Pattern Recognition 141 (2023) 109628

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Deepfacelab: Integrated, flexible and extensible face-swapping framework

Kunlin Liu[a], Ivan Perov[b], Daiheng Gao[b], Nikolay Chervoniy[b], Wenbo Zhou[a,1,*], Weiming Zhang[a,2]

[a] Department of Cyber Science and Technology, University of Science and Technology of China, Hefei, 230026, Anhui, China
[b] Freelancer

ARTICLE INFO

Article history:
Received 23 October 2022
Revised 14 January 2023
Accepted 21 April 2023
Available online 28 April 2023

Keywords:
Face swapping
Practical machine learning
Open source

ABSTRACT

Face swapping has drawn a lot of attention for its compelling performance. However, current deepfake methods suffer the effects of obscure workflow and poor performance. To solve these problems, we present DeepFaceLab, the current dominant deepfake framework for practical face-swapping. It provides the necessary tools as well as an easy-to-use way to conduct high-quality face-swapping. It also offers a flexible and loose coupling structure for people who need to strengthen their pipeline with other features without writing complicated boilerplate code. We detail the principles that drive the implementation of DeepFaceLab and introduce its pipeline. DeepFaceLab could achieve cinema-level results with high fidelity as our supplemental video shows. We also demonstrate the advantage of our system by comparing our approach with other face-swapping methods.

Deepfake defense not only requires the research of detection but also requires the efforts of generation methods. As for a popular and practical toolkit, we encourage users to promote harmless deepfake-entertainment content on social media, reminding the public of the existence of deepfake when they are looking for entertainment.

1. Introduction

Face swapping is a task to swap one person's face to another, preserving other attributes like expressions, head poses, and backgrounds (as shown in Fig. 1). Face swapping is an eye-catching task in generating fake content by transferring a source face to the destination while maintaining the destination's facial movements and expression deformations. Since deep learning has empowered the realm of computer vision in recent years, manipulating digital images, especially the manipulation of human portrait images, has improved rapidly and achieved photorealistic results in most cases.

In 2018, DeepFakes [1] introduced a complete production pipeline in replacing a source person's face with the target person's along with the same facial expression such as eye movement, and facial muscle movement. However, the results produced by DeepFakes are poor somehow, and so are the results with contemporary Nirkin et al.'s automatic face swapping [2]. With the fast development of generation ability, the fundamental motivation behind face manipulation techniques, Generative Adversarial Networks (GANs) [3], could synthesize more and more realistic faces which are entirely indistinguishable from the human vision system, i.e., StyleGAN [4] and Alias-free GAN [5]. Recent advances have led to improved face manipulation techniques, such as GAN face synthesis [6,7] and facial attribution editing [8]. More and more face-swapping methods [9–14] are proposed subsequently, but these methods mainly focus synthesizing swapped images with one-shot source portrait. Despite the performance improvement, however, existing models usually suffer from heavy identity mismatch. These methods adopt complex model architectures and numerous loss functions to constrain face shape, which represents most of the identity information. For example, previous studies have focused on using handcrafted components such as mask-based mixing [10] or 3D face-shape modeling [11,13] to enhance the performance. Further research proposes instead a new identity embedding model having improved smoothness, which achieved

* Corresponding author.
E-mail addresses: lkl6949@mail.ustc.edu.cn (K. Liu), welbeckz@ustc.edu.cn (W. Zhou), zhangwm@ustc.edu.cn (W. Zhang).

[1] Wenbo Zhou participated in the design of the methodology framework, identified potential ethical and safety risks and proposed solutions, and provided guidance to the first author in revising the paper.

[2] Weiming Zhang participated in the design of the methodology framework, provided guidance to the student in writing the paper and designing the experimental plan.

https://doi.org/10.1016/j.patcog.2023.109628

Fig. 1. Face swapping results generated by DeepFaceLab (DFL). Left: Source face. Middle: Destination face for replacement. Our results appear on the right, demonstrating that DFL could handle occlusion, complex illumination, and side face with high fidelity.

state-of-the-art performance. [15] However, no matter how these models are adjusted, the identity information is always obtained from one shot source image. On the other hand, the identity information these methods leverage is guessed by a prior model.

Different from these methods, we propose DeepFaceLab (DFL for short), a multi-shot face-swapping system. The DFL leverage enormous data from the destination person and source person to execute face swapping. We first collect at least hundreds of portraits of the source person which have different angles and backgrounds, and these data will provide abundant source identity information. Then, we train neural networks to make the neural networks carry the specific identity information. With the specific design, The DFL could achieve cinema-level face-swapping results with high fidelity.

As a potential technique, face-swapping always faces ethical problems. Face-swapping techs show great performance on some commercial applications such as Zhubobao[3] which helps sellers to have good looks. Commercial mobile applications such as FaceApp[4] and FacePlay[5] which allow general netizens to create fake images and videos effortlessly significantly boost the spreading of these swapping techniques. These content generation and modification technologies may affect public discourse quality and infringe upon the citizens' portrait rights, especially given that deepfake may be used maliciously as a source of misinformation, manipulation, harassment, and persuasion. Identifying manipulated media is a technically demanding and rapidly evolving challenge that requires collaborations across the entire tech industry and beyond.

Research into effective deepfake detection has led to a variety of proposed methods, such as PRRNet [16], which focuses on pixel-wise deepfake detection and achieves good performance, and deep dual-level networks [17], which offer more robust detection. It was later pointed out that most face swapping methods had bi-granularity artifacts, and a BiG-based detection method [18] was proposed. Leveraging networks to learn and enhance multiple tampering traces can also be an effective way for manipulation detection [19].

However, deepfake defense not only requires the research of detection but also requires the efforts of generation methods. In order to further awaken people's awareness of facial-manipulation videos and provide convenience for forgery detection researchers, we established an open-source deepfake project, DeepFaceLab (DFL for short), which is used to build high-quality face-swapping videos for entertainment and greatly help the development of forgery detection by providing high-quality forgery data.

Research on media anti-forgery detection is being invigorated and dedicating growing efforts to forgery face detection. DFDC[6] is a typical example, a million-dollar competition launched by Facebook and Microsoft. Training robust forgery detection models requires high-quality fake data. Data generated by our methods are involved in the DFDC dataset [20].

On the other hand, detection after being attacked is not the unique manner for reducing the malicious influence of deepfake. It is always too late to detect the spreading spoofing content. In our perspective, for both academia and the general public, helping netizens know what deepfake is and how a cinema-level swapped video is generated is much better. Making general netizens realize the existence of deepfake and strengthening their identification ability for spoof media published on social media is much more critical than agonizing the fact whether spoof media is true or not.

This paper introduces DeepFaceLab, an integrated open-source system with a human-in-the-loop design of the pipeline, achieving photorealistic face-swapping results without painful tuning. DFL has turned out to be very popular with the public. For instance, many visual effect artists create DFL-based videos and publish them on their YouTube channels, which attracts enormous hits so far.

