DEEPFAKE DETECTION: CURRENT CHALLENGES AND NEXT STEPS

Siwei Lyu

Computer Science Department
University at Albany, State University of New York

ABSTRACT

High quality fake videos and audios generated by AI-algorithms (the deep fakes) have started to challenge the status of videos and audios as definitive evidence of events. In this paper, we highlight a few of these challenges and discuss the research opportunities in this direction.

***Index Terms*—** DeepFake videos, detection techniques, digital media forensics

1. INTRODUCTION

Falsified videos created by AI algorithms, in particular, deep neural networks (DNNs), are a recent twist to the disconcerting problem of online disinformation. Although fabrication and manipulation of digital images and videos are not new [1], the rapid development of DNNs in recent years has made the process to create convincing fake videos increasingly easier and faster. DNN generated fake videos first caught the public's attention in late 2017, when a Reddit account with name *Deepfakes* began posting synthetic pornographic videos generated using a DNN-based face-swapping algorithm. Subsequently, the term DeepFake have been used more broadly to refer to any AI generated impersonating videos.

Currently, there are three major types of DeepFake videos.

- Head puppetry entails synthesizing a video of a target persons whole head and upper-shoulder using a video of a source persons head, so the synthesized target appears to behave the same way as the source.
- Face swapping involves generating a video of the target with the faces replaced by synthesized faces of the source while keeping the same facial expressions.
- Lip syncing is to create a falsified video by only manipulating the lip region so that the target appears to speak something that s/he does not speak in reality.

Figure 1 shows some example frames of each type of DeepFake videos aforementioned. As the first examples of DeepFakes, face swapping has been commercialized and mainstreamed through readily available software freely available on GitHub, *e.g.*, `FakeApp` [2], `DFaker` [3], `faceswap-GAN` [4], `faceswap` [5], and `DeepFaceLab` [6]. There are also emerging online services that can generate DeepFake videos on demand (`https://deepfakesweb.com`), and there are many online discussion fora on DeepFakes. Furthermore, several start-up companies also commercialized tools that can potentially be used to make DeepFakes, such as Synthesia[1] and Canny AI[2].

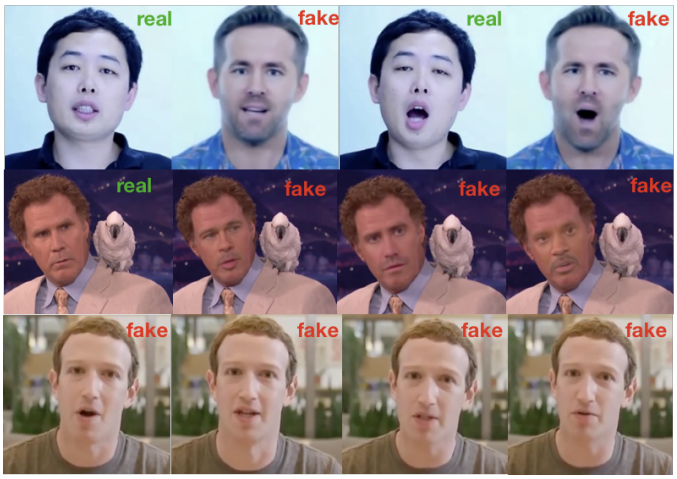

Fig. 1. *Examples of DeepFake videos: (top) Head puppetry, (middle) face swapping, and (bottom) lip syncing.*

While there are interesting and creative applications of the DeepFake videos, due to the strong association of faces to the identity of an individual, they can also be weaponized. Well-crafted DeepFake videos can create illusions of a person's presence and activities that do not occur in reality, which can lead to serious political, social, financial, and legal consequences [7]. The potential threats range from revenge pornographic videos of a victim whose face is synthesized and spliced in, to realistically looking videos of state leaders seeming to make inflammatory comments they never actually made, a high-level executive commenting about her company's performance to influence the global stock market, or an online sex predator masquerades visually as a family member or a friend in a video chat. The high stakes spawn wide media coverage of this topic in the past two years, and the US congress has had two public hearings to this problem.

