Generative Artificial Intelligence in Deepfake Face Synthesis

Tianyi Wang[1], Cyril Leung[2]

[1]Hong Kong University, China
The University of British Columbia, Canada

Abstract

Generative artificial intelligence (AI) models have been prevalently leveraged in a plethora of downstream tasks because of the recent popular large language models (LLMs). While LLMs greatly step forward the research progress in analyzing languages, deep generative AI models are also inspirative in the remaining domains, especially computer vision. Deepfake, a technique that manipulates facial content in a target image with the guiding information (identity, expression, pose, and attributes) from a source image, has brought both industrial opportunities and privacy threats to human lives by deriving hyper-realistic synthetic media. In this survey, we perform a comprehensive study on the evolution of deep generative AI models in Deepfake face synthesis research and provide informative discussions on the pros and cons of the Deepfake technique, concluding with the potential future research directions in the domain.

Keyword: Deepfake, generative artificial intelligence model, face synthesis.

I. Introduction

Generative artificial intelligence (AI), a concept that depicts the capability of artificial intelligence in generating new media contents such as image, audio, and text using generative models, has experienced magnificent development up to the current stage. Generative AI models are designed to learn feature patterns from existing data corpora and produce synthetic content following similar

patterns. Throughout the history of generative AI models, various novel concepts and tasks have successively appeared. Among the new terms that are derived in the background of generative AI, besides the large language models (LLMs) that have caught significant public attention recently, Deepfake[1], a facial image synthesis technique using deep generative models, has caused massive attraction to the public due to its impacts in both positive and negative aspects.

Relying on the deep generative AI models, Deepfake is able to perform image manipulation by modifying the identities, expressions, poses, and attributes of the facial content. Consequently, along with the evolution of generative AI models from auto-encoder [1] to stable diffusion [2], the Deepfake facial manipulation technique has brought public benefits and privacy risks to human lives. For instance, to avoid reshooting or removing the relevant episodes, tainted celebrities banned due to controversial behaviors are face-swapped in the TV productions they performed before being released to the public[2]. Additionally, Marvel movie lovers may swap their faces onto movie clips and perform as their favorite superheroes and superheroines [3]. On the other hand, victim groups attacked by misusing Deepfake include politicians[3,4], celebrities[5,6], and even every human being[7] due to easy access to the image forgery applications and tools. To make wise utilization of Deepfake to bring convenience to our lives and prevent it from causing privacy issues simultaneously, it is necessary to truly get to know the technique itself.

In this paper, we conduct a comprehensive study based on the evolution of generative AI in the domain of Deepfake face synthesis. First, we present a detailed review of the evolution history in regard to the generative AI models for Deepfake face synthesis. In particular, the two representative sub-tasks, face swapping and face reenactment (Fig. 1), are each detailedly summarized via walking through the evolution experience of the auto-encoders, generative adversarial networks, and diffusion models.

[1] https://github.com/deepfakes
[2] http://tinyurl.com/7f5rwn67
[3] https://www.youtube.com/watch?v=cQ54GDm1eL0
[4] https://tinyurl.com/bdzyuam9
[5] https://www.bbc.com/news/technology-42905185
[6] https://www.bbc.com/news/technology-42912529
[7] https://tinyurl.com/yh82nz45

Then, we demonstrate the potential opportunities and risks by describing the benefits brought by Deepfake to society and the available countermoves regarding misusing the technique, respectively. In the end, we brief the future directions of generative AI models in Deepfake-related research with a conclusion.