The contributions of DeepFaceLab can be summarized as threefolds:

- A state-of-the-art multi-shot face-swapping framework consists of a maturity pipeline is proposed, aiming to achieve photorealistic face-swapping results.
- DeepFaceLab open-sourced the code in 2018 and always kept up to the progress in the computer vision area, making a positive contribution for defending deepfake, which has drawn broad attention in the open-source community and VFX areas.
- A series of high-efficiency components and tools are introduced in DeepFaceLab to build better face-swapping videos.

2. Characteristics of deepfacelab

DeepFaceLab's success stems from weaving previous ideas into a design that balances speed and ease of use and the booming of computer vision in face recognition, alignment, reconstruction, segmentation, etc. There are Four main characteristics behind our implementation:

[3] https://www.zhubobao.com.
[4] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341.
[5] https://apps.apple.com/us/app/faceplay-cosplay-video-maker/id1559859897.
[6] https://deepfakedetectionchallenge.ai/.

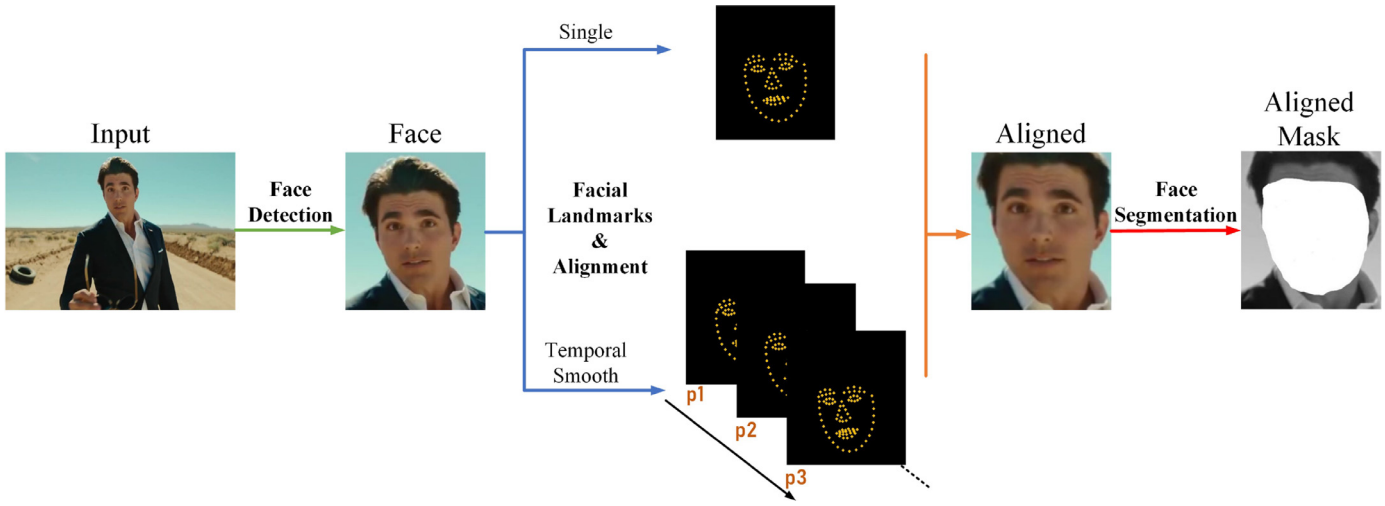

Fig. 2. Overview of extraction phase in DeepFaceLab (DFL for short).

Convenience DFL strives to make the usage of its pipeline, including data loader and processing, model training, and post-processing, as easy and productive as possible. Unlike other face swapping systems, DFL provides a complete command-line tool with every aspect of the pipeline that could be implemented in the way that users choose. Notably, the complexity inherent and many hand-picked features for fine-grained control such as the canonical face landmark for face alignment should be handled internally and hidden behind DFL. People could achieve the smooth and photorealistic face-swapping results without the need for hand-picked features if they follow the settings of the workflow, but only with the need of two videos: the source video (src) and the destination video (dst) without the requirement to pair the same facial expression between src and dst.

Wide engineering support Some practical measures were added to improve the performance: multi-GPU support, half-precision training, usage of pinned CUDA memory to improve throughput, use of multiple threads to accelerate graphics operations and data processing. To support different hardware environments, DFL allows model customization. Users can adjust model parameters and types to adapt to their environments. Even a machine with 2GB VRAM can also conduct a successful face-swapping.

Extensibility To strengthen the flexibility of DFL's workflow and attract the interests of the research community, any component of DFL that does not meet requirements is alternative. Most of DFL's modules are designed to be interchangeable. For instance, people could provide a newer face detector to achieve higher performance in detecting faces with extreme angles or outlying areas. Some interesting applications, i.e., virtual human swapping, could be extended with minor modification.

Practicality Good datasets is essential for DFL. Generally, the larger the datasets, the better the final results. However, results that are directly extracted from src and dst are always with noises and lights, which may be prejudicial to the final quality. In consideration of the complex situations of input data, DFL provides a series of measures to clean up datasets. With these measures, DFL achieve robust practicality. DFL could support 10,000-level datasets and conduct cinema-level face-swapping basing on large datasets.

3. Pipeline

DeepFaceLab provides a set of workflow which form the flexible pipeline. In DFL, we can abstract the pipeline into three main phases: extraction, training, and mergence. These three parts are presented sequentially. Besides, it is noteworthy that DFL falls in a typical one-to-one face-swapping paradigm, which means there are only two kinds of data: src and dst, the abbreviation for source and destination, are used in the following narrative.

3.1. Extraction

The extraction phase is the first phase in DFL, aiming to extract faces from src and dst data. As shown in Fig. 2, this phase consists of many algorithms and three processing parts, i.e., face detection, face alignment, and face segmentation. As a practical system, DFL provides several extraction modes (i.e, half – face, full – face, whole – face), which represents the face coverage area of the extraction phase, to meet different demands.

Face Detection The first step in Extraction is to find the target face in the given data: src and dst. DFL regards S3FD [21] as its default face detector. S3FD can be replaced with other face detection algorithms painlessly, i.e RetinaFace [22]. In consideration of the missed and wrong detection, we could use Machine Video Editor[7] to further process the data as a human-in-the-loop way.

Face Alignment The second step is face alignment. After numerous experiments and failures, we realized that an effective facial landmarks algorithm essential in producing an excellent successive footage shot and film. DFL provides two canonical types of facial landmark extraction algorithms to solve this: (a) heatmap-based facial landmark algorithm 2DFAN [23] (for faces with standard pose) and (b) PRNet [24] with 3D face prior information (for faces with large Euler angle (yaw, pitch, roll), e.g., A face with a large yaw angle, means one side of the face is out of sight). After facial landmarks are retrieved, we also provide an optional function with a configurable time step to smooth facial landmarks of consecutive frames in a single shot to ensure stability further.

