Survey and enhancement on DeepFaceLab

A0247473X Li JiaLin

Abstract

Deep fake video has been more and more popular on entertainment with the improvement of fake video quality. DeepFaceLab is one of the most powerful deep fake frameworks, which provides 95% current fake videos on the website with cinema-quality results. However, our survey shows missing of training data still severely restricts the fidelity of result image. In this work, we do data argumentation to DeepFaceLab to enhance the model.

Keywords

Deep fake video; DeepFaceLab; Data augmentation

Directory

1 Introduction

1.1 Deep Fake video

Recently, fake videos have become one of the most popular types of fake media, which includes image-level face swapping, human expression transfer, fake pictures generation and so on. Fake videos mostly specialize to face swapping fake video. While the most basic fake videos just crop and exchange faces, the more powerful one can make it entirely indistinguishable from human vision system, modify emotions and even put words and actions in the mouth and other body parts. Such “deep” hybridized or generated videos are so called “Deep fake videos” which use a series of machine learning, AI and neural network techniques.

The first widely known examples AI-manipulated face-swapped video appeared in November, 2017. A Reddit user with the user name “deepfakes” uploaded some videos grafting faces of famous female actors like Gal Gadot and Scarlett Johansson onto other actors’ bodies. In 2019, artists Bill Posters and Daniel Howe published a deep fake video of Mark Zuckerberg talking about amassing power and control. [1] The video caused a sensation for it looked like an almost real video. Someone is fascinated with the infinite possibilities of deep fake videos while others feel worried about its potential damage to the society.

Based on the interests of the public nowadays, deep fake videos mainly focus towards two directions: fake video detection and forgery. Deep fake video forgery works on making fake video more realistic and more effective to generate and detection algorithms target on protecting people from false propaganda. For example,

distinguish popular funny fake video of famous politicians like Donald Trump and Obama from their real speeches.

In this work, we pay attention to the famous algorithms and public open-source repository of deep fake video forgery. We will do surveys to evaluate and try to improve.

1.2 DeepfaceLab

Deepfacelab is one of the most popular frameworks on deep fake video forgery, which creates more than 95% deep fake videos on the Internet. [2] DeepfaceLab is very powerful. It supports a series of AI face manipulation method, from the most basic swapping face to de-aging face, replacing the head and even manipulating politicians lips. Thanks to the state-of-art pipeline, DeepfaceLab produces realistic swapping face videos but requires far less amount of training data than other related work. What's more, DeepfaceLab does well on the software engineering level. It integrates several kinds of fantastic algorithms into each part of the model so that people can switch them easily. It also can be easily installed, adapt most of the develop environment and support multi-GPU and multi-CPU.

According to comprehensive evaluation, DeepfaceLab can be valued as one of the best deep fake video frameworks. So in this work, we will do some experiments to evaluate the performance of DeepfaceLab under possible circumstances and conduct insufficient points from the whole pipeline where we can try to make up.

1.3 StyleGAN2

StyleGAN2 is an amazing model developed by Nvidia Lab. [3] Although it costs very high to train such a model to a relatively perfect state, StyleGAN2 performs perfectly on generating not only human face images, but also car, animal and house, etc. It can generate 1024*1024 face images with high resolution. The generated human face is entirely indistinguishable from our eyes. The emotion is very realistic. In this work, StyleGAN2 is trained to generate possible fake images.

2 Model

In this work, we focus on the most popular category of deepfake: face-swapping and fake video. Except the pre and post processing, the basic face-swapping pipeline can be divided into three phases: extraction, training and conversion. Besides, DeepfaceLab do a one-to-one face swapping. The data is only composed of src and dst, i.e. the faces of source images will replace the faces of destination images and retain the background of destination images.

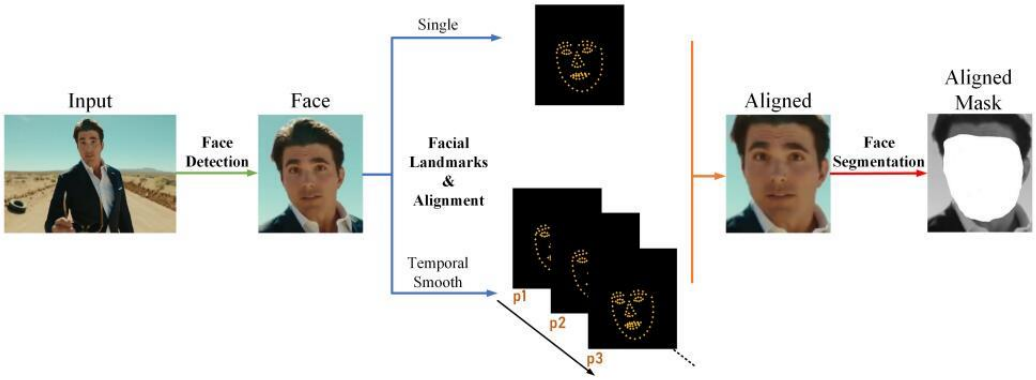

Figure 1 [4]