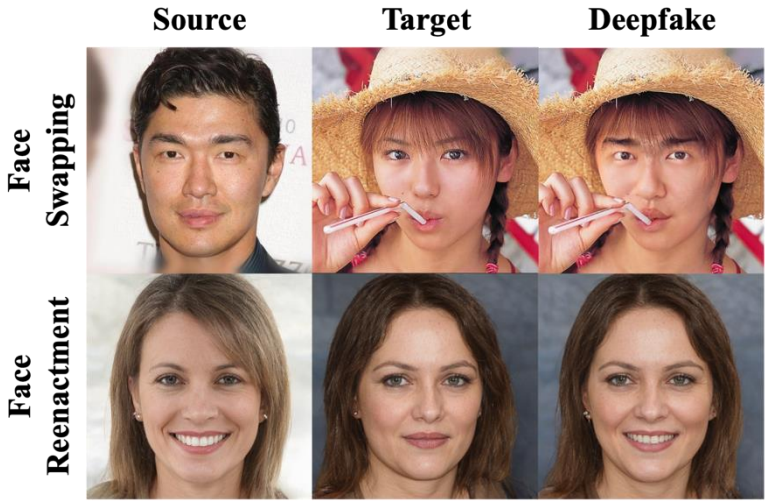


Fig. 1. Deepfake visualization illustrations by face swapping [4] in row 1 and face reenactment [5] in row 2.

II. Evolution of Deepfake Face Synthesis

In 2017, the Reddit user ‘deepfakes’ announced that, with the help of deep neural networks, he could automatically execute face identity swapping from a source person onto a target one while maintaining all other contents of the target image unchanged. Since then, the term ‘Deepfake’ has become popular and experienced explosive improvements in various domains. During the evolution of Deepfake, two types of face synthesis tasks with different focuses are researched individually, namely, *face swapping* and *face re-enactment*.

Face Swapping

Face swapping refers to the task that transfers a desired identity from a source face to a target facial image and preserves the rest image content including expression, movement, and background scene. Ever since the first occurrence of Deepfake face swapping was released in 2017, the early follow-up research mainly focuses on subject-specific approaches, which swap fixed pairs of facial identities that the models have seen during training using an auto-encoder.

In specific, for a pair of facial identities, a shared encoder extracts identity-independent features from either identity to preserve the unchanged information, and two unique decoders each take the work of synthesizing faces with the desired facial identity. During the training phase, the faces of each identity are first encoded and then reconstructed identically via the corresponding decoder. This pipeline enforces the decoders' ability to reconstruct the wanted facial identities. Consequently, the facial identity of the input face can be favorably modified to the other identity in the inference phase.

Later approaches mainly boost the face swapping performance by calibrating the fine-grained feature extraction and refinement pipelines. In particular, Korshunova et al. [6] proposed the integral of content, style, and light loss functions that collaborate with convolutional neural networks (CNNs) as the core architecture within the auto-encoder for producing highly photorealistic results. Nirkin et al. [7] employed 3D facial features for image segmentation and demonstrated the high efficiency and low computation cost with merely fully connected layers. DeepFaceLab [8] provides two canonical types of facial landmark extraction algorithms [9][10] to resolve the unsatisfactory performance caused by unstable face alignment protocols. Li et al. [11] attempted additional restrictions on low resolution, color mismatch, inaccurate face masks, and temporal flickering, which helps construct one of the most challenging high-quality Deepfake detection datasets, Celeb-DF.

To derive synthetic images with reasonable fidelity, researchers introduced generative adversarial network (GAN) [12] and adjusted the face swapping algorithms in the adversarial learning pipelines. Zhu et al. [13] aimed to improve the consistency in illuminations and skin colors, and they introduced a mix-and-segment discriminator (MSD) by randomly mixing output image patches for consistency verification in adversarial learning. Nirkin, Keller, and Hassner [14] utilized a multi-scale discriminator consisting of multiple discriminators for different levels of image resolution and achieved high-resolution face swapping results. Then, considering the high computational cost of training a new model for every pair of faces, the concept of identity-

agnostic face swapping has been raised. Specifically, by introducing a well-trained identity encoder, the generative model is able to extract identity information from arbitrary source images and perform face swapping on the target. FaceShifter [15] adopts the well-trained face recognition model, ArcFace [16], as the identity encoder following the algorithms of adaptive instance normalization (AdaIN) [17] and spatially-adaptive normalization (SPADE) [18] for identity feature injection, and accomplishes high-quality face swapping in challenging cases with occlusions by conducting a novel heuristic error acknowledging refinement network (HEAR-Net). Chen et al. [19] designed SimSwap with a novel ID injection module (IIM) that is based mainly on ArcFace, AdaIN, and a weak feature matching loss that preserves facial attributes to establish a balance between identity and attributes. The ID injection concept has since become commonplace and has been frequently adopted in subsequent face swapping techniques.