Then we adopt a classical point pattern mapping and transformation method proposed by [25] to calculate a similarity transformation matrix used for face alignment.

Face Segmentation After face alignment, a data folder with face of standard front/side-view (aligned src or aligned dst) is obtained. We employ a fine-grained Face Segmentation network (TernausNet [26]) on top of (aligned src or aligned dst), through which a face with either hair, fingers, or glasses could be segmented exactly. It is optional but useful to remove irregular occlusions to keep the network in the training process robust to hands, glasses, and any other objects which may cover the face somehow.

[7] https://github.com/MachineEditor/MachineVideoEditor.

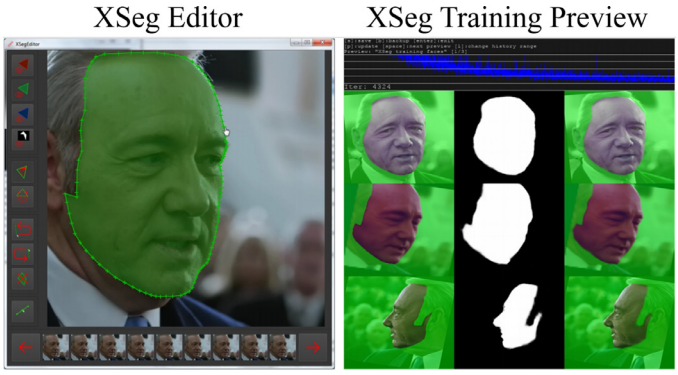

Fig. 3. The preview of XSeg. Users can label the masks they want by XSegEditor. With the help of XSeg, users can use it to eliminate the occlusion of hands, glasses, and any other objects which may cover the face somehow and control specific areas for swapping.

However, since some state-of-the-art face segmentation models fail to generate fine-grained masks in some particular shots, the XSeg was introduced in DFL. As shown in Fig. 3, XSeg allows everyone to train their model for the segmentation of a specific face set (aligned src or aligned dst) through a few-shot learning paradigm (Fig. 3 is the schematic of XSeg). With the human-in-the-loop design, face swapping could be implemented in any area.

As the above workflow executed sequentially, everything DFL needs in the training phase is already prepared: cropped faces with their corresponding coordinates in its original images, facial landmarks, aligned faces, and pixel-wise segmentation masks from src (Since the extraction procedure of dst is the same with src, there is no need to elaborate that in detail).

3.2. Training

The training phase plays the most vital role in achieving photo-realistic face-swapping results of DFL.

No need for facial expressions of aligned src and aligned dst being strictly matched, DFL aims to provide an efficient algorithm paradigm to solve this unpaired problem along with maintaining high fidelity and perceptual quality of the generated face. Motivated by the previous methods, we propose two structures, DF structure, and LIAE structure, to address this issue.

As shown in Fig. 4, DF structure consists of an Encoder as well as Inter with shared weights between src and dst, two Decoders which belong to src and dst separately. The generalization of src and dst is achieved through the shared Encoder and Inter, which solves the aforementioned unpaired problem easily. DF structure can finish the face-swapping task and produce results that fits src data. However, this structure cannot inherit enough information from dst, such as lighting.

To further enhance the problem of light consistency, we propose Lightly Improved Auto-Encoder(LIAE for short).

As depicted in Fig. 5, LIAE structure is a more complex structure with a shared-weight Encoder, Decoder and two independent Inters. The main difference compared to the DF is that InterAB is used to generate both latent code of src and dst while InterB only output the latent code of dst. Here, F_{src}^{AB} denotes the latent code of src produced by InterAB and we generalize this representation to F_{dst}^{AB}, F_{dst}^{B}. After getting all the latent codes from InterAB and InterB, LIAE then concatenate these feature maps through channel: $F_{src}^{AB}||F_{src}^{AB}$ is obtained for a new latent code representation of src and $F_{dst}^{AB}||F_{dst}^{B}$ for dst as the same way.

Then $F_{src}^{AB}||F_{src}^{AB}$ and $F_{dst}^{AB}||F_{dst}^{B}$ are put into the Decoder and we get the predicted src (dst) alongside with their masks. The motivation of concatenating F_{dst}^{B} with F_{dst}^{AB} is to shift the direction of latent code in direction of the class (src or dst) we need, through which InterAB obtained a compact and well-aligned representation of src and dst in the latent space.

Except for the structure of the model, some useful tricks are effective for improving the quality of the generated face: A weighted sum mask loss in general SSIM [27] can be added to make each part of the face carry different weights under the AE training architecture, for example, we add more weights to the eye area than the cheek, which aims to make the network concentrate on generating a face with vivid eyes.

As for losses, DFL uses a mixed loss (DSSIM (structural dissimilarity) [28] + MSE) by default. The motivation for this combination is to get benefits from both: DSSIM generalizes human faces faster while MSE provides better clarity. This combination of losses serves to find a compromise between generalization and clarity.

Besides, we adopt a fancy true face mode TrueFace, which serves for the generated face of better likeness to the dst in the mergence phase. For LIAE structure, we enforce F_{src}^{AB} approaches F_{dst}^{AB}. And for DF structure, counterparts turn out to be F_{src} and F_{dst}. Two rarely used methods have been validated by DFL: Convolutional Aware Initialization [29] along with Learning Rate Dropout [30], which greatly enhanced the final quality of the fake face.

3.3. Mergence

The mergence phase is the last but not least phase. Previous methods often ignore the importance of this phase. As depicted in Fig. 6, users can swap faces of src to dst and vice versa. In the case of src2dst, the first step of the proposed face-swapping scheme in the mergence phase is to transform the generated face alongside with its mask from dst Decoder to the original position

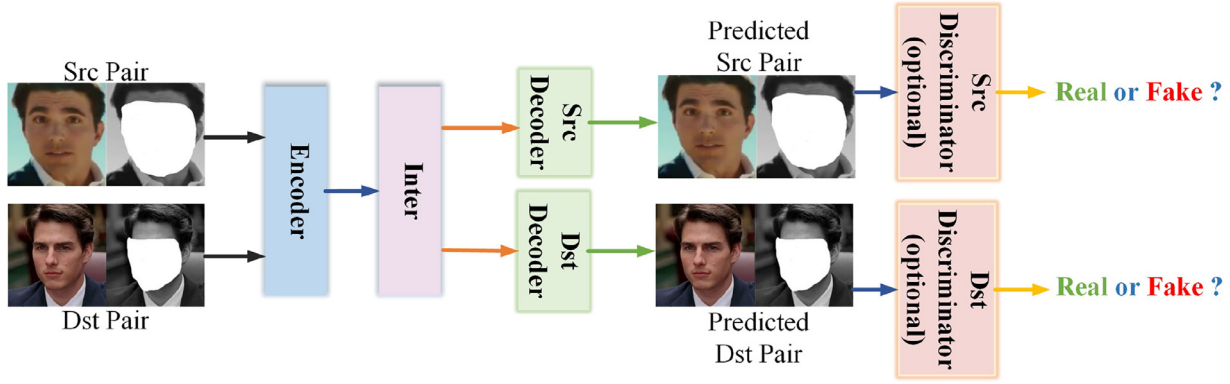

Fig. 4. Overview of DF structure.