With the escalated concerns over DeepFakes, there is a

[1] `https://www.synthesia.io/`.
[2] `https://www.cannyai.com/`.

surge of interest in developing DeepFake detection methods with significant progress witnessed in the past two years. This includes (1) a slew of effective detection methods developed in less than two years, mostly based on deep learning [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]; (2) the availability of several large-scale DeepFake video datasets [11, 19, 15, 20, 21]; and (3) two public challenges dedicated to DeepFake detection, namely, the DARPA MFC18 Synthetic Data Detection Challenge and the Facebook *DeepFake Detection Challenge*[3].

Notwithstanding this progress, there are a number of critical problems that are yet to be resolved for existing DeepFake detection methods. Furthermore, in the foreseeable future, it is expected that the generation of DeepFake videos will continue evolving, it is thus important to anticipate such new developments and improve the detection methods accordingly. The main objective of this paper is to highlight a few of these challenges and discuss the research opportunities in this direction.

2. CURRENT DEEPFAKE DETECTION METHODS

Current DeepFake detection methods mostly target face-swapping videos, which account for the majorities of DeepFake videos circulated online. Many of the existing methods are formulated as frame-level binary classification problems. Based on the features that are used, these methods fall into three major categories. Methods in the first category are based on inconsistencies exhibited in the **physical/physiological** aspects in the DeepFake videos. The method in work of [10] exploits the observation that many DeepFake videos lack reasonable eye blinking due to the use of online portraits as training data, which usually do not have closed eyes for aesthetic reasons. Incoherent head poses in DeepFake videos are utilized in [11] to expose DeepFake videos. In [22], the idiosyncratic behavioral patterns of a particular individual are captured by the time series of facial landmarks extracted from real videos are used to spot DeepFake videos. The second category of DeepFake detection algorithms (*e.g.*, [12, 13]) use **signal-level** artifacts introduced during the synthesis process. Also, as synthesized faces are spliced into the original video frames, state-of-the-art DNN splicing detection methods, *e.g.*, [23, 24, 25, 26], can be applied. The third category of DeepFake detection methods (*e.g.*, [8, 9, 16, 18]) are **data-driven**, which directly employ various types of DNNs trained on real and DeepFake videos but capturing specific artifact.

2.1. Limitations

Albeit impressive progress has been made in the performance of detection of DeepFake videos, there are several concerns over the current detection methods that suggest caution.

[3] https://deepfakedetectionchallenge.ai.

Fig. 2. *Visual artifacts of DeepFake videos in existing datasets, including low-quality, visible splicing boundaries, color mismatch, visible parts of the original face, and inconsistent face orientations.*

Quality of DeepFake Datasets. The availability of large-scale datasets of DeepFake videos is an enabling factor in the development of DeepFake detection method. However, a closer look at the DeepFake videos in existing datasets reveals some stark contrasts in visual quality to the actual DeepFake videos circulated on the Internet. Several common visual artifacts that can be found in these datasets are highlighted in Fig.4, including low-quality synthesized faces, visible splicing boundaries, color mismatch, visible parts of the original

face, and inconsistent synthesized face orientations. These artifacts are likely the result of imperfect steps of the synthesis method and the lack of curating of the synthesized videos before included in the datasets. Moreover, DeepFake videos with such low visual qualities can hardly be convincing, and are unlikely to have real impact. Correspondingly, high detection performance on these dataset may not bear strong relevance when the detection methods are deployed *in the wild*. A related issue is that DeepFake detection methods trained using different DF datasets have trouble extending the performance to different datasets [27].

In a recent work [27], we present a new large-scale challenging DeepFake video dataset, *Celeb-DF*, which contains $5,639$ high-quality DeepFake videos of celebrities generated using improved synthesis process. We conduct a comprehensive evaluation of DeepFake detection methods and datasets to demonstrate the escalated level of challenges.

Performance Evaluation. Currently, the problem of detecting DeepFake videos is commonly formulated, solved, and evaluated as a binary classification problem, where each video is categorized as real or a DeepFake. Such dichotomy is easy to set up in controlled experiments, where we develop and test DeepFake detection algorithms using videos that are either pristine or made with DeepFake generation algorithms. However, the picture is murkier when the detection method is deployed in real world. For instance, videos can be fabricated or manipulated in ways other than DeepFakes, so not being detected as a DeepFake video does not necessarily suggest the video is a real one. Also, a DeepFake video may be subject to other types of manipulations and a single label may not comprehensively reflect such. Furthermore, in a video with multiple subjects' faces only one or a few are generated with DeepFake for a fraction of the frames. So the binary classification scheme needs to be extended to multi-class, multi-label, and local classification/detection to fully handle the complexities of real world media forgeries.