Based on the fixed ID injection protocol, the demand for identity consistency is gradually satisfied by the subsequent studies. Zhu et al. introduced MegaFS [20] with a face transfer module (FTM) and conducted stable control toward identity-related facial attributes for high-resolution face swapping up to 1,024 utilizing StyleGAN2 [21]. Gao et al. [22] proposed InfoSwap to optimize the information bottleneck trade-off to extract the most expressive identity information from the pre-trained face recognition model, deriving superior identity-consistent performance in swapped faces. Wang et al. [23] devised HifiFace with geometric supervision by 3D morphable models (3DMM) to preserve the face shape of the source facial identities. By leveraging a style-based generator, StyleSwap [24] is able to maintain high output fidelity and optimize identity consistency concurrently. RAFSwap [25] raises the idea of region-aware face swapping and adopts transformer [26] for identity-relevant local and global feature interactions. This idea has achieved state-of-the-art identity consistency performance at a time. With the help of supervised contrastive learning, Smooth-Swap [27] achieves promising face swapping results in regard to identity preservation. Zeng et al. [28] developed FlowFace based on the pre-trained face masked auto-encoder [29] to capture facial appearances and identity information, and better

identity preservation effects are obtained with correctly reconstructed face shapes. Meanwhile, several studies aim to reduce the computational cost in various aspects. Shu et al. [30] attempted a few-shot head swapper, HeSer, and achieved reasonable in-the-wild performance while largely reducing the demand for data corpora. Xu et al. [31] invented a lightweight identity-aware dynamic network (IDN) that efficiently executes identity-agnostic face swapping by dynamically updating the model parameters. This method is capable of real-time face swapping on mobile phones with a significantly small number of parameters. Wang et al. [4] proposed AP-Swap that achieves the uttermost facial attribute preservation performance by completely introducing facial landmarks to guide target facial attribute reconstruction. This utilization of facial landmarks concurrently gains huge efficiency improvements.

Recently, the mushroomed research progress in denoising diffusion probabilistic models (DDPMs) [32][33] has caught the attention of Deepfake researchers, and multiple face swapping algorithms are designed accordingly. With simple conditional inputs, conditional diffusion models (CDMs) can generate images with high degrees of semantic consistency. Although abundant applications [34][35][36][37] are released for various image generation tasks, the research on face swapping using CDMs has been just initialized. Kim et al. [38] proposed a dual-condition face generator (DCFace) to reconstruct facial images with desired identities and styles provided by the source and target images. To ensure stable feature extraction from source images, a patch-wise style extractor is designed to disable the identity information in target images. Zhao et al. [39] trained a DDPM conditioned in an inpainting task on the identity feature and facial landmark for face swapping (DiffSwap), collaborating with a midpoint estimation method for identity feature learning at a low cost.

Face Reenactment

As opposed to face swapping, face reenactment refers to a task in facial synthesis that involves transferring facial attributes from a source face to a target one, while preserving the appearance and the identity of the target face. In other words, the goal for content modifying and maintaining

on the target face is completely opposite in regard to face swapping. In the early stages, the face reenactment work is mainly exploited on fixed facial identities in a one-to-one manner with the help of GAN architectures. In particular, Xu et al. [40] leveraged CycleGAN [41] and PatchGAN [42] to transfer head poses and facial expressions, devoting a preliminary effort to the face reenactment task. Then, ReenactGAN [43] advances the research stage to the many-to-one manner face reenactment. To avoid potential structural artifacts, the authors mapped the source faces onto a boundary latent space while collaborating with the transformer before transferring the facial attributes to the target images. A popular method, Neural Textures [44], being adopted to construct one of the most popular Deepfake datasets, FaceForensics++ (FF++) [45], combines the traditional graphics pipeline with learnable components to interpret high-dimensional feature maps for photo-realistic re-rendering even with source images that have noisy and incomplete surface geometries.