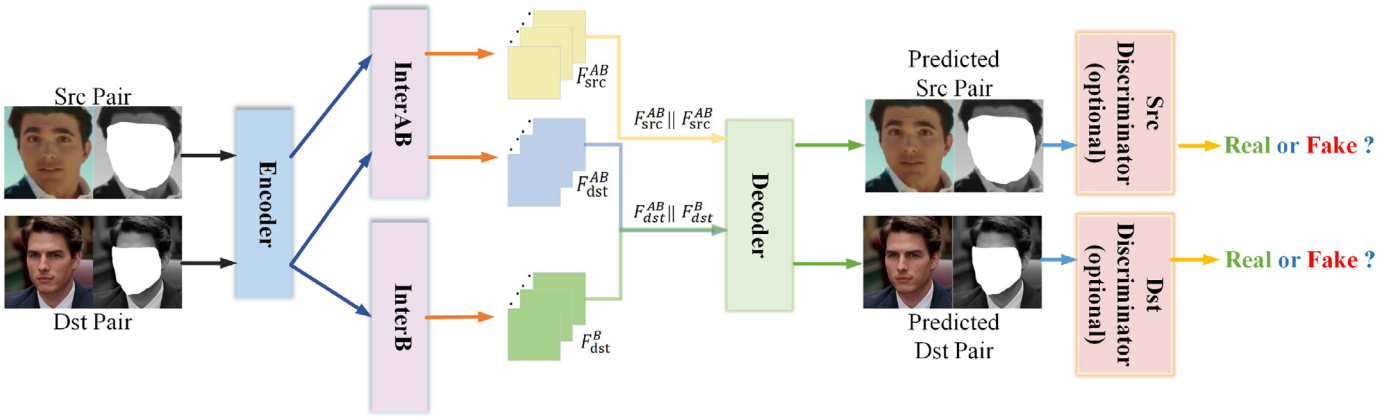

Fig. 5. Overview of LIAE structure. ○‖○ represents the concatenation of latent vectors.

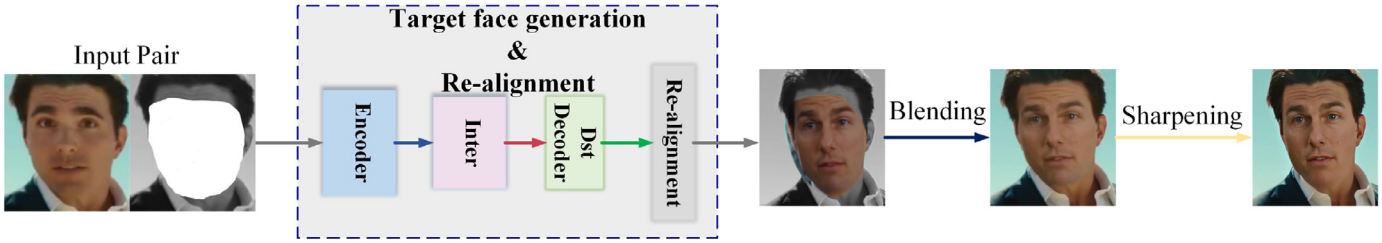

Fig. 6. Overview of mergence phase in DeepFaceLab (DFL).

of the target image in src due to the reversibility of Umeyama [25]. The following piece is blending, with the ambition for the re-aligned reenacted face to seamlessly fit with the target image along its outer contour. To remain consistent complexion, DFL provides five more color transfer algorithms (i.e., Reinhard color transfer: RCT [31], iterative distribution transfer: IDT [32] and etc.) to approximate the color of the reenacted face to the target. Any blending must account for different skin tones, face shapes, and illumination conditions, especially at the junctions between reenacted face with the delimited region and target face. DFL implemented this by Poisson blending [33]. Finally, sharpening is indispensable. A pre-trained face super-resolution neural network was added to sharpen the blended face since it is noted that the generated faces in almost current state-of-the-art face-swapping works, more or less, are smoothed and lack minor details (i.e., moles, wrinkles).

Besides, there is also an automatic post-processing method, NICE, which could greatly enhance the final results of DFL [34]. Briefly, it leverages a U-Net to replace the whole swapping network so that the results could inherit more attributes from the dst, as shown in Fig. 7. However, it requires extra training, more details could be referred to [34].

4. Evaluation

In this section, we compare the DeepFaceLab with several other common face-swapping frameworks and two representative works. We find that DFL has competitive performance among them under identical experimental conditions. It is noteworthy that all experiments are conducted in case of insufficient data samples for DFL. The DFL can support hundred-thousand-scale dataset and provide vivid results. More video results can be seen in supplemental materials video. We use Adam optimizer (lr=0.00005, $\beta_1 = 0.5$, $\beta_2 = 0.999$) to optimize our model. All of our networks were trained on a single NVIDIA GeForce GTX 1080Ti GPU and an Intel Core i7-8700 CPU.

4.1. Qualitative results

Fig. 8 offers face-swapping results of representative open-source projects (DeepFakes [1], Nirkin et al. [2] and Face2Face [35]) taken from FaceForensics++ dataset [36]. Examples of different expressions, face shapes, and illuminations are selected in our experiment.

Previous experiments show that DFL could achieve comparable results with insufficient data. However, with the help of massive data, DFL could achieve cinema level face-swapping. As shown in Fig. 9, the faces can be well swapped into different identities and achieve cinema-level results. Even virtual human could be regarded as a illegal identity which could be used for face swapping. With the help of XSeg, we can swap faces under occlusion. With the support of abundant data, extreme-pose results are also well generated as shown in Fig. 10.

4.2. Quantitative results

We compare our results with videos of FaceForensics++ in quantitative experiments. In practice, the naturalness and realness of the results of the face-swapping method are hard to describe with some specified quantitative indexes. However, pose and expression indeed embodies valuable insights of the face-swapping results. Besides, SSIM is used to compare the structure similarity as well as perceptual loss [37] is adopted to compare high-level differences between the target subject and the swapped subject.

To measure the accuracy of the pose, we calculate the Euclidean distance between the Euler angles (extracted through FSA-Net [38]) of I_t and I_{output}. Besides, the accuracy of the facial expression is measured through the Euclidean distance between the 2D landmarks (2DFAN [23]). We use the default face verification method of DLIB [39] for the comparison of identities.

To be statistically significant, we compute the mean and variance of those measurements on the 100 frames (uniform sampling over time) of the first 500 videos in FaceForensics++, averaging

Fig. 7. Overview of NICE mergence. Compared with simple mergence, more attributes of dst, i.e. gaze direction and motion blur, could be preserved accurately with the help of NICE.

