated and high-spatial information. A deep network called auto encoder decoder is the suitable alternative of deep learning which is extensively useable for spatial declining and image contraction. This method of generating DeepFake was first used by Reddit with FakeApp where encoder- decoder pair was applied [15], [16]. In this method, an encoder is designed for reducing image outline and decoder is designed for regeneration of face image. For switching the source face with the target face, two pairs encoder-decoder is needed. Each pair of encoder-decoder is used to instruct an image set. The spatial characteristics of encoder are aggregated within two pairs of networks.

In short, auto encoder and GAN by H. Haung et al [17]. accepted to update the powerful explanation unusually for face manipulation while an excellent level of realistic image has been attained.

E.Zakharov et al [20]. introduced image tampering can be attained with the help of sketch or T. Park et al [21]. a text description. For the manipulation of an image StyleGAN accepted to modify the painting style, swapping of apples and oranges by P.Isola et al [22]. For expression modification swapping faces many methods are launched by Y. Choi, et al [23]. C. Chan et al [24]. newly launched methods are subject to the motion transformation from source person to target person.

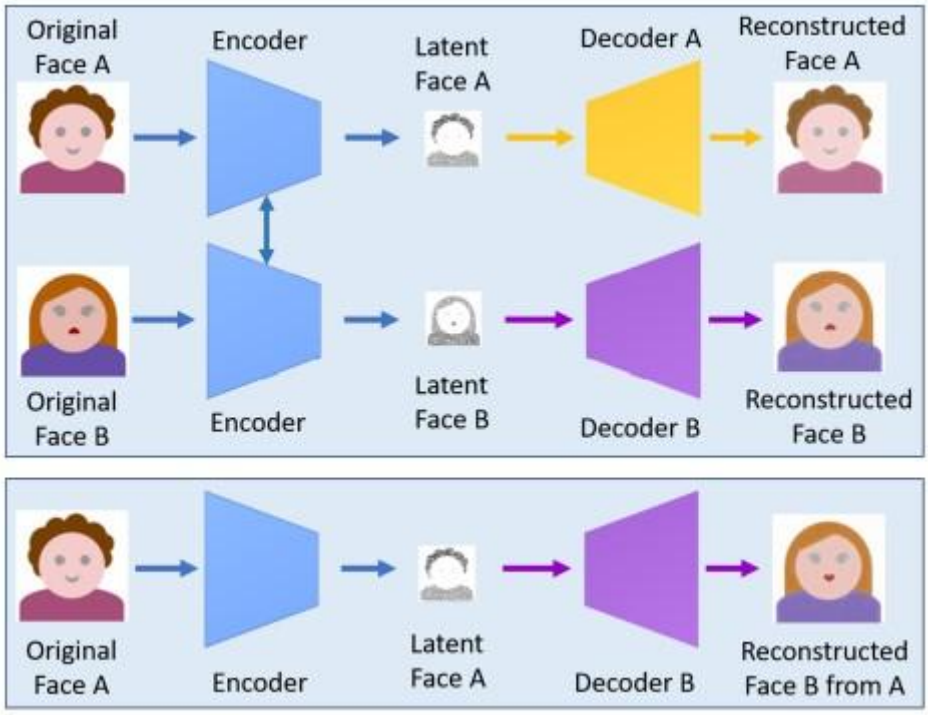

Fig3: Encoder- Decoder pair. To create deepfake 2 pairs of encoder and decoder are used. In first pair original faces are encoded and in second pair Face of A is decoded with face of B. [3]

IV. Detection of DeepFakes

Nowadays, a threat of generating tampered media using DeepFake tools become a critical question. Unusually, non-professional individuals can also easily generate fake videos as enough information is obtained from internet. With DeepFake tools, non-existing faces can be created by using GAN technique and tampering of faces in video clips line up for modifying specifications. As soon as this threat of DeepFake came into picture detecting DeepFake methods have been launched. The forged video synthesis action was acquired by natural attributes known as handcrafted features which were derived from traces and differences in untimely experiments. For detection of forged images and videos different methods are used which are as follows:

A. Detection of Forged Image

The faces from images can be swapped by using the set of data from the abundance. In video fusion, face swapping has more proposals like conversion into pictures and unusually in security purposes which leads to fascinate.