Later, the goal shifts to the many-to-many face reenactment manner, and researchers are interested in splitting facial identities and attributes such that the attributes can be purely and correctly transferred even for unseen facial identities. Specifically, to resolve the identity preservation issue for unseen identities, Ha et al. [46] heavily adopted the core idea of the transformer for the source image, target image, and target facial landmarks. Since then, adopting facial landmarks in the face reenactment pipeline guarantees precision at some levels [47][48][49]. By combining appearance-based and warping-based methods, FLNet [50] generates faithful face reenactment outputs by simply adopting only the source faces and facial landmark differences between source and target. CrossID-GAN [51] performs multi-ID face reenactment by gathering identity-invariant motion patterns from facial landmarks. Concurrently, more complex algorithms have occurred to analyze deeper image features. Zeng et al. [52] introduced a novel self-supervised hybrid model, DAE-GAN, that learns to reenact faces naturally from large amounts of unlabeled videos that provide hints for identity and pose features disentanglement. Besides, Burkov et al. [53] conducted a latent pose representation

encoder and an identity encoder that are supervised solely by image reconstruction losses. Despite achieving satisfactory facial output with favorable model simplicity, the output images are unavoidably turned into an all-black background. Further, to deal with cases containing extreme facial expressions and poses, Liu presented a large pose identity-preserving face reenactment network, LI-Net [54], that treats the pose and expression features separately and carefully by the face rotation module and the expression enhancing generator, respectively.

With the demand for boosting model ability, the few-shot and one-shot topics have become popular and several studies have been conducted accordingly. Zhang et al. [55] attempted to satisfy the one-shot scenario in this task, and they devised a disentangle-and-compose framework to correctly extract expression and pose features out of a single target image. Yao et al. [56] decomposed identity, expression, and pose features from the source and driving images, then recombined them to reconstruct the 3D mesh with desired features. Based on the optical flow between the original and desired 3D meshes, the output facial image can be easily derived under few-shot and one-shot conditions. Zakharov et al. [57] demonstrated a one-shot face reenactment pipeline that guides the expression and pose transfer by obtaining high- and low-frequency texture components from source and target images, respectively. Yao et al. [58] designed a novel appearance adaptive normalization mechanism to globally predict adaptive parameters of different layers using a skip-connected network, preventing the unresolved issue caused by non-pixel-level alignment. Meanwhile, some studies focus on facial expression synthesis without requiring a source image. Providing a target image and a desired facial expression label, the algorithms are capable of modifying the facial expression within the target image. Specifically, GANimation [59] utilizes action unit (AU) values with attention masks [60] to transfer facial images to standard facial expressions such as happy and angry. FReeNet [61] builds a unified landmark converter (ULC) that utilizes an auto-encoder to adapt standard expressions to the target faces.

Similar to the evolution of generative AI models in the face swapping task, recently, a few diffusion-based models have been raised thanks to the development of the large models. In detail, Collaborative Diffusion [62] can edit the facial attributes of the images upon receiving prompts by collaborating pre-trained uni-modal diffusion models. DiffusionRig [63], conditioned on crude 3D face models estimated from single in-the-wild images, illustrates outstanding performance by feeding a small number of portrait photos of a target identity. However, such an adoption trend on the DDPMs has just started and thus there is a considerable potential research gap.

III. Opportunities and Risks

Like many other new technologies, the flourishing of Deepfake can bring convenience to human lives in many industries such as educational media and digital communications [64]. On the contrary, huge negative consequences have gradually appeared on the dark side. In this section, we briefly discuss the potential opportunities and risks brought by Deepfake from both ends.

