Seeing is No Longer Believing: How Deepfakes May Shape the Future of Identity Credibility in Media

Raphael Antonius Frick
Fraunhofer SIT | ATHENE Center
Darmstadt, Hesse, Germany
raphael.frick@sit.fraunhofer.de

Niklas Bunzel
Fraunhofer SIT | ATHENE Center
Darmstadt, Hesse, Germany
niklas.bunzel@sit.fraunhofer.de

Lukas Graner
Fraunhofer SIT | ATHENE Center
Darmstadt, Hesse, Germany
lukas.graner@sit.fraunhofer.de

Abstract

Deep learning techniques and frameworks for manipulating or generating identities in multimedia, particularly visual data like images and videos, have advanced to a point where even individuals without specialized technical knowledge can create “deepfakes” and use them in real-world situations. This paper examines the potential threats posed by these advancements, offering an overview of manipulation techniques and state-of-the-art detection methods to counteract them. Additionally, it discusses the requirements, trade-offs, and limitations associated with these specific methods, providing a thorough understanding of the current landscape in deepfake technology and its detection, as well as possible trends for the future.

CCS Concepts

• **Security and privacy** → *Social network security and privacy*; • **Information systems** → *Multimedia information systems.*

Keywords

Deepfakes, Threat Assessment, Multimedia Forensics, Adaptive Attacks

ACM Reference Format:

Raphael Antonius Frick, Niklas Bunzel, and Lukas Graner. 2025. Seeing is No Longer Believing: How Deepfakes May Shape the Future of Identity Credibility in Media. In *Proceedings of the 2025 Deepfake, Deception, and Disinformation Security Workshop (3D-Sec ’25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3733813.3764371

1 Introduction

With the recent developments in deep learning technology, methods to create or manipulate the identity of a person shown in visual data (i.e., image or video data) continue to increase in quality. Higher resolutions, faster computation times and less data needed to target a specific identity are some of the advancements observed in state-of-the-art manipulations. Such AI-based methods, like face swapping and voice conversion, have been used to create videos featuring actors in new movies without requiring their physical presence on set [41]. However, because these techniques have also been used to create illicit pornography, identity fraud and disinformation on the past, they can have a major impact on society. Since images and videos are frequently shared on the internet and most profoundly on social media platforms, it is often easy to obtain the necessary reference material for a particular identity. Hence, targeted identity manipulation becomes feasible for an attacker and several attacks have already been reported in the media [11, 34, 42], contributing to society’s mistrust of visual media. On the other hand, a vast amount of deep learning based detection methods are presented as countermeasures in scientific literature [1, 14, 47]. However, the suitability for real-world application remains questionable on most methods because of concerns on their generalizability, explainability and the easy circumvention by adaptive attackers. Therefore, this paper reviews the current state-of-the-art in deep learning-based identity manipulation and detection methods. It highlights leading techniques for artificially creating and manipulating identities in visual media, describes key detection approaches, and discusses current and possible emerging threats posed by deepfakes.

3D-Sec ’25, Taipei, Taiwan

ACM ISBN 979-8-4007-1902-8/25/10
https://doi.org/10.1145/3733813.3764371

2 Deepfakes

The term “deepfake” was first introduced in 2017 to describe a type of facial forgery where a person’s facial texture was replaced with that of another individual using deep learning. Over time, the term has broadened to encompass a wide range of generation and manipulation techniques, including those that do not use deep learning. In parallel, the term AI-generated content (AIGC) is now widely used. Under this broad framing, all AIGC can be viewed as deepfakes, while some deepfakes are not AIGC because they are created without AI. In this section, however, we adopt an updated, narrower definition and review image- and video-focused methods. Hereby, we define deepfakes as media artificially generated or altered with deep learning in ways that affect the depicted identities.

Image Generation: Modern methods generate face images either randomly or with specific guidance. GANs like StyleGAN [28] sample from a latent space to create realistic but mostly uncontrolled faces. Diffusion models [20, 36] use text prompts or reference images for more directed synthesis. Fine-tuning (e.g., LoRA [27]) and techniques like ControlNet [50] further enable control over identity, style, or image structure.

Image Editing: GANs can edit facial attributes (e.g., age, hair color) by mapping images to latent space and modifying its vectors [24, 32], though this may blur details or unintentionally change multiple traits. Diffusion models improve control, allowing precise attribute changes using inversion [5] or inpainting. Neural inpainting fills selected image regions with new content, now producing high-quality results using deep learning. Diffusion models can condition this process on text prompts for targeted edits [16].