2.1 Extraction

Figure 1 shows the basic process of face extraction, which involves many algorithms and models. The first step is face detection aiming to find faces in the picture. DeepFaceLab use S3FD [5] as the default detector. The second step is face alignment between src and dst images. In most of the excellent works of face alignment, face landmark is the critical point. Face landmark is those critical points on human face to represent the whole face. Face alignment is actually the alignment of face landmarks, i.e. to minimize the distance loss between landmarks of src and dst images. The whole face is aligned when landmarks are aligned. DeepfaceLab provides two canonical types of extraction algorithms: heat-map based facial landmark algorithm 2DFAN [6] and PRNet [7], which works better when faces have large Euler angle. The last step is face segmentation. This step crops only face from the picture according to the landmarks so that we can maintain the hair style of the dst images. DeepFaceLab use TernausNet [8] to do the segmentation. After extraction, we can focus AI

manipulation on only face.

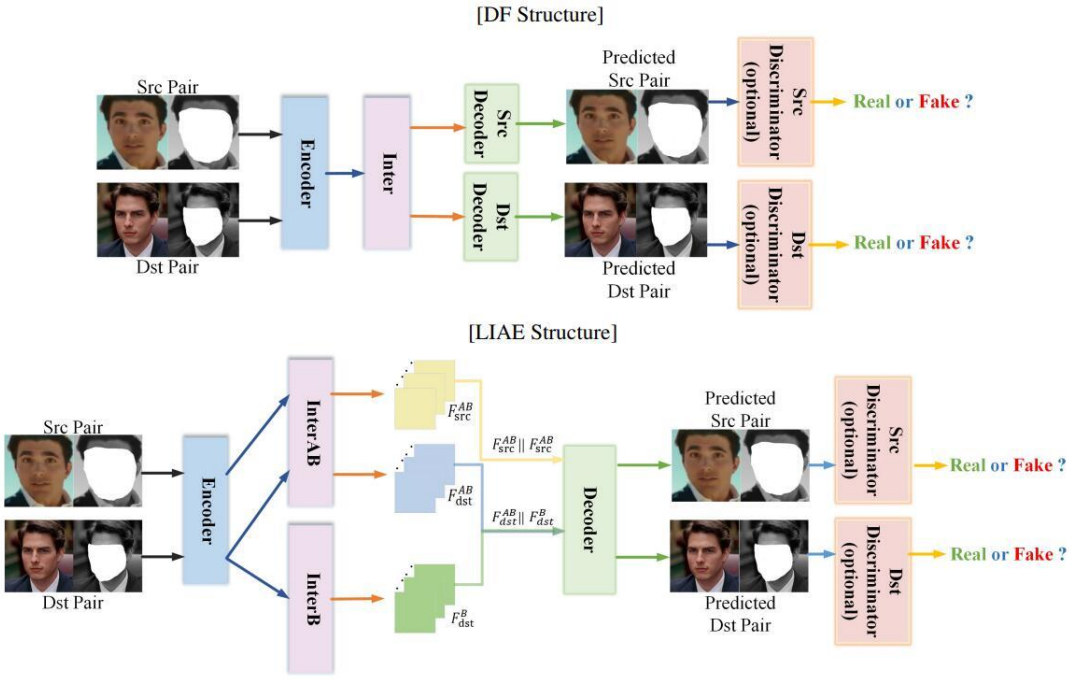

Figure 2 [9]

2.2 Training

Training phase is the most vital part in the whole pipeline. As shown in above picture Figure 2, DeepFacLab provides two model structures to choose: DF structure and LIAE structure, which is more complex but more robust to light consistency. The train process is easily to be explained in Figure 3. In this circumstance, we try to replace the face of src image (Obama) to dst image (one of my best friends HanWei). In the training process, we separately train in two pipelines, one in src and the other in dst. The encoder actually works like a neural network, which extract features from the image automatically. It projects the information of image to the latent space.