A. Opportunities

Benefiting from the publicly available source code implementations, various Deepfake mobile applications have been published on the internet. In the early times, FakeAPP[8] requires sufficient input images in order to contribute a reasonable synthetic output. Later, a popular Chinese application, ZAO[9], demonstrates superior face swapping performance while requiring a single image, especially for the ones without complex occlusions and bad light conditions. Other later software such as ReFace[10], Wombo[11], and FaceApp[12] continuously adopts and encapsulates state-of-the-art Deepfake algorithms, and has brought mass attention in people's entertainment.

[8] http://tinyurl.com/yj3r6ce2
[9] https://apps.apple.com/cn/app/id1465199127
[10] https://hey.reface.ai/
[11] http://tinyurl.com/rkf29d5s
[12] http://tinyurl.com/4rak9fsu

With the swift development of generative AI models, it is becoming easier to produce high-quality and hyper-realistic synthetic media. Generally, the Deepfake technique benefits both individuals and industries. On the one hand, individual film lovers may inject themselves into Hollywood movies and play the role of their favorite characters by performing face swapping. People can also digitally bring a deceased friend or family back and have conversations to make up for regrets via face reenactment upon their still photographs. As to the e-commerce industry, market brands can show fashion outfits with a diversity of faces using Deepfake, rather than for the customers to personally visit the off-line stores. Meanwhile, for global educational purposes, a malaria awareness campaign produces Deepfake David Beckham with multilingual speeches, attracting increasing attention from the soccer fan group[13]. Referencing this event, further educational activities can be arranged in the future.

Risks

Privacy Threats and Crisis of Confidence. In contrast to the positive benefits of Deepfake, misusing the technology can lead to severe consequences. Besides the popular 'fake Obama' video[14] circulating on the internet for educational and entertainment purposes, recently, a fake president Zelensky[15] has caused a crisis of confidence toward the government in Ukraine during the Russia-Ukraine war. In that fake video, President Zelensky tells Ukrainians to put down their weapons and give up resistance. In addition, there is already a large number of female celebrities suffering reputational loss due to Deepfake. Representative victims include the famous singer Ariana Grande and the well-known actress Emma Watson [65], who have been face swapped into porn videos that are largely distributed. Furthermore, the privacy threats of malicious Deepfake are closely approaching every human being. In June 2022, a lady was sentenced to probation for three years because of her harassment of the rivals on her daughter's cheerleader

[13] http://tinyurl.com/yj3r6ce2
[14] https://www.youtube.com/watch?v=cQ54GDm1eL0
[15] http://tinyurl.com/ywka3brs

team using Deepfake[16]. Unfortunately, due to the lack of evident and credible forgery detection tools, the justification for generating Deepfake videos is unable to be confirmed[17].

Due to the hyper-realistic quality that is indistinguishable by human eyes, Deepfake has already been ranked as the most serious artificial intelligence crime threat since 2020[18]. To resolve the ongoing and potential threats, the battle between malicious facial manipulation and Deepfake detection has been consistently continued [66]. In a nutshell, to reduce the threats and efficiently protect people's privacy and reputation, the current research domain is continuously focusing on conducting reliable deep-neural-network-based (DNN-based) Deepfake detectors that are generalizable to unseen data, interpretable with evidence that follows common sense, and robust when dealing with complex real-life scenarios [3].

Generative AI Models and the Environmental Issue. In the big data era, most generative AI models follow a common guideline, that is, more training data can usually advance better model performance given acceptable data qualities. On the other hand, the sufficient data quantity also benefits the data-consuming generative models for them to grow toward larger scales, where the additional modules and the corresponding model weights are designed to deeply and broadly analyze the latent features that are unfortunately ignored by the lightweight models in the early stages. Admittedly, the progress in data and model scales has brought out the best in each other for a period of time. However, such a fast development causes a significantly large demand for computational power. For example, a recent study [4] compares the number of parameters (Params.) and the number of floating point operations (FLOPs) of state-of-the-art (SOTA) face swapping models in the past years, and an obvious efficiency loss can be observed in Table I for most of the ones that come up with complex modules to deal with deeper features. As a result, the environmental issue has been frequently raised in discussions [71]. Therefore, besides

[16] http://tinyurl.com/mst7wk6v
[17] http://tinyurl.com/4w8c62ja
[18] http://tinyurl.com/4wpxfs6t

consistently improving the ability of generative AI models, the new research demand lies in the topic of green generative AI.