Video Generation: This technique involves generating video content from a text description. AI models trained on large datasets

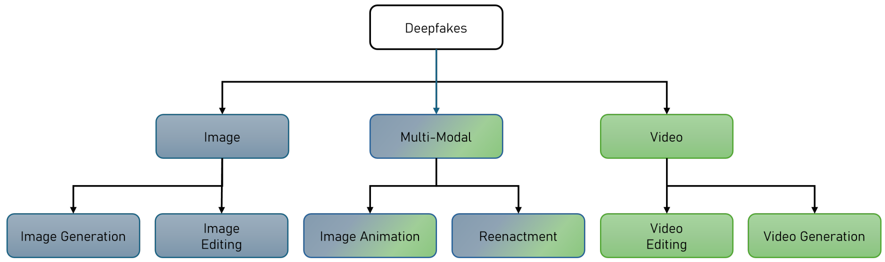

Figure 1: Overview on deepfake techniques enabling the synthesis and manipulation of medial identities.

of video and text can create short video clips that match the description provided [29, 39]. While still in its developmental stages, this technology represents a future direction for deepfakes, where entirely new video content can be created from scratch based on textual input.

Video Editing: Face swapping replaces a person's face with another's in images or videos, using autoencoders to map facial features and textures for realistic results [13, 31]. This can be done in real-time or on pre-recorded content. Video inpainting applies deep learning to modify specific regions in video, enabling object removal or addition with high quality [51, 52].

Multi-Modal: Facial reenactment maps one person's expressions onto another's face in videos or images, enabling realistic expression and mouth movement transfer, with some methods supporting real-time synthesis [22, 46]. Lip-syncing changes a person's lip movements to match different audio, often combined with voice conversion or text-to-speech to create convincing synthetic speech [4, 8, 23, 35]. Unlike reenactment, lip-syncing does not support real-time processing. Full-body deepfakes manipulate not only faces but also body movements and gestures, using motion capture and deep learning to fabricate actions in videos [25, 44]. Image animation uses diffusion models to turn still images into videos, with possible control via text prompts for specific animation outcomes [9, 45].

3 Threat Landscape

Deepfake methods, though originally developed for entertainment, are nowadays widely misused. Face swapping quickly became a tool for non-consensual pornography and defamation, affecting both celebrities and the public [7]. Real-time face swaps are now used increasingly in fraud, such as CEO scams [6]. Image synthesis, face swapping, and reenactment also enable the spread of fake news [11, 42].With the advancement of diffusion-models, image generation has shifted to video generation. However, these large-scale diffusion models also raise data privacy issues due to being dependent on large amounts of data during training.

4 Deepfake Detection & Regulation

Detection methods are either model-based (using handcrafted features) or data-driven (deep neural networks) [38, 47]. For video face swaps, model-based methods analyze behavioral, signal-based, or geometric features [3, 40, 43]. Data-driven models generally achieve higher accuracy, while model-based methods offer more interpretability.

Detection is challenging due to limited generalizability and robustness, especially after post-processing [21], though training with augmentations helps [18]. Methods must be tailored to specific forgery types, hence, no universal solution exists. Most focus on face swaps and diffusion-based fakes, while few target lip sync [37] or facial reenactment [30], making these harder to detect.

The EU AI Act and the US Executive Order emphasize transparency, safety, and anti-misuse for AI-generated content, requiring clear disclosure and content authentication. Stricter rules apply for high-risk uses, with penalties for non-compliance [15, 26]. These initiatives aim to ensure trustworthy and responsible AI deployment. Open-source deepfake tools make regulation difficult, as protections can be bypassed. Effective detection systems, e.g., integrated into content platforms, will therefore still play huge importance in the future.

5 Adversarial Attacks on Deepfake Detection

Adversarial attacks threaten deepfake detectors by subtly modifying inputs to fool the system. White-box attacks, which require access to model gradients, are unrealistic in real-world settings, while black-box attacks rely on model outputs but need many queries [10, 12], limiting real-time use. Another key risk is transferability: adversarial examples made for one model can often fool others [17], especially when generated with an ensemble of models [19]. Universal perturbations [33], which generalize across images, further increase the threat to detection systems, though their efficacy against deepfake detection remains to be analyzed.

6 Future Developments

In the following, we will present various experiments, aiming to identify potential emerging trends of deepfakes. For the experiments, we took advantage of a system featuring an Intel i7-8700K CPU (3.7 GHz), NVIDIA GeForce 2080 Ti GPU (11 GB RAM), and 32 GB RAM.

Obstruction-Aware Real-Time Face Swapping: Current real-time face swapping solutions struggle with obstructing hands and objects. To enhance occlusion robustness, we developed a SegFormer-based [49] segmentation model, trained on synthetic images combining hand overlays from the 11k Hands dataset [2] with FaceSynthetics faces [48]. We used a mask strategy that defines the visible face region for seamless integration into DeepFaceLive (see Figure 2). These were used to train an *nvidia/mit-b2* and *nvidia/mit-b0* variant of SegFormer. Although both models achieve high accuracies on validation sets and are highly efficient (B0, 0.9996 AUC,

Figure 2: Obstructed deepfake before (left) and after proposed face segmentation (right).