Explainability of Detection Results. Current DeepFake detection methods are usually designed to perform batch analysis over a large collection videos. However, when the detection methods are used in the field by journalists or law enforcement, we usually need only to analyze a small number of videos. Numerical score corresponding to the likelihood of a video being generated using a synthesis algorithm is not as useful to the practitioners if it is not corroborated with proper reasoning of the score. In such scenarios, it is very typical to request a justification for the numerical score for the analysis to be acceptable for publishing or used in court. However, many data-driven DF detection methods, especially those based on the use of deep neural networks, usually lack explainability due to the black box nature of the DNN models.

Temporal Aggregation. Most existing DeepFake detection methods are based on binary classification at the frame level, *i.e.*, determining the likelihood of an individual frame as real or of DeepFake. Although simple and straightforward, there are two issues of this methodology. First, the temporal consistency among frames are not explicitly considered, as (i) many DeepFake videos exhibit temporal artifacts and (ii) real or DeepFake frames tend to appear in continuous intervals. Second, it necessitates an extra step when video-level integrity score is needed: we have to aggregate the scores over individual frames to compute such a score.

Social Media Laundering. A large fraction of online videos are now spread through social networks, *e.g.*, FaceBook, Instagram, and Twitter. To save network bandwidth and also to protect the users' privacy, these videos are usually striped off meta-data, down-sized, and then heavy compressed before they are uploaded to the social platforms. These operations, commonly known as *social media laundering*, are detrimental to recover traces of underlying manipulation, and at the same time increase the false positive detections, *i.e.*, classifying a real video as a DeepFake. So far, most data-driven DeepFake detection methods that use signal level features are much affected by social media laundering. A practical measure to improve the robustness of DeepFake detection methods to social media laundering is to actively incorporate simulations of such effects in training data, and also enhance evaluation datasets to include performance on social media laundered videos, both real and synthesized.

3. FUTURE DIRECTIONS

Besides continuing improving to solve the aforementioned limitations, we also envision a few important directions of DeepFake detection methods that will receive more attention in the coming years.

Other Forms of DeepFakes. Although face swapping is currently the most widely known form of DeepFake videos, it is by no means the most effective. In particular, for the purpose of impersonating someone, face swapping DeepFake videos have several limitations. Psychological studies [citation] show that human face recognition largely relied on information gleaned from face shape and hairstyle. As such, to create convincing impersonating effect, the person whose face is to be replaced (the target) has to have similar face shape and hairstyle to the person whose face is used for swapping (the donor). Second, as the synthesized faces need to be spliced into the original video frame, the inconsistencies between the synthesized region and the rest of the original frame can be severe and difficult to conceal.

In these respects, the other two forms of DeepFake videos, namely, head puppetry and lip-syncing, are more effective and thus should become the focus of subsequent research in DeepFake detection. Methods studying whole face synthesis or reenactment have experienced fast development in recent years. Although there have not been as many easy-to-use and free open-source software tools generating these types of DeepFake videos as for the face-swapping videos, the continuing sophistication of the generation algorithms will change

Fig. 3. *Example frames from the Celeb-DF dataset. Left column is the frame of real videos and right five columns are corresponding DeepFake frames generated using different donor subject.*

the situation in the near future. Because the synthesized region is different from face swapping DeepFake videos (the whole face in the former and lip area in the latter), detection methods designed based on artifacts specific to face swapping are unlikely to be effective for these videos. Correspondingly, we should develop detection methods that are effective to these types of DeepFake videos.

Audio DeepFakes. AI-based impersonation are not limited to imagery, recent AI-synthesized content-generation are leading to the creation of highly realistic audios [28, 29]. Using synthesized audios of the impersonating target can significantly make the DeepFake videos more convincing and compounds its negative impact. As audio signals are 1D signals and have very different nature from images and videos, different methods need to be developed to specifically targeting such forgeries. This problem has drawn attention in the speech processing community recently with part of the most recent Global ASVspoofing Challenge[4] dedicated to AI-driven voice conversion detection, and a few dedicated methods for audio DeepFake detection, *e.g.*, [30], have also shown up recently. In the coming years, we expect more developments in these areas, in particular, those can leverage features in both visual and audio features of the fake videos.