Model	Year	Resolution	FLOPs ↓	Params. ↓
Deepfakes[19]	2017	64	1.9G	82.1M
SimSwap [19]	2020	224	57.7G	107.3M
FaceShifter [15]	2020	256	97.4G	421.0M
InfoSwap [22]	2021	512	222.8G	121.6M
RAFSwap [25]	2022	256	83.6G	305.2M
UniFace [67]	2022	256	137.2G	92.4M
FaceSwapper [68]	2022	256	279.3G	97.3M
FSLSD [69]	2022	1,024	256.2G	324.1M
FaceDancer [70]	2023	256	58.1G	89.7M

Table I. Complexity evaluation of the face swapping models. The table is partially adopted from the latest study [4].

IV. Future Research Directions

In this survey, we comprehensively review the evolution of generative AI models in Deepfake face synthesis domain. For both categories, face swapping and face reenactment, in the domain, a similar developing history experiences auto-encoder, GAN, and DDPM as the model scale, data scale, and synthetic performance are all boosted continuously. As a new technology that has pros and cons, the opportunities and risks of Deepfake are illustrated accordingly.

Nevertheless, there remain considerable research gaps from both aspects. Regarding Deepfake face synthesis, although achieves relatively stable performance in common and easy scenarios such as the frontal face condition, faces are usually produced with unexpected artifacts, distortions, and unwanted content when encountering complex cases such as side faces and occlusions. Meanwhile, preserving a low computational cost simultaneously to resolve the environmental issue becomes a new topic, and new solutions are eagerly desired. Furthermore, although faces generated by the diffusion-based algorithms are generally clear and smooth with no obvious irregular trace, the existing models are not stable such that the high-resolution synthetic images overdo artifacts clean-up by accidentally removing common facial textures that are commonly contained in facial images. Therefore, future

[19] Source code at https://github.com/deepfakes

research is expected to maintain a reasonable balance regarding unwanted artifacts and common facial textures, and at the same time, to save the computational cost.

On the other hand, to deal with malicious Deepfake manipulation on facial images, a reliable detector that can be truly helpful in human lives, and even for relevant court-case prosecutions, is expected to demonstrate satisfactory ability in terms of transferability, interpretability, and robustness. In specific, research studies may focus on devising a detection model that maintains high accuracy when generalized to unseen testing data and evaluated on real-life noisy or low-quality data, while the evidence that a falsification as made based is easy to follow. Furthermore, to address the bottleneck that the current passive detectors are encountering, it is also meaningful to proactively insert invisible signals to untampered media in advance to prevent potential attacks. In particular, one can add invisible perturbations into image features that the generative models mostly rely on, and thus disables the synthetic workflow. Additionally, robust and semi-fragile watermarks can be utilized to prove the credibility by verifying the their existence after potential malicious Deepfake facial image manipulation.

References

[1] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10674–10685.

[3] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, “Deepfake detection: A comprehensive study from the reliability perspective,” 2023.

[4] T. Wang, Z. Li, R. Liu, Y. Wang, and L. Nie, “An efficient attributepreserving framework for face swapping,” IEEE Transactions on Multimedia, 2024.

[5] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, “Stylemask: Disentangling the style space of stylegan2 for neural face reenactment,” in 2023 IEEE International Conference on Automatic Face and Gesture Recognition, 2023, pp. 1–8.

[6] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” in 2017 IEEE International Conference on Computer Vision, 2017, pp. 3697–3705.

[7] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, “On face segmentation, face swapping, and face perception,” in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 98–105.

[8] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Um´e, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P.Wu, B. Zhou, and W. Zhang, “Deepfacelab: Integrated, flexible and extensible face-swapping framework,” 2021.

