Face Swapping in the Deepfake Era: Technologies, Applications, and Ethical Concerns

Kaige Zhang *

Department of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China

* Corresponding Author Email: zhangkaige725@gmail.com

Abstract. The technology of face-swapping has developed from conventional manual editing to the emergence of deepfake technology powered by deep learning and then to the stage of diversified development, achieving improvements in terms of speed, realism and stability, and being applied to film, television, entertainment and virtual makeup, while also giving rise to problems such as false information and privacy violation attracting attention to the ethical and legal restrictions; among the mainstream models, GHOST, a deep-learning-based framework combining GANs and autoencoders, performs well in high-precision face-swapping by extracting and mapping facial features, keeping details and working stably under complex lighting but with large parameters and slow inference; Inswapper, an advanced pre-trained model trained on millions of diverse facial images, performs well in 3D facial structure inference and feature separation, making realistic transfers possible and being integrated into various tools flexibly but with potential blurring in results; SimSwap, a GAN-based framework with a generator and a discriminator, uses an ID Injection Module and weak feature loss for arbitrary swapping and target attribute keeping, although requiring considerable computing resources and possibly producing artifacts at extreme angles.

Keywords: Face swapping, GHOST, Inswapper, SimSwap.

1. Introduction

Recently there have been many advances in the field of facial generation/editing. A 2023 report on the digital media industry suggests that more than 60% of VFX studios around the world are already using AI face-swapping technology for such purposes as restoring characters faces or replacing stunt doubles [1]. The technology has matured enough to satisfy the ISO/IEC 30107-3 standard for biometric presentation attack detection.

Although the existing face-swapping systems have made remarkable progress in terms of technology, they have major deficiencies in user experience. For example, mainstream mobile apps like ZAO 141414 (*https://zao.ai/*) and Reface 151515 (*https://re-face.com/*) can only perform single face-swapping on static images, without supporting complex scenarios such as dynamic video stream processing and real-time interactive feedback. In academia, the emphasis is still on improving the technical indicators of generated images, such as PSNR and SSIM, with little consideration given to non-technical factors such as users entertainment needs. Such a structural gap between technological development and market demand makes it difficult for users to obtain persistent interactive experience, leading to the products retention rate always below the industry average.

User demand research data shows that the new generation of digital natives is increasingly demanding interactive entertainment experiences. On short video platforms, where face-swapping effects are applied over 200 million times per day, 78% of users indicate that the existing effect templates are highly homogeneous and insufficiently customizable. Meanwhile, the rising popularity of augmented reality (AR) technology has fostered a need for combining real-time facial tracking with dynamic face-swapping. However, existing systems have not yet surmounted critical technical challenges such as millisecond-level latency control and multi-angle lighting adaptation. This limitation not only hinders the application of face-swapping technology in emerging areas such as live e-commerce and virtual idols, but also pushes many potential users towards other digital entertainment products.

2. Historical Evolution of Face Swapping

2.1. Traditional Image Processing Era

The evolution of face swapping technology is a process of constant breaking through of technical bottlenecks, extending the boundary of applications, and stimulating discussion of ethical and social implications. In its infancy, people tried to do face swapping based on conventional image processing techniques in a manual or semi-automatic way. For instance, professionals working on some simple image editing works manually adjusted the positions of facial features and the shape of contours by some special software to realize preliminary face swapping. The procedure was tedious, the effect was crude, and it could only be applied to low-resolution images with few operations, which was far from enough to satisfy the requirements in complex situations. It was confined to the narrow field of artistic creation, providing a primitive accumulation for further technological progress.

2.2. Rise of Deepfake Technology

With the advent of deep learning algorithms, face-swapping technology took a giant stride and stepped into the age of deepfake technology. The birth of Deepfakes technology in 2017 was a milestone. It adopted GANs and obtained facial features from massive amounts of data via the game-like learning between the generator network and the discriminator network, accomplished the exchange of facial identity and attribute information, and produced realistic face-swapped images/videos. On this basis, the desktop application *FakeApp* reduced the operation threshold so that users without technical knowledge could easily generate face-swapped videos, which facilitated the rapid diffusion of face-swapping technology on the Internet and was a pivotal moment for its popularization.

2.3. Diversified Development and Current Situation

Since then, the face-swapping technology keeps branching out and growing into different technical categories. Face Swapping can seamlessly blend identity and attribute information of different faces; Face Reenactment concentrates on transferring pose and expression attributes without identity change; Facial Attribute Editing alters local features such as hair color, age, etc.; Face Synthesis generates new faces from random noise, demonstrating its potential in virtual character generation. Nowadays, with the evolution of large AI models, the technology has gained speed, realism and stability, finding a wide range of applications in film and television, entertainment, virtual makeup, etc. Meanwhile, it also leads to problems of fake news, privacy violation, etc., which urges people to pay more attention to the ethical and legal restrictions on technological development [2].