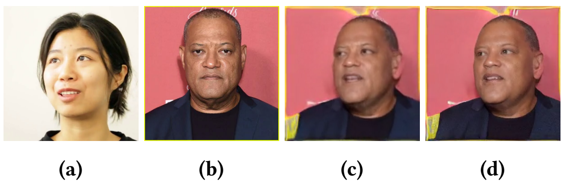

Figure 3: Influence of various face reconstruction models on the visual quality: (a) Input Driving Sample, (b) Target Sample, (c) Output LIA, (d) Output LIA + RestoreFormer++

	Total Run Time			Estimated Processing Time	
Approach	t_{1f}	t_{60s}	t_{120s}	1 Frame	1 Second (25 fps)
FOMM	16 s	93 s	168 s	0.051 s	1.275 s
SAFA	35 s	370 s	724 s	0.227 s	5.663 s
LivePortrait	24 s	194 s	372 s	0.115 s	2.867 s
TPSM	15 s	181 s	354 s	0.112 s	2.796 s

Table 1: Total run time for processing videos of different lengths, and estimated processing time per frame and per second (25 fps) for each approach.

42ms, 24fps; B2, 0.9998 AUC, 58ms, 17fps), they still require further improvement be robust against fast movements and it resulting motion-blur.

From Live Face Swapping to Live Facial Reenactment: Studies (redacted for review) have shown that facial reenactment methods are particularly effective at spoofing automatic verification systems. However, most state-of-the-art reenactment methods are either not real-time capable or, if they are, produce lower visual quality. Our benchmarks (Table 1) confirm that only FOMM achieves near real-time speeds (20 fps), while others are slower. Real-time variants, such as Live Portrait Monitor[1], exist but often with quality trade-offs.

We demonstrate that visual quality can be significantly improved using face reconstruction methods that are real-time enabled after optimization. Integrating RestoreFormer++ into DeepFaceLive, which is based on LIA [46], and optimizing with backends like TensorRT or TorchInductor, the enhanced synthesis can be performed in real-time with up to 14 fps while providing high-quality outputs. Both examples show that face swapping and facial reenactment can become even more convincing and artifact-free in real-time as technology improves. In parallel, video-based diffusion models are expected to play a growing role in offline identity spoofing. Although currently slow and costly, these models will likely become more efficient and able to handle longer videos as hardware advances.

[1]Live Portrait Monitor

7 Conclusion

The rapid progress in deep learning has made it increasingly easy to create and manipulate visual identities in images and videos. State-of-the-art deepfake methods now deliver higher quality, require less data, and can operate in real-time, opening up new opportunities in creative industries but also enabling large-scale misuse. As deepfakes become more convincing and accessible, they pose significant risks, ranging from identity fraud and non-consensual content to disinformation and diminishing of trust in visual media.

At the same time, detection and regulation face growing challenges. While detection methods have advanced, real-world applicability is hindered by issues with generalizability, robustness, and adaptability against sophisticated attacks. Regulatory measures such as the EU AI Act and US Executive Orders are important steps, but enforcement remains difficult due to the open-source nature of many deepfake tools.

Our experiments demonstrate that both the creation and detection of deepfakes are evolving rapidly. Techniques like obstruction-aware real-time face swapping and optimized face reconstruction significantly improve quality and usability, while adversarial attacks highlight persistent vulnerabilities in detection systems.

Ultimately, as the boundaries between authentic and synthetic media blur, society must adapt both technologically and legally. Ensuring identity credibility in the age of deepfakes will require continuous innovation in detection, robust regulatory frameworks, and widespread public awareness.

Acknowledgements

This research work has been funded by BMBF and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–7.
[2] Mahmoud Afifi. 2019. 11K Hands: Gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools Appl.* 78, 15 (aug 2019), 20835–20854.
[3] Shruti Agarwal, Tarek El-Gaaly, Hany Farid, and Ser-Nam Lim. 2020. Detecting Deep-Fake Videos from Appearance and Behavior. arXiv:2004.14491 [cs.CV]
[4] Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice Conversion With Just Nearest Neighbors. In *Interspeech*.
[5] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. 2024. Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Semantic Directions. arXiv:2403.17064 [cs.CV]
[6] BBC. 2023. MrBeast and BBC stars used in deepfake scam videos. https://www.bbc.com/news/technology-66993651. Visited: 10.10.2024.
[7] BBC. 2024. Inside the deepfake porn crisis engulfing Korean schools. https://www.bbc.com/news/articles/cpdlpj9zn9go. Visited: 10.10.2024.
[8] James Betker. 2023. Better speech synthesis through scaling. arXiv:2305.07243 [cs.SD]
[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127 [cs.CV]
[10] Niklas Bunzel and Lukas Graner. 2023. A Concise Analysis of Pasting Attacks and their Impact on Image Classification. In *53rd Annual IEEE/IFIP International*

Conference on Dependable Systems and Networks, DSN 2023 - Workshops, Porto, Portugal, June 27-30, 2023. IEEE, 136–140.