[9] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2d & 3d face alignment problem?” in 2017 IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.

[10] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, “Joint 3d face reconstruction and dense alignment with position map regression network,” in European Conference of Computer Vision, 2018, p. 557–574.

[11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3204–3213.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in The 27th Neural Information Processing Systems Advances, 2014, pp. 2672–2680.

[13] H. Zhu, C. Fu, Q. Wu, W. Wu, C. Qian, and R. He, "Aot: Appearance optimal transport based identity swapping for forgery detection," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 21699–21712.

[14] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7184–7193.

[15] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5073–5082.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp.4685–4694.

[17] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in 2017 IEEE International Conference on Computer Vision, 2017, pp. 1510–1519.

[18] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[19] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, p. 2003–2011.

[20] Y. Zhu, Q. Li, J.Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4832–4842.

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8107–8116.

[22] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, “Information bottleneck disentanglement for identity swapping,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3403–3412.

[23] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, “Hififace: 3d shape and semantic prior guided high fidelity face swapping,” in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 1136–1142.

[24] Z. Xu, H. Zhou, Z. Hong, Z. Liu, J. Liu, Z. Guo, J. Han, J. Liu, E. Ding, and J. Wang, “Styleswap: Style-based generator empowers robust face swapping,” in European Conference of Computer Vision, 2022.

[25] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, and Y. Liu, “Region-aware face swapping,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7622–7631.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[27] J. Kim, J. Lee, and B.-T. Zhang, “Smooth-swap: A simple enhancement for face-swapping with smoothness,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10769–10778.

[28] H. Zeng, W. Zhang, C. Fan, T. Lv, S. Wang, Z. Zhang, B. Ma, L. Li, Y. Ding, and X. Yu, “Flowface: Semantic flow-guided shape aware face swapping,” in Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, 2023.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Doll´ar, and R. Girshick, “Masked autoencoders are scalable vision learners,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15979–15988.

[30] C. Shu, H. Wu, H. Zhou, J. Liu, Z. Hong, C. Ding, J. Han, J. Liu, E. Ding, and J. Wang, "Few-shot head swapping in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022.

[31] Z. Xu, Z. Hong, C. Ding, Z. Zhu, J. Han, J. Liu, and E. Ding, "Mobilefaceswap: A lightweight framework for video face swapping," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[32] Z. Chang, G. A. Koulieris, and H. P. H. Shum, "On the design fundamentals of diffusion models: A survey," 2023.

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 6840–6851.

[34] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. M¨ uller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023.

[35] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in International Conference on Learning Representations, 2022.

[36] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[37] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human preference score: Better aligning text-to-image models with human preference," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2096–2105.

[38] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12715–12725.

[39] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, “Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion,” in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8568–8577.

[40] R. Xu, Z. Zhou, W. Zhang, and Y. Yu, “Face transfer with generative adversarial network,” 2017.

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

[42] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5967–5976.

[43] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, “Reenactgan: Learning to reenact faces via boundary transfer,” in European Conference of Computer Vision, ser. Lecture Notes in Computer Science, vol. 11205, 2018, pp. 622–638.

[44] J. Thies, M. Zollhofer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” ACM Trans. Graph., vol. 38, no. 4, 2019.

[45] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “Faceforensics++: Learning to detect manipulated facial images,” in 2019 IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.

[46] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, “Marionette: Few-shot face reenactment preserving identity of unseen targets,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 10893–10900, 2020.

[47] K. Yang, K. Chen, D. Guo, S.-H. Zhang, Y.-C. Guo, and W. Zhang, “Face2faceρ: Real-time high-resolution one-shot face reenactment,” in European Conference of Computer Vision, 2022, p. 55–71.

[48] G.-S. Hsu, C.-H. Tsai, and H.-Y. Wu, “Dual-generator face reenactment,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 632–640.

[49] Q. Ren, Z. Lu, H.Wu, J. Zhang, and Z. Dong, “Hr-net: A landmark based high realistic face reenactment network,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 11, pp. 6347–6359, 2023.