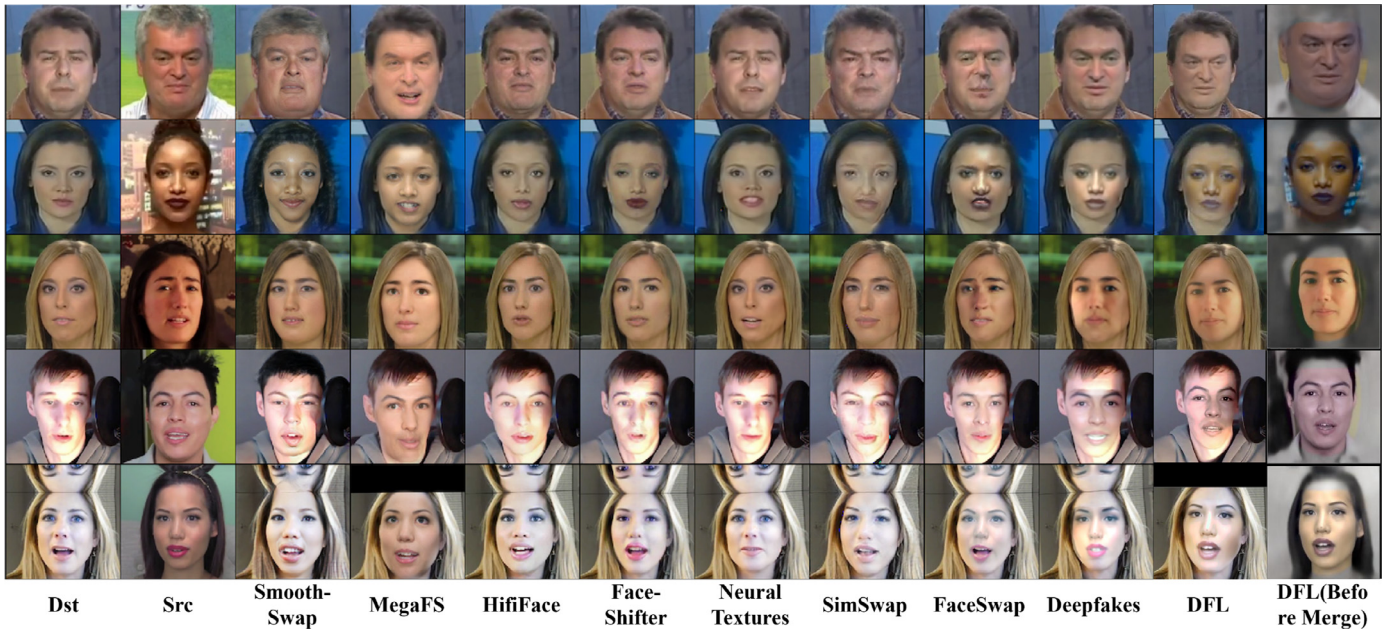

Fig. 8. Comparison of the face-swapping results of various models on the FaceForensics++ dataset. The results from our models show the most consistent identity and shape change, reflecting the characteristics of the source identities. To further show the consistency, we also provide the closest image from source and our before-merge results.

Table 1
Quantitative face swapping results on FaceForensics++ [36] face images.

Method	SSIM ↑	perceptual loss ↓	verification ↓	landmarks ↓	pose ↓
DeepFakes	0.71 ±0.07	0.41 ±0.05	0.69 ±0.04	1.15 ±1.10	4.75 ±1.73
Nirkin et al.	0.65 ±0.08	0.50 ±0.08	0.66 ±0.05	**0.35 ±0.18**	6.01 ±3.21
DFL (ours)	**0.73 ±0.07**	**0.39 ±0.04**	**0.61 ±0.04**	0.73 ±0.36	**1.12 ±1.07**

them across the videos. Here, DeepFakes [1] and Nirkin et al. [2] are chosen as the baselines to compare. It should be noted that all the videos produced by DFL were followed by the same settings with 4.1.

From the indicators listed in Table 1, DFL is more adept at retaining pose and expression than baselines. Besides, with the empowerment of super-resolution in the mergence phase, DFL often produces I_{output} with vivid eyes and sharp teeth, but this phenomenon couldn't be reflected clearly in the SSIM-like score for they only take a small part of the whole face.

4.3. Ablation study

To compare the visual effects of different model choices, GAN settings and etc., we perform several ablation tests. The ablation

Fig. 9. Cinema-level results of DFL. DFL could achieve significant performance of face swapping. With the help of XSeg, DFL could solve samples under heavy occlusion. To protect the right of portraits, we don't involve the *Src* and *Dst* here.

Fig. 10. Extreme-pose results of DFL. With the support of more than 10,000 portraits of single identity, the results show strong ability to deal with extreme poses. To protect the right of portraits, we don't involve the *Src* and *Dst* here.

Fig. 11. The ablation experiments of different model structures (with GAN and TrueFace). (Here, we provide training previews instead of the converted faces, which aims to make a fair comparison in model architectures of DFL meanwhile avoid the impact of post-processing from the conversion phase.).

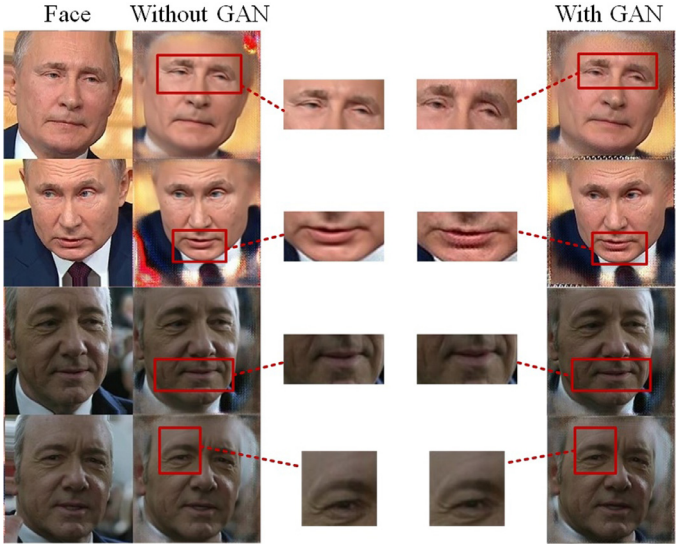

Fig. 12. Ablation experiments of different training paradigms: non-GAN-based and GAN-based (The image on the left is the original face, a reconstruction image produced by a model that trained without GAN listed to its right, far right is produced by a model trained with GAN). GAN enforces the model to become more sensible in capturing the sharp details, i.e., wrinkles and moles. Meanwhile, it significantly reduces the vagueness compared to the model without the empowerment of GAN.

study is conducted on top of three essential parts: network structure, training paradigm, and latent space constraint.