The Deep-learning approaches like CNN and GAN are used for swapping faces, this technique has made face swapping farther difficult to verify. Zang et al [25]. adapted the bag of word s to cite a set of attributes and provided it into a variety of classifiers like SVM, Random Forest (RF) AND Multilayer Perceptron (MLP) for distinguishing between forged faces from the real.

The DeepFakes which are generated by using GAN technique are more complicated and not easy to verify as they are genuine or forged. X. Xuan et al [26]. replaced another approach in which pre- and post-processing functions are useable for expansion of information and to enhance the portability.

The performance exhibits as CNN-created images distributes some conventional defects which permits one to artifacts its nature even over concealed structure, details and methods for training. In inconsistent right plane, CNN which includes universal image consistency data which exhibits positive conception work application that is another explanation Z.Liu et al [27].

Fig4: Eyes color is different (top), Teeth are abruptly shaped. [1]

B. Detection of Forged Videos

The fame information is powerfully derogated after video compression hence utmost image detection techniques cannot be applicable for video. Some of the methods are designed only for motionless frames and video frames have sensual attributes and these are different among frame sets. Actually, the fake faces which are created with CGI and deep learning tools are not much distinct hence, they both are deficient in distinctive appearances which are characteristics of human faces captured by genuine cameras.

D.-T. Dang-Nguyen et al [28] work's detection counts dimensional transitory distortion of a 3D model which applies for faces. Specially, real faces are complicated and have different dimensional directions and this encourages more concern of 3D model. There are two methods to detect manipulated videos which are as: detection using sequential attributes through frames and method such that investigates into frames using Deep learning.

1. Sequential attribute along video frames:

Manipulating the utilization of spatio-sequential attributes of video series to expose deepfakes. Sabir et al [28]. reviewed that sequential reasoning is not imposed efficiently in the emulsion procedure of deepfakes.

By using some distinct visual artifacts, deepfakes leverages in video can be described. These types of fakes can be seen in the eyes and teeth. For example, missing or depicted as white spot in eye or abruptly shaped teeth which can be seen as white spots as shown in fig. (4). This was observed by F. Matern et al [30].

Y. Li, et al [31]. introduced a system which is depends on blinking of eye and has a distinct rate and time period in persons which is not emulated in forged videos. Some other methods are depended on deformation fragments [32], countenance feature position [33] or head posture unreliability [34].

In [32], to match the source face in video more deformed fragmentations are needed as deepfake techniques can only created finite closure images. CNN works on face section and its adjacent region. Nevertheless, deformation permits unusual clones which are detected by CNN.

In [33], GAN based system can create high quality of realistic faces and with lots facts but only deficient an exact discipline over areas of some parts of face. Because of this area of facial section, such as eyes, mouth and nose can be employed as the discriminative features for detecting the genuineness of the GAN images.

The main expediency of this technique is that the visual artifacts are not strained by resizing and compression. Also, some fake media can be identified by means of handcrafted solutions which incorporate declined risk.

2. Detection of Video frames using Deep learning

To detect deepfake videos the methods using sequential attributes along frames are depend on deep recurrent network models. Generally, video frames are defragmented and evaluates inside a single frame to achieve discriminant attributes. These attributes are then allocated to deep or shallow classifiers to distinguish between fake and realistic videos.

In general, deepfake videos are generated with finite closure, which need similar face deformation to achieve the parallel pattern of real faces. In [35], two elementary surveying has four levels of pooling and convolutions and then developed by robust system with one latent layer. The second surveying is alternatively derived from a different outset that has extended convolutions.

V. Conclusion

Ongoing achievements of digital leverages, especially DeepFakes this survey deals with the manipulation type, methods to generate DeepFakes and detect DeepFake images and videos separately.

Generally, most of the manipulating faces can be detect and controlled easily as they may be generated with CGI. In fact, the fake media generated using deep learning tools then detection task might be difficult. However, this scenario can be changed with the help of continuously improving detection techniques.