[50] K. Gu, Y. Zhou, and T. S. Huang, “Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis,” in The Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, pp. 10861–10868.

[51] P. Huang, F. Yang, and Y. Wang, “Learning identity-invariant motion representations for cross-id face reenactment,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7082–7090.

[52] X. Zeng, Y. Pan, M. Wang, J. Zhang, and Y. Liu, “Realistic face reenactment via self-supervised disentangling of identity and pose,” in AAAI Conference on Artificial Intelligence, 2020.

[53] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky, “Neural head reenactment with latent pose descriptors,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[54] J. Liu, P. Chen, T. Liang, Z. Li, C. Yu, S. Zou, J. Dai, and J. Han, “Linet: Large-pose identity-preserving face reenactment network,” in 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1–6.

[55] Y. Zhang, S. Zhang, Y. He, C. Li, C. C. Loy, and Z. Liu, “One-shot face reenactment,” in British Machine Vision Conference, 2019.

[56] G. Yao, Y. Yuan, T. Shao, and K. Zhou, “Mesh guided one-shot face reenactment using graph convolutional networks,” in Proceedings of the 28th ACM International Conference on Multimedia, 2020, p. 1773–1781.

[57] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky, “Fast bi-layer neural synthesis of one-shot realistic head avatars,” in European Conference of Computer Vision, 2020.

[58] G. Yao, Y. Yuan, T. Shao, S. Li, S. Liu, Y. Liu, M. Wang, and K. Zhou, “One-shot face reenactment using appearance adaptive normalization,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, pp. 3172–3180, 2021.

[59] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Ganimation: Anatomically-aware facial animation from a single image,” in European Conference of Computer Vision, 2018, pp. 835–851.

[60] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[61] J. Zhang, X. Zeng, M. Wang, Y. Pan, L. Liu, Y. Liu, Y. Ding, and C. Fan, “Freenet: Multi-identity face reenactment,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5325–5334.

[62] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, “Collaborative diffusion for multi-modal face generation and editing,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[63] Z. Ding, C. Zhang, Z. Xia, L. Jebe, Z. Tu, and X. Zhang, “Diffusionrig: Learning personalized priors for facial appearance editing,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[64] M. Westerlund, “The emergence of deepfake technology: A review,” Technology Innovation Management Review, vol. 9, pp. 40–53, 2019.

[65] T. Wang and K. P. Chow, “Noise based deepfake detection via multi-head relative-interaction,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 12, pp. 14 548–14 556, 2023.

[66] F. Juefei-Xu, R.Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, “Countering malicious DeepFakes: Survey, battleground, and horizon,” International Journal of Computer Vision, vol. 130, no. 7, pp. 1678– 1734, 2022.

[67] C. Xu, J. Zhang, Y. Han, G. Tian, X. Zeng, Y. Tai, Y. Wang, C. Wang, and Y. Liu, “Designing one unified framework for highfidelity face reenactment and swapping,” in European Conference of Computer Vision, 2022, pp. 54–71.

[68] Q. Li, W. Wang, C. Xu, and Z. Sun, “Learning disentangled representation for one-shot progressive face swapping,” 2022.

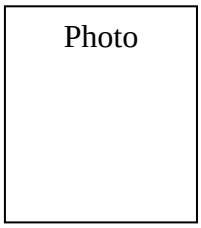
[69] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, “Highresolution face swapping via latent semantics disentanglement,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7642–7651.

[70] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund, “Facedancer: Pose- and occlusion-aware high fidelity face swapping,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3454–3463.

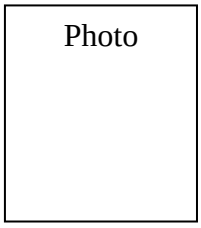
[71] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “Deep learning’s diminishing returns: The cost of improvement is becoming unsustainable,” IEEE Spectrum, vol. 58, no. 10, pp. 50–55, 2021.

Photo

First A. Author’s biography.



Second B. Author’s biography.



Third C. Author’s biography.