Aside from DF structure and LIAE structure, we also enhance them to DFHD and LIAEHD by adding more feature extraction layers and residual blocks than the original version, which enriches the model structures for comparison. The qualitative results of different model structures can be seen in Fig. 11, and the qualitative results of different training paradigms are depicted in Fig. 12. As shown in Fig. 11, we can see that LIAE can inherit a well-shaped face shape from dst and generate more advanced results than DF, which solves the unmatched face shape problem gracefully.

Moreover, to probe whether the introduction of GAN works in DFL, we compare GAN-based with non-GAN-based in Fig. 12, it is apparent that the details of the face are more realistic and lumpy than non-GAN-based generated.

Quantitative ablation results are reported in Table 2. The experiment settings of the training are almost the same as previous settings except for the structure of the model.

Verification results from Table 2 show that source identities are preserved across networks with the same structure. With more shortcut connections added to the model (i.e., DF to DFHD, LIAE to LIAEHD), scores of landmarks and pose decrease without GAN. Meanwhile, the generated results could have a better chance to get rid of the influence of the source face.

Also, we found that TrueFace is effectively relieved the instability of GAN, through which a more photo-realistic result without much degradation is then achieved. Besides, SSIM progressively increases with more shortcut connections, TrueFace and GAN also do good to it in varying degrees.

5. Discussion

5.1. Integrity

Previous methods often lack enough integrity. As mentioned before, DFL consists of three main phases, extraction, training, and mergence. Each phase plays an indispensable role and has various kinds of alternative techniques. Thanks to the long development progress, DFL has become the most mature face-swapping system in the world. Taking face segmentation techs as examples, we provide several alternative face segmentation options.

Considering the complex requirements for face segmentation, DFL provides an automatic algorithm, TernausNet [26] as default. As mentioned in Section 3.1, TernausNet can remove irregular occlusions efficiently. However, this model may fail to generate fine-grained masks in some particular shots. So we develop a high-efficiency face segmentation tool, XSeg, which allows everyone to customize to suit specific requirements by few-shot learning. With the help of XSeg, users can define the swapping masks by training a customized segmentation model in the training and mergence phase. It is noteworthy that DFL is not a simple combination of current state-of-the-art methods. Instead, the most efficient tools are developed by ourselves according to users' requirements. Besides, we also provide manual extraction methods, which allow users to extract faces in transition frames.

5.2. Potential

Besides, previous methods always lack potential. Previous methods usually focus on synthesizing high-quality results by feeding two videos or hundreds of images. However, making good face-swapping results in this manner underestimate the user's ability. DFL supports mega-scale datasets, up to ~100k images. With the help of enormous data, final swapped results can achieve signif-

Table 2
Quantitative ablation results on FaceForensics [36] face images.

Method	SSIM ↑	verification ↓	landmarks ↓	pose ↓
DF	0.73 ± 0.07	0.61 ± 0.04	0.73 ± 0.36	1.12 ± 1.07
DFHD	0.75 ± 0.09	0.61 ± 0.04	0.71 ± 0.37	1.06 ± 0.97
DFHD (GAN)	0.72 ± 0.11	0.61 ± 0.04	0.79 ± 0.40	1.33 ± 1.21
DFHD (GAN + TrueFace)	0.77 ± 0.06	0.61 ± 0.04	0.70 ± 0.35	0.99 ± 1.02
LIAE	0.76 ± 0.06	**0.58 ±0.03**	0.66 ± 0.32	0.91 ± 0.86
LIAEHD	0.78 ± 0.06	0.58 ± 0.03	**0.65 ±0.32**	0.90 ± 0.88
LIAEHD (GAN)	0.79 ± 0.05	0.58 ± 0.03	0.69 ± 0.34	1.00 ± 0.97
LIAEHD (GAN + TrueFace)	**0.80 ±0.04**	0.58 ± 0.03	0.65 ± 0.33	**0.83 ±0.81**

Fig. 13. The limitation of DFL. *Src* and *Dst* denote the source data and the destination data, $Recon_Src$ and $Recon_Dst$ denote the reconstruction results of the DFL, *Swapped* denotes the swapped results of DFL. The swapped results prefer to generate data which belong to the *Src* domain. When the data of *Src* is insufficient, the quality of swapped results will deteriorate.

icant quality improvements. With the help of several customized tools, DFL can provide face-swapping and support lip manipulation, head replacement, de-age, and more. Users can obtain these functions by slightly finetuning our framework. We also encourage users to use video editing tools for post-processing, such as DaVinci Resolve and After Effect, to enhance the final video's visual quality further.

Since the attention deepfake-related productions received grew exponentially, DFL, the most commonly used deepfake generation tool for VFX artists, has played an irreplaceable role. The emergence of DFL certainly adds entertainment to the world meanwhile of high economic value in the post-production industry when it comes to replacing the stunt actor with pop stars. Besides, It is not a joke that replacing actors completely with Deepfake technology using an impersonator, further cut studio costs. DFL provides the potential to create new actors, where producers can choose elements they like from multiple actors, such as the body movement of one, the facial expressions of another, and the voice of another actor, creating an entirely computer-generated actor. This idea could fulfill any director's dream while eliminating the high cost of established actors.

5.3. Model fingerprints

As demonstrated in previous research, machine learning models always prefer to generate results with specific fingerprints. On social media, attributing images to their sources can potentially deter malicious organizations and hold them accountable by leading legal proceedings [40]. To further allow the traceability of generated results, we take model watermark into consideration when we release models or datasets [41].

For the shared dataset, the previous method proposes to embed artificial fingerprints into the training data and shows a surprising discovery on the transferability of such fingerprints from training data to final models, which in turn enables reliable detection and attribution of deepfakes [42].

For the released pretrained model, the model could be distributed with a user-specific watermark. With the help of a watermark, we could trace the producer of malicious face-swapping media [43].

Nevertheless, it is important to take into account adversarial attack and defense, as these could have a significant effect on extracting model fingerprints [44–47].

5.4. Limitations

Comparing to previous methods, DFL is a multi-shot face-swapping system and the quality of final results depends in part on the quality of input data. As shown in Fig. 13, the results generated by DFL cannot accurately catch the gazing direction and always show right-looking. This is because the source data is extracted from one single right-looking video and DFL don't provide any facial prior in the architecture. To solve this problem, we can involve much more data of the source identity.

5.5. Broader impact

Any Face swapping algorithm runs the risk of producing biased or unsuitable content—our work is undoubtedly not an exception. We have to admit that deepfake techniques may affect public discourse quality and infringe upon the citizens' portrait rights, primarily since deepfake may be used maliciously as a source of misinformation, manipulation, harassment, and persuasion.

However, inspired by some distinguished researchers in this area: "Suppressing the publication of such methods would not stop their development, but rather make them only available to a limited number of experts and potentially blindside policymakers if it goes without any limits," we believe that it's our responsibility to publish DeepFaceLab to the academia community formally.