To provide robust tool, alternative for traditional image/video detectors, Tursman et al. introduced a real time system [36]. Such achievement can further defend media denotations from harm.

REFERENCES

[1] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910-932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.

[2] Tolosana, Ruben & Vera-Rodriguez, Ruben & Fierrez, Julian & Morales, Aythami & Ortega-Garcia, Javier. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion. 64. 10.1016/j.inffus.2020.06.014.

[3] Nguyen, Thanh & Nguyen, Cuong M. & Nguyen, Tien & Nguyen, Duc & Nahavandi, Saeid. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey.

[4] [4]Bloomberg (2018, September 11). How faking videos became easy and why that's so scary. Available at https://fortune.com/2018/09/11/deepfakes-obama-video/

[5] Chesney, R., and Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Affairs, 98, 147.

[6] Schroepfer, M. (2019, September 5). Creating a data set and a challenge for deepfakes. Available at https://ai.facebook.com/blog/deepfakedetection-challenge.

[7] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[8] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso Fernandez, "Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation," IEEE Transactions on Information Forensics and Security, vol. 13, no. 8, pp. 2001–2014, 2018.

[9] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image to-Image Translation," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[10] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-Time Face Capture and Reenactment of RGB Videos," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[11] J. Thies, M. Zollhofer, and M. Nießner, "Deferred Neural Rendering: ¨ Image Synthesis using Neural Textures," ACM Transactions on Graphics, vol. 38, no. 66, pp. 1–12, 2019.

[12] Punnappurath, A., and Brown, M. S. (2019). Learning raw image reconstruction-aware deep image compressors. IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2019.2903062.

[13] Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019). Energy compaction-based image compression using convolutional autoencoder. IEEE Transactions on Multimedia. DOI: 10.1109/TMM.2019.2938345. [14] Chorowski, J., Weiss, R. J., Bengio, S., and Oord, A. V. D. (2019). Unsupervised speech representation learning using wavenet autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 27(12), pp. 2041-2053.

[14] Faceswap: Deepfakes software for all. Available at https://github.com/deepfakes/faceswap

[15] FakeApp 2.2.0. Available at https://www.malavida.com/en/soft/fakeapp/

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[17] H. Huang, P. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," arXiv:1803.04469v2, 2018.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in International Conference on Learning Representations, 2018.

[19] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," arXiv preprint arXiv:1905.08233v2, 2019.

[20] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially adaptive normalization," in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346. [22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[21] J. Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-toimage translation using cycle-consistent adversarial networks," in IEEE International Conference on Computer Vision, 2017.

[22] C. Chan, S. Ginosar, T. Zhouy, and A. Efros, "Everybody dance now," in International Conference on Computer Vision, 2019.

[23] Zhang, Y., Zheng, L., and Thing, V. L. (2017, August). Automated face swapping and its detection. In 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP) (pp. 15-19). IEEE.

[24] X. Xuan, B. Peng, W. Wang, and J. Dong, "On the generalization of GAN image forensics," in Chinese Conference on Biometric Recognition, 2019.

[25] Z. Liu, X. Qi, and P. Torr, "Global texture enhancement for fake face detection in the wild," arXiv preprint arXiv:2002.00133v3, 2020.

[26] D.-T. Dang-Nguyen, G. Boato, and F. De Natale, "3D-model-based video analysis for computer generated faces identification," IEEE Transactions on Information Forensics and Security, vol. 10, no. 8, pp. 1752–1763, Aug 2015.

[27] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," IEEE Trans. Inf. Forensics Security, vol. 7, no. 3, pp. 1003–1017, 2012.

[28] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in IEEE WACV Workshop on Image and Video Forensics, 2019.

[29] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye," in IEEE Workshop on Information Forensics and Security, 2018.

[30] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in IEEE CVPR Workshops, 2019.

[31] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing GAN-synthesized faces using landmark locations," in ACM Workshop on Information Hiding and Multimedia Security, June 2019, pp. 113–118.

[32] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head pose," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[33] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in IEEE International Workshop on Information Forensics and Security, 2018, pp. 1–7.

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, "Densely connected convolutional networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4700–4708.