3. Face-swapping Algorithm Models

3.1. Ghost

GHOST is a strong face swapping framework based on deep learning which combines the benefits of GANs and autoencoders to obtain high quality results [3]. It differs from previous approaches in that it combines the two techniques to extract and align a large variety of facial features from both source and target images, resulting in identity-aware, realistic-looking face swaps.

It usually starts with a fast and robust face detection. It not only helps to identify some key points on the face (eyes, nose, mouth), but also to capture finer details like the shape of the eyebrows, small muscle movements, ambient lighting etc [4]. GHOST exploits these structural and expressive details to transfer the expression of the target face, yet keeping the identity of the source [5].

Another attractive point of GHOST is that it can cooperate with the state-of-the-art facial recognition system *ArcFace*, which is famous for its accuracy. ArcFace applies angular margin loss and convolutional network to extract feature vectors highly representative of facial identity [6]. During face-swapping, ArcFace firstly extracts the identity vector of the source face. Then this vector

is blended with expression features borrowed from the target. The obtained result is passed to GHOST, which produces results keeping the same facial identity, while transferring expressions convincingly [7].

GHOST can be applied to a variety of other scenarios. In the film and video industry, it is already being used to map real faces onto digital characters to produce more realistic and authentic virtual performances. Because GHOST is optimized for being lightweight and efficient, its output may not always be visually on par with more heavy-weight models used in professional settings.

3.2. SimSwap

SimSwap is a state-of-the-art face-swapping framework based on GANs that leverages conditional GANs from deep learning to generate high-quality results [8]. It consists of two main components - the generator and the discriminator that interact with each other in an interwoven way. The former is responsible for encoding and processing the facial attributes whereas the latter is in charge of decoding and reconstructing them.

The generator itself consists of three main components. The first one is the encoder that extracts the important features of the target image. The second component is the ID Injection Module (IIM) which is a dedicated system to inject the source face identity information [9]. The last component is the decoder that reconstructs the face-swapped image with high fidelity. During the analysis of face features, SimSwap usually depends on ArcFace to extract the identity vectors of the source face accurately. These vectors are then injected into the targets features by several ID blocks - special versions of residual modules that utilize *AdaIN* instead of the usual batch normalization [10]. This helps the model to align source identity with targets features leading to more natural face-swapping effects.

The discriminator is responsible for fine-tuning the generated images. It takes the output of the generator and contrasts it with real images, measuring the distance between them via an error loss function. Should the generated images be within an acceptable tolerance, the discriminator passes this feedback to the generator, which in turn employs this information to enhance its results in an iterative learning process [11].

SimSwap has several advantages. Perhaps its most striking feature is the proposed IIM, which enables face swap between any two faces without the necessity of training a separate model for each pair. This greatly alleviates the burden of computation resources and relaxes the hardware requirement, making it more amenable to the public. Moreover, the weak feature matching loss adopted by SimSwap helps to preserve the integrity of facial details, such as expression, pose, and illumination variation. This facilitates the swapped face to blend into the target image better and avoids distortion generally incurred in image generation tasks.

Despite its advantages, SimSwap has some limitations. When the target attributes are complex (e.g., change lighting in dynamic scenes, large resolution gap between source and target), artifacts can be observed in the generated faces that cause distortion. Sometimes, the model fails to correctly infer the details near the boundary of the source face, e.g., hairline, slight face contour, etc., which results in obvious inconsistency. Moreover, the complex architecture and carefully designed loss function need a lot of computation resources. It can consume a lot of resources, which makes it impractical on low-computing-power devices such as smartphones, low-end PC, etc.

3.3. Inswapper

Inswapper is a pre-trained AI model that has been trained on millions of facial images, capturing a wide variety of angles, expressions, and lighting conditions [12]. This extensive training enables the model to accurately identify and extract facial features, even in complex or challenging situations, providing a solid foundation for high-quality face-swapping.

What makes Inswapper special is that it can infer 3D structure of faces from 2D images. While most models concentrate only on 2D, Inswapper applies special algorithms to analyze spatial and geometrical properties of faces. Such 3D understanding helps the model to generate less distorted

images even for non-frontal facial poses - it works much better than other models for not frontal-facing faces.

Besides that, Inswapper has a sophisticated way of separating identity features (e.g. a unique face shape of a person) from pose features (head angles, smile, etc.). It allows to mix identity of one person with expression, pose and lighting of another resulting in very natural-looking final face swap. The model is able to adapt features of the original face to its surroundings, thus improving realism of the swap.