Though several methods, such as model watermark technologies, were proposed to reduce the negative effects caused by our work from the technical side, the negative influences still exist. In order to eliminate such negative effects thoroughly, we encourage DFL users to generate deepfake-entertainment content and spread them to raise the netizens' alerts to deepfake technology. According to the data from Youtube, deepfake-entertainment content has gained more than 100,000,000 clicks. We should not focus on the potential malicious use of deepfake techniques and ban deepfake-related topics. Instead, we could promote friendly deepfake-entertainment content to general netizens on social media, reminding them of the existence of deepfake when they are watching these videos.

6. Conclusion

The rapidly evolving DeepFaceLab has become a popular face-swapping tool in the deep learning practitioner community by freeing people from laborious, complicated data processing, trivial detailed work in training, and mergence phase. We choose to remind the public by using entertainment videos to avoid the destructive influence of deepfake. As more and more people participate in the development of DeepFaceLab, deepfake-entertainment content has been trending in social media. With the increase of deepfake-entertainment content, we hope general netizens can realize the existence of deepfake and be alert to their potential malicious uses.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledge

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62121002, 62072421, 62002334 and Key Research and Development program of Anhui Province under Grant 2022k07020008.

Appendix A. Architecture of LIAEHD

The layout as well as every specified submodule of the **DF** are depicted in Fig. A.14. According to the result, it's fairly easy to see the difference between the original **DF** and enhanced edition **DFHD** lies in that **DFHD** have more feature extraction layers and of varied stacking orders. Three typical traits of the structure are:

- We use pixelshuffle (depth2space) to do upsampling instead of transposed convolution neither bilinear sampling followed by convolution, which aims to eliminate the artifact and checkboard effects.
- Identity shortcut connection, which derived from Resnet, are frequently used in composing the module of Decoder. This is because model with more shortcut connections always have many independent effective paths at the same time, which makes the model with ensemble-like behaviour.
- We normalize the images between 0 and 1 other than -1 to 1. Then Sigmoid as the last layer of the Decoder output rather than Tanh.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2023.109628

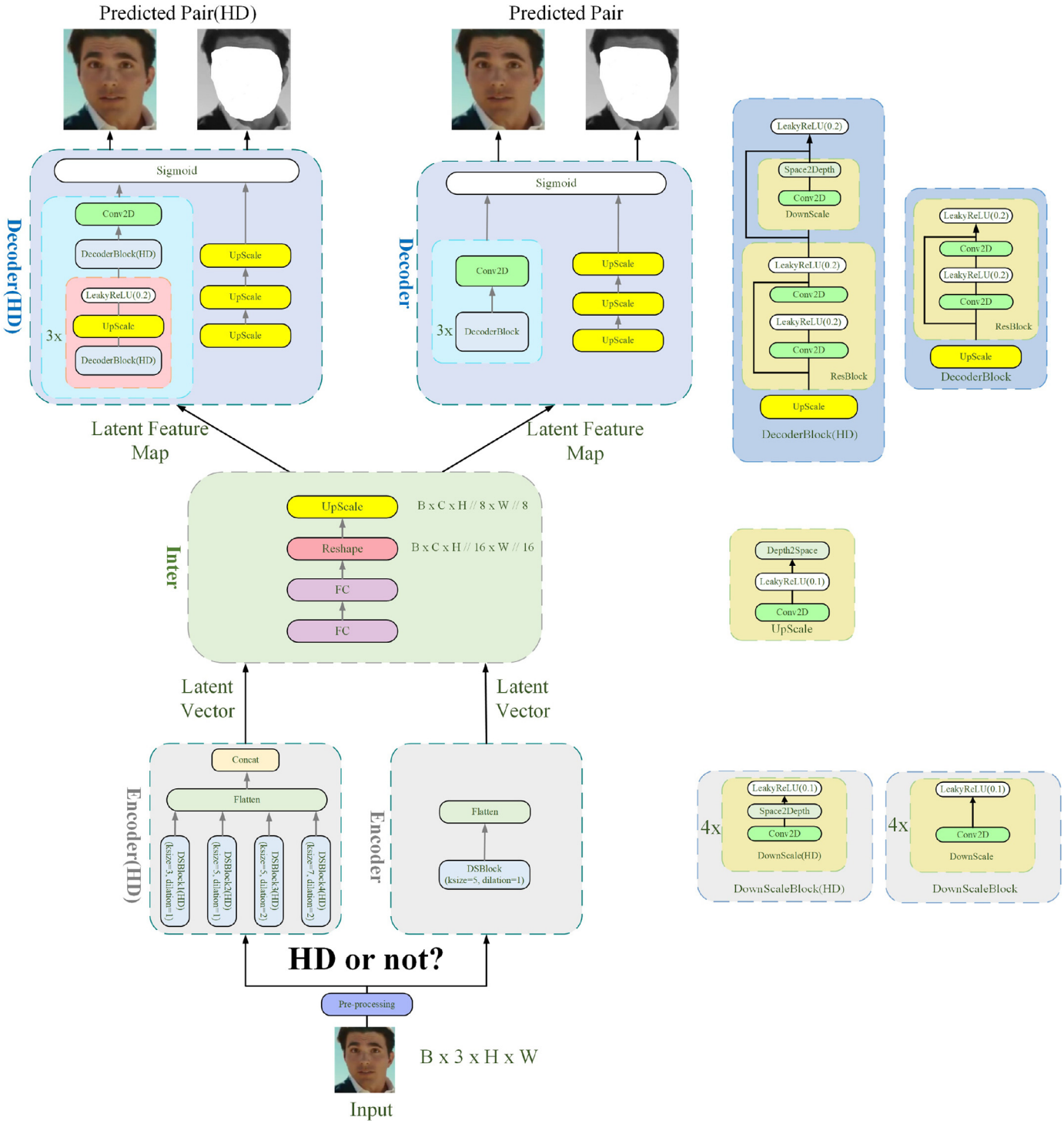

Fig. A1. A detailed overview of **DF** in DeepFaceLab. Modules (Encoder, Inter and Decoder) of **DF** are completely same with **LIAE**, which means both InterAB and InterB of **LIAE** owns the same structure and settings.

References

[1] Deepfakes, Deepfakes, 2017, (https://github.com/deepfakes/faceswap).

[2] Y. Nirkin, I. Masi, A.T. Tran, T. Hassner, G. Medioni, On face segmentation, face swapping, and face perception, in: IEEE Conference on Automatic Face and Gesture Recognition, 2018.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[4] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[5] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, arXiv preprint arXiv:2106.12423 (2021).

[6] Y. Fang, W. Deng, J. Du, J. Hu, Identity-aware cyclegan for face photo-to-sketch synthesis and recognition, Pattern Recognit. 102 (2020) 107249.

[7] N. Liu, T. Zhou, Y. Ji, Z. Zhao, L. Wan, Synthesizing talking faces from text and audio: an autoencoder and sequence-to-sequence convolutional neural network, Pattern Recognit. 102 (2020) 107231.

[8] S. Zhao, J. Li, J. Wang, Disentangled representation learning and residual gan for age-invariant face verification, Pattern Recognit. 100 (2020) 107097.