[11] Ella Cao and Eduardo Baptista. [n. d.]. *'Deepfake' scam in China fans worries over AI-driven fraud*. https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/

[12] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, Bhavani Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha (Eds.). ACM, 15–26.

[13] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. 2024. SimSwap++: Towards Faster and High-Quality Identity Swapping. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 1 (2024), 576–592.

[14] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*. Springer, 219–229.

[15] European Commission. 2024. Artificial Intelligence Act. https://artificialintelligenceact.eu/the-act/. Visited: 10.10.2024.

[16] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. 2024. LatentPaint: Image Inpainting in Latent Space with Diffusion Models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4322–4331.

[17] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. 2019. Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 321–338.

[18] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. 2020. The DeepFake Detection Challenge Dataset. *ArXiv* abs/2006.07397 (2020).

[19] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting Adversarial Attacks With Momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 9185–9193.

[20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206 [cs.CV]

[21] Raphael Antonius Frick and Martin Steinebach. 2024. One Detector to Rule Them All? On the Robustness and Generalizability of Current State-of-the-Art Deepfake Detection Methods. *Electronic Imaging* (2024).

[22] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168* (2024).

[23] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[24] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Trans. Image Process.* 28, 11 (2019), 5464–5478.

[25] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. 2023. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[26] THE WHITE HOUSE. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Visited: 10.10.2024.

[27] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

[28] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 4401–4410.

[29] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).

[30] Prabhat Kumar, Mayank Vatsa, and Richa Singh. 2020. Detecting Face2Face Facial Reenactment in Videos. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2578–2586.

[31] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. 2023. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recogn.* 141, C (Sept. 2023), 12 pages.

[32] Yunfan Liu, Qi Li, Qiyao Deng, Zhenan Sun, and Ming-Hsuan Yang. 2023. GAN-Based Facial Attribute Manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 12 (2023), 14590–14610.

[33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal Adversarial Perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 86–94.

[34] Federal Bureau of Investigation. [n. d.]. *Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. https://www.ic3.gov/Media/Y2023/PSA230605

[35] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 484–492.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[37] Sahibzada Adil Shahzad, Ammarah Hashmi, Sarwar Khan, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. 2022. Lip Sync Matters: A Novel Multimodal Forgery Detector. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 1885–1892.

[38] Kaede Shiohara and T. Yamasaki. 2022. Detecting Deepfakes with Self-Blended Images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 18699–18708.

[39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[40] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. 2021. Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features. arXiv:2104.04480 [cs.CV]

[41] Jonathan Swartz and Sean Walker. 2022. Deep Learned Face Swapping in Feature Film Production. In *ACM SIGGRAPH 2022 Talks* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 2 pages.

[42] The Telegraph. [n. d.]. *Deepfake video of Volodymyr Zelensky surrendering surfaces on social media*. https://www.youtube.com/watch?v=X17yrEV5sl4

[43] Gao-Jian Wang, Wei Li, Qian Jiang, Xin Jin, and Xiao-Hui Cui. 2021. Using Grayscale Frequency Statistic to Detect Manipulated Faces in Wavelet-Domain. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2986–2993.

[44] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. 2024. VividPose: Advancing Stable Video Diffusion for Realistic Human Image Animation. *arXiv preprint arXiv:2405.18156v1* (2024).

[45] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. 2024. MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation. arXiv:2401.04468 [cs.CV]

[46] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. 2024. LIA: Latent Image Animator. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–16.

[47] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. 2023. AltFreezing for More General Video Face Forgery Detection. arXiv:2307.08317 [cs.CV]

[48] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. 2021. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3681–3691. GIT repo: https://github.com/microsoft/FaceSynthetics.

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *CoRR* abs/2105.15203 (2021). arXiv:2105.15203 GIT repo: https://github.com/NVlabs/SegFormer.

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

[51] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. 2023. AVID: Any-Length Video Inpainting with Diffusion Model. *arXiv preprint arXiv:2312.03816* (2023).

[52] Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. 2023. ProPainter: Improving Propagation and Transformer for Video Inpainting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*.