An Overview of Deepfake Technologies: from Creation to Detection in Forensics

Luca Guarnera[1] and Sebastiano Battiato[1]

University of Catania, Italy
luca.guarnera@unict.it, battiato@dmi.unict.it

Abstract: Advancements in Artificial Intelligence (AI) techniques have given rise to significant challenges in the field of Multimedia Forensics, particularly with the emergence of the Deepfake phenomenon. Deepfakes are images, video and audio generated or altered by powerful generative models such as Generative Adversarial Networks (GANs) [5] and Diffusion Models (DMs) [12]. While GANs have long been recognized for their ability to generate high-quality images, DMs offer distinct advantages, providing better control over the generative process and the ability to create images with a wide range of styles and content [2]. In fact, DMs have shown the potential to produce even more realistic images than GANs. The AI-generated contents span diverse domains, including films, photography, video games, and virtual reality productions. A major concern of the Deepfake phenomenon is the application on important people such as politicians and celebrities to spread misinformation. However, the most alarming aspect is the misuse of GANs and DMs to create pornographic Deepfakes, posing a serious security threat. Notably, a staggering 96% of Deepfakes available on the internet fall into this pornographic category. The malicious use of Deepfakes extends to issues such as misinformation, cyberbullying, and privacy violation. In addition, Deepfakes have been applied in the fields of art and entertainment, sparking ethical discussions about the limits of creativity and authenticity. To counteract the illicit use of this powerful technology, novel forensic detection techniques are required to identify whether multimedia data has been manipulated or altered using GANs and DMs. Regarding image deepfake detection methods in the state of the art, the primary focus lies in binary detection, distinguishing between Real and AI-generated images [14, 16]. Notably, some methods in the state of the art have already demonstrated the ability to effectively differentiate between various GAN architectures [4, 7, 6, 15] and several DM engines [13, 1, 9]. These researches showed that generative models leave unique fingerprints in the generated multimedia data, which can be used not only to identify Deepfakes, but also to recognize the specific architecture used during the creation process [11]. This can be extremely important in forensics in order to reconstruct the history of the multimedia data under analysis (forensic ballistics) [8]. In order to create increasingly sophisticated deepfakes detection solutions, several challenges have been proposed by the scientific community such as the Deepfake Detection Challenge (DFDC) [3] and the Face Deepfake Detection Challenge [10]. The latter has also launched a new challenge among researchers in the field: reconstructing the orginal image from deepfakes; a task that can be extremely important in forensics.

References

[1] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[3] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[4] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, 7(8):128, 2021.

[5] 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*, pages 2672–2680, 2014.

[6] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667, 2020.

[7] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Fighting deepfake by exposing the convolutional traces on images. *IEEE Access*, 8:165085–165098, 2020.

[8] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake style transfer mixture: A first forensic ballistics study on synthetic images. In *International Conference on Image Analysis and Processing*, pages 151–163. Springer, 2022.

[9] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. *arXiv preprint arXiv:2303.00608*, 2023.

[10] Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh MQ Bui, Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, et al. The face deepfake detection challenge. *Journal of Imaging*, 8(10):263, 2022.

[11] Luca Guarnera, Oliver Giudice, Matthias Nießner, and Sebastiano Battiato. On the exploitation of deepfake model recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 61–70, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[13] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint arXiv:2210.06998*, 2022.

[14] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019.

[15] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[16] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.