Intent Inference. Even though the potential negative impacts of DeepFake videos are tremendous, in reality, the majority of DeepFake videos are not created not with a malicious intent. Many DeepFake videos currently circulated online are of a pranksome, humorous, or satirical nature. As such, it is important to expose the underlying intent of a DeepFake in the context of legal or journalistic investigation. Inferring intention may require more semantic and contextual understanding of the content, few forensic methods are designed to answer this question, but this is certainly a direction that future foren-

[4] https://www.asvspoof.org/.

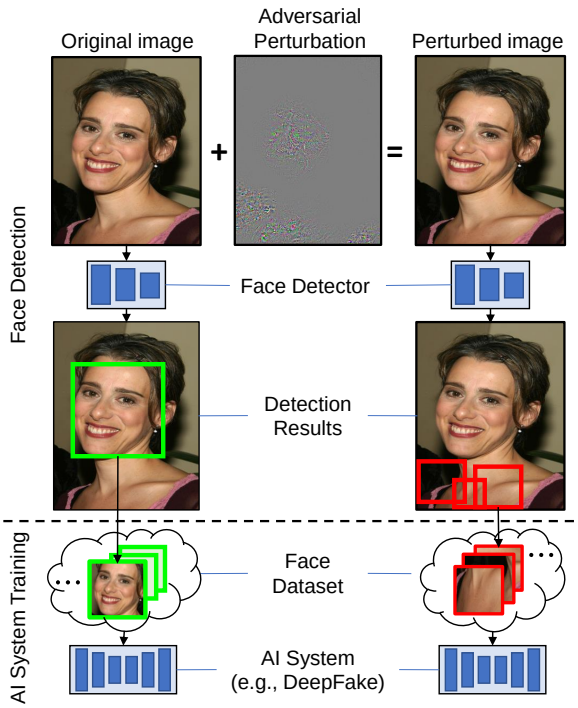

Fig. 4. *Overview of the proposed method of disrupting AI face synthesis. Our aim is to use the adversarial perturbations (amplified by* 30 *for better visualization) to distract DNN-based face detectors, such that the quality of the obtained face set as training data to the AI face synthesis is reduced.*

sic methods will focus on.

Anti-forensics. With the increasing effectiveness of DeepFake detection methods, we also anticipate developments of corresponding anti-forensic measures, which take advantage of the vulnerabilities of current DeepFake detection methods to conceal revealing traces of DeepFake videos. The data-driven deep neural network based DeepFake detection methods are particularly susceptible to anti-forensic attacks due to the known vulnerability of general deep neural network classification models. Anti-forensic measures can also be developed in the other aspect, to disguise a real video as a DeepFake video by adding simulated signal level features used by current detection algorithms, a situation we term as *fake DeepFake*. Further DeepFake detection methods must improve to handle such intentional and adversarial attacks.

Human Performance. Although the potential negative impacts of online DeepFake videos are widely recognized, currently there is a lack of formal and quantitative study of the perceptual and psychological factors underlying their deceptiveness. Interesting questions such as if there exist an *uncanny valley*[5] for DF videos, what is the *just noticeable difference* between high-quality DeepFake videos and real videos to human eyes, or what type/aspects of DeepFake videos are more effective in deceiving the viewers, have yet to be answered. To pursue these questions, it calls for close collaboration among researchers in digital media forensics and in perceptual and social psychology. There is no doubt that such studies are invaluable to research in detection techniques as well as a better understanding of the social impact that DeepFakes can cause.

Protection measures. However, given the speed and reach of the propagation of online media, even the currently best forensic techniques will largely operate in a postmortem fashion, applicable only after AI synthesized fake face images or videos emerge. We aim to develop *proactive* approaches to protect individuals from becoming the victims of such attacks, which complement to the forensic tools.

One such method we have recently studied [31] is to add specially designed patterns known as the *adversarial perturbations* that are imperceptible to human eyes but can result in detection failures. The rationale is as follows. High-quality AI face synthesis models need large number of, typically in the range of thousands, sometimes even millions, training face images collected using automatic face detection methods, *i.e.*, the *face sets*. Adversarial perturbations "pollute" a face set to have few actual faces and many non-faces with low or no utility as training data for AI face synthesis models, Fig. 4. The proposed adversarial perturbation generation method can be implemented as a service of photo/video sharing platforms before a user's personal images/videos are uploaded or as a standalone tool that the user can use, to process the images and videos before they are uploaded online.