Inswapper is already included in some popular tools such as Deep-Live-Cam, where the model does fast face detection and processing, completing the face swap operation with remarkable speed and accuracy. The application then applies sophisticated image restoration models such as GFPGAN to refine facial details, sharpen details, remove artifacts, and enhance overall image quality. Inswapper has the benefit of being able to be easily incorporated into various workflows and tailored to specific needs. It also provides models like mobile0.25, allowing the user to choose the most suitable model depending on the hardware capabilities and desired performance. This leaves space for professional optimization according to specific necessities.

However, there are some drawbacks with Inswapper. Sometimes the face swapping images turn out to be a little bit blurry. It is a usual problem for face swapping models - if you swap faces on images with significantly different poses or lightning conditions, the result will likely be inaccurate since the model has to somehow combine conflicting details. Nevertheless, it is still one of the most effective and universal tools in its class.

4. Conclusion

Finally, the dataset on which these models are trained is just as important as the model itself. It is crucial to have a rich and balanced dataset with respect to ethnicity, age, lighting, and expression. The dataset can be augmented through techniques such as flipping, rotation, changing brightness, etc. This allows for better generalization to real world images.

Face-swapping technology has gained widespread traction across different industries, particularly in film and television, where it's used to create seamless visual effects. Despite its progress, the technology is far from perfect, and there's still considerable potential for improvement.

One of the most promising approaches to improve face-swapping models is to strengthen the backbone architecture. Using more powerful convolutional neural networks such as *ResNet* and *DenseNet* can greatly improve the model capacity to learn fine-grained details on faces. Such networks are designed to build deep yet efficient architectures that allow the model to capture subtle details in facial geometry. If equipped with some form of attention mechanism (e.g., spatial or channel attention), the model can better focus on the relevant information and thus better preserve identity and transfer expressions.

Another aspect that could be considered is multi-level feature fusion, where high-level structure and low-level details are combined to balance the global facial consistency and local details such as subtle facial motions/expression. The consistency can also be improved at the preprocessing step by performing some simple operations such as contrast adjustment, denoising, lighting normalization, etc., which help alleviate the variation caused by different cameras/camera settings and environments

Loss function design also plays a critical role in the training process. A tailored loss function that prioritizes identity preservation while minimizing distortion can greatly enhance visual fidelity. Meanwhile, adopting optimization techniques like *AdamW* not only speeds up training but also reduces the risk of overfitting—a common challenge in deep learning.

Finally, the dataset on which these models are trained is just as important as the model itself. To achieve robust results, it is critical that the dataset contains variation in terms of ethnicity, age, lighting conditions, and facial expressions. The dataset can be augmented through techniques such as flipping, rotation, changing brightness, etc., to help the model generalize better to unseen scenarios.

References

[1] Du P, Li C, Dong C. Face Swapping for Film and Television Video based on FaceNet and Local Translation Warp. 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS). IEEE, 2022: 290-295.

[2] Tuysuz M K, Kılıç A. Analyzing the legal and ethical considerations of deepfake technology. Interdisciplinary Studies in Society, Law, and Politics, 2023, 2(2): 4-10.

[3] Groshev A, Maltseva A, Chesakov D, et al. GHOST—a new face swap approach for image and video domains. IEEE Access, 2022, 10: 83452-83462.

[4] Nirkin Y, Masi I, Tuan A T, et al. On face segmentation, face swapping, and face perception. Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 98-105.

[5] Li L, Bao J, Yang H, et al. Advancing high fidelity identity swapping for forgery detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5074-5083.

[6] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4690-4699.

[7] Agarwal A, Sen B, Mukhopadhyay R, et al. FaceOff: A video-to-video face swapping system. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 3495-3504.

[8] Chen R, Chen X, Ni B, et al. Simswap: An efficient framework for high fidelity face swapping. Proceedings of the 28th ACM international conference on multimedia. 2020: 2003-2011.

[9] Blanz V, Scherbaum K, Vetter T, et al. Exchanging faces in images. Computer Graphics Forum. Oxford, UK and Boston, USA: Blackwell Publishing, Inc, 2004, 23(3): 669-676.

[10] Natsume R, Yatagawa T, Morishima S. Fsnet: An identity-aware generative model for image-based face swapping. Asian Conference on Computer Vision. Cham: Springer International Publishing, 2018: 117-132.

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[12] Thies J, Zollhofer M, Stamminger M, et al. Face2face: Real-time face capture and reenactment of rgb videos. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2387-2395.