[9] DeepFakes(https://github.com/deepfakes/faceswap), 2021.

[10] R. Chen, X. Chen, B. Ni, Y. Ge, Simswap: an efficient framework for high fidelity face swapping, ACM Multimedia (2020), doi:10.1145/3394171.3413630.
[11] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Faceshifter: towards high fidelity and occlusion aware face swapping, arXiv preprint arXiv:1912.13457 (2019).
[12] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: image synthesis using neural textures, ACM Trans. Graph. (TOG) 38 (4) (2019) 1–12. Comment: Video: https://youtu.be/z-pVip6WeyY SIGGRAPH 2019
[13] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, R. Ji, HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 1136–1142, doi:10.24963/ijcai.2021/157.
[14] Y. Zhu, Q. Li, J. Wang, C.-Z. Xu, Z. Sun, One Shot Face Swapping on Megapixels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4834–4844.
[15] K. Jiseob, L. Jihoon, Z. Byoung-Tak, Smooth-swap: a simple enhancement for face-swapping with smoothness, arXiv preprint arXiv:2112.05907 (2021).
[16] Z. Shang, H. Xie, Z. Zha, L. Yu, Y. Li, Y. Zhang, Prrnet: pixel-region relation network for face forgery detection, Pattern Recognit. 116 (2021) 107950, doi:10.1016/j.patcog.2021.107950.
[17] W. Pu, J. Hu, X. Wang, Y. Li, S. Hu, B. Zhu, R. Song, Q. Song, X. Wu, S. Lyu, Learning a deep dual-level network for robust deepfake detection, Pattern Recognit. 130 (2022) 108832, doi:10.1016/j.patcog.2022.108832.
[18] H. Chen, Y. Li, D. Lin, B. Li, J. Wu, Watching the big artifacts: exposing deepfake videos via bi-granularity artifacts, Pattern Recognit. 135 (2023) 109179, doi:10.1016/j.patcog.2022.109179.
[19] X. Lin, S. Wang, J. Deng, Y. Fu, X. Bai, X. Chen, X. Qu, W. Tang, Image manipulation detection by multiple tampering traces and edge artifact enhancement, Pattern Recognit. 133 (2023) 109026, doi:10.1016/j.patcog.2022.109026.
[20] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The deepfake detection challenge (dfdc) dataset, 2020, arXiv:2006.07397
[21] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, S.Z. Li, S3fd: single shot scale-invariant face detector, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 192–201.
[22] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, S. Zafeiriou, Retinaface: single-stage dense face localisation in the wild, arXiv preprint arXiv:1905.00641 (2019).
[23] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.
[24] Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, Joint 3d face reconstruction and dense alignment with position map regression network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 534–551.
[25] S. Umeyama, Least-squares estimation of transformation parameters between two point patterns, IEEE Trans. Pattern Anal. Mach. Intell. (1991) 376–380.
[26] V. Iglovikov, A. Shvets, Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation, arXiv preprint arXiv:1801.05746 (2018).
[27] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[28] A. Loza, L. Mihaylova, N. Canagarajah, D. Bull, Structural similarity-based object tracking in video sequences, in: 2006 9th International Conference on Information Fusion, IEEE, 2006, pp. 1–6.
[29] A. Aghajanyan, Convolution aware initialization, arXiv preprint arXiv:1702.06295 (2017).
[30] H. Lin, W. Zeng, X. Ding, Y. Huang, C. Huang, J. Paisley, Learning rate dropout, arXiv preprint arXiv:1912.00144 (2019).
[31] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (5) (2001) 34–41.
[32] F. Pitié, A.C. Kokaram, R. Dahyot, Automated colour grading using colour distribution transfer, Comput. Vis. Image Understand. 107 (1–2) (2007) 123–137.
[33] P. Pérez, M. Gangnet, A. Blake, Poisson Image Editing, in: ACM SIGGRAPH 2003 Papers, 2003, pp. 313–318.
[34] K. Liu, P. Wang, W. Zhou, Z. Zhang, Y. Ge, H. Liu, W. Zhang, N. Yu, Face swapping consistency transfer with neural identity carrier, Future Internet 13 (11) (2021), doi:10.3390/fi13110298.
[35] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: real–time face capture and reenactment of rgb videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2387–2395.
[36] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: learning to detect manipulated facial images, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1–11.
[37] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European conference on computer vision, Springer, 2016, pp. 694–711.
[38] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, Y.-Y. Chuang, Fsa-net: learning fine-grained structure aggregation for head pose estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1087–1096.
[39] D.E. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (Jul) (2009) 1755–1758.
[40] S. Girish, S. Suri, S. Rambhatla, A. Shrivastava, Towards discovery and attribution of open-world gan generated images, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14074–14083, doi:10.1109/ICCV48922.2021.01383.
[41] J. Zhang, D. Chen, J. Liao, H. Fang, W. Zhang, W. Zhou, H. Cui, N. Yu, Model watermarking for image processing networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 12805–12812.
[42] N. Yu, V. Skripniuk, S. Abdelnabi, M. Fritz, Artificial gan fingerprints: Rooting deepfake attribution in training data (2020).
[43] J. Zhang, D. Chen, J. Liao, W. Zhang, H. Feng, G. Hua, N. Yu, Deep model intellectual property protection via deep watermarking, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[44] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, D. Tao, Perceptual-sensitive gan for generating adversarial patches, in: Proceedings of the AAAI conference on artificial intelligence, 2019.
[45] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, H. Yu, Bias-based universal adversarial patch attack for automatic check-out, ECCV, 2020.
[46] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, et al., Robustart: benchmarking robustness on architecture design and training techniques, arXiv preprint arXiv:2109.05211 (2021).
[47] C. Zhang, A. Liu, X. Liu, Y. Xu, H. Yu, Y. Ma, T. Li, Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity, IEEE Trans. Image Process. (2021).

Kunlin Liu received his B.S. degree in 2018 from University of Science and Technology in China. He is currently pursuing the Ph.D. degree in electronic engineering in University of Science and Technology of China. His research interests include multimedia manipulation and AI security.

Ivan Perov is a machine learning enthusiast and devote himself into multimedia manipulation. He is the owner of a famous github project, DeepFaceLab.

Daiheng Gao is a freelancer who has been working on face swapping technology.

Nikolay Chervoniy is a freelancer who has been working on face swapping technology.

Wenbo Zhou received his B.S. degree in 2014 from Nanjing University of Aeronautics and Astronautics, China, and Ph.D. degree in 2019 from University of Science and Technology of China, where he is currently postdoctoral researcher. His research interests include information hiding and AI security.

Weiming Zhang received his M.S. degree and Ph.D. degree in 2002 and 2005 respectively from the Zhengzhou Information Science and Technology Institute, P.R. China. Currently, he is a professor with the School of Information Science and Technology, University of Science and Technology of China. His research interests include multimedia security.