4. CONCLUSION

We predict that several future technological developments will further improve the visual quality and generation efficiency of the fake videos. Firstly, one critical disadvantage of the current DeepFake generation methods are that they cannot produce good details such as skin and facial hairs. This is due to the loss of information in the encoding step of generation. However, this can be improved by incorporating GAN models[32] which have demonstrated performance in recovering facial details in recent works [33, 34]. Secondly, the synthesized videos can be more realistic if they are accompanied with realistic voices, which combines video and audio synthesis together in one tool.

In the face of this, the overall running efficiency, detection accuracy, and more importantly, false positive rate, have to be improved for wide practical adoption. The detection methods also need to be more robust to real-life post-processing steps, social media laundering, and counter-forensic technologies. There is a perpetual competition of technology, know-hows, and skills between the forgery makers and digital media forensic researchers. The future will reckon the predictions we make in this work.

[5]The uncanny valley in ths context refers to the phenomenon whereby a DeepFake generated face bearing a near-identical resemblance to a human being arouses a sense of unease or revulsion in the viewers.

5. REFERENCES

[1] Hany Farid, *Digital Image Forensics*, MIT Press, 2012.

[2] "FakeApp," `https://www.malavida.com/en/soft/fakeapp/`, Acessed Nov 4, 2019.

[3] "DFaker github," `https://github.com/dfaker/df`, Accessed Nov 4, 2019.

[4] "faceswap-GAN github," `https://github.com/shaoanlu/faceswap-GAN`, Accessed Nov 4, 2019.

[5] "faceswap github," `https://github.com/deepfakes/faceswap`, Accessed Nov 4, 2019.

[6] "DeepFaceLab github," `https://github.com/iperov/DeepFaceLab`, Accessed Nov 4, 2019.

[7] Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21*.

[8] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[9] David Güera and Edward J Delp, "Deepfake video detection using recurrent neural networks," in *AVSS*, 2018.

[10] Yuezun Li, Ming-Ching Chang, and Siwei Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[11] Xin Yang, Yuezun Li, and Siwei Lyu, "Exposing deep fakes using inconsistent head poses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[12] Falko Matern, Christian Riess, and Marc Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019.

[13] Yuezun Li and Siwei Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[14] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, "Recurrent-convolution approach to deepfake detection-state-of-art results on faceforensics++," *arXiv preprint arXiv:1905.00582*, 2019.

[15] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *ICCV*, 2019.

[16] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[17] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2019.

[18] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen, "Use of a capsule network to detect fake images and videos," *arXiv preprint arXiv:1910.12467*, 2019.

[19] Pavel Korshunov and Sébastien Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[20] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler, "Deepfakes detection dataset by google & jigsaw," .

[21] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[22] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li, "Protecting world leaders against deep fakes," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[23] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis, "Two-stream neural networks for tampered face detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[24] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis, "Learning rich features for image manipulation detection," in *CVPR*, 2018.

[25] Yaqi Liu, Qingxiao Guan, Xianfeng Zhao, and Yun Cao, "Image forgery localization based on multi-scale convolutional neural networks," in *ACM Workshop on Information Hiding and Multimedia Security (IHMMSec)*, 2018.

[26] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury, "Hybrid lstm and encoder-decoder architecture for detection of image forgeries," *IEEE Transactions on Image Processing (TIP)*, 2019.

[27] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020.

[28] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.

[29] Yu Gu and Yongguo Kang, "Multi-task WaveNet: A multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions," in *Interspeech*, Hyderabad, India, 2018.

[30] Ehab AlBadawy, Siwei Lyu, and Hany Farid, "Detecting ai-synthesized speech using bispectral analysis," in *Workshop on Media Forensics (in conjunction with CVPR)*, Long Beach, CA, United States, 2019.

[31] Yuezun Li, Xin Yang, Baoyuan Wu, and Siwei Lyu, "Hiding faces in plain sight: Disrupting ai face synthesis with adversarial perturbations," 2019.

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[33] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *The International Conference on Learning Representations (ICLR)*, 2017.

[34] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.