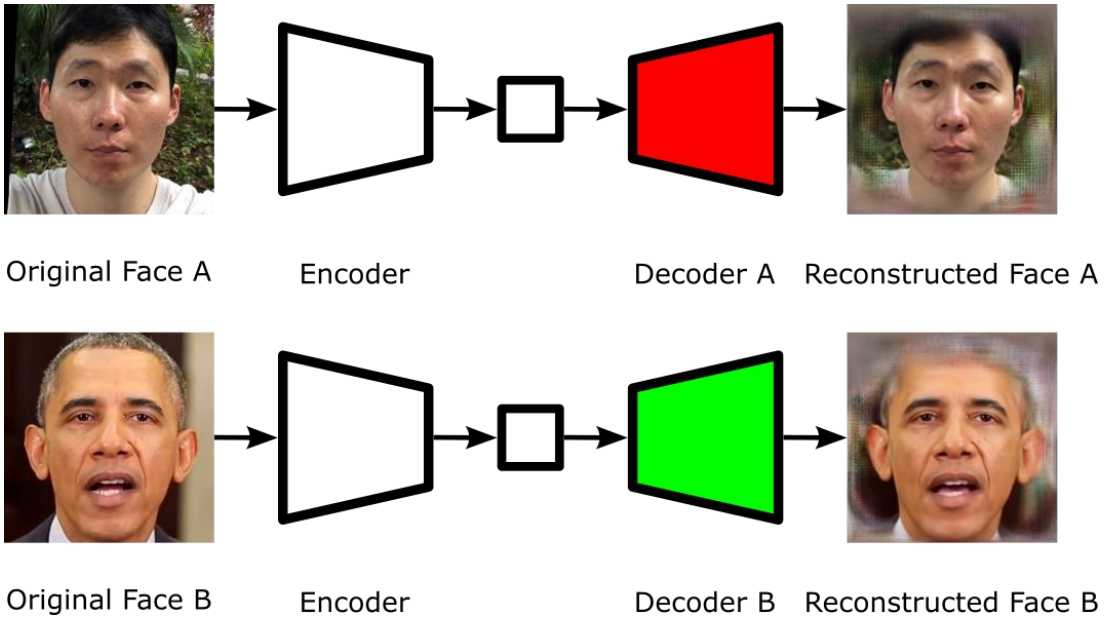

Figure 3

2.3 Conversion

This stage is the last but unignorable step. Figure 4 shows the core of conversion: we join src's decoder with dst's encoder. Using information encoded by dst's encoder, decoder of src tries to reconstruct a face with src's style. In this way can we obtain a swapping face with abundant details and realistic emotion. After the basic conversion, we still need to do blending and sharpening work to make the final result better.

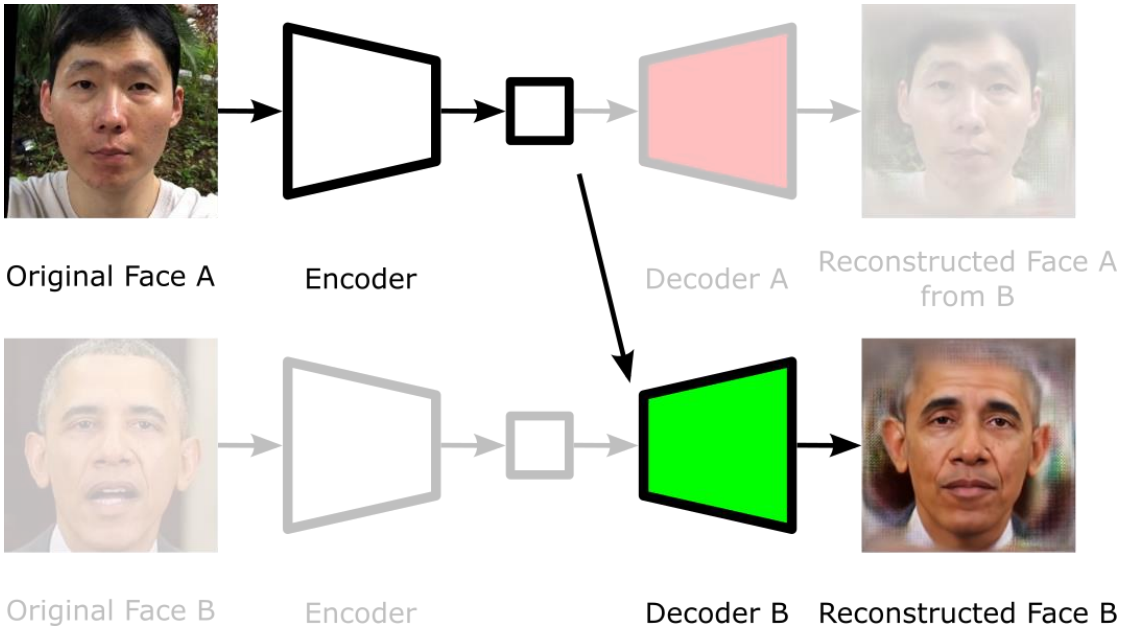

Figure 4

3 Experiment

In this part, we do experiments on DeepFaceLab to survey the possible drawbacks of DeepFaceLab where we can try to make up.

3.1 Data and develop environment

The src data is images of American president Obama from videos of public speech on Youtube [10]. The dst data is images of me and HanWei. That is to say, replace Obama's face to HanWei's face. All the images are the segmentation of frames of video using ffmjpeg. We try to find videos on the website with higher resolution so that images will be more realistic. However, it is not easy to find video of politicians that is clear enough. Finally, I only get 36000 images from 3 high resolution speeches or interviews of Obama on Youtube. The dst images are around 600 images.

Test environment is on a single NVIDIA A100 GPU with 40GB VRAM.

3.2 Test

The main reasons that influence the generate video quality are: amounts of src and dst images, training iterations of model, distribution of src and dst images. Figure 5 shows the result under only 300 images of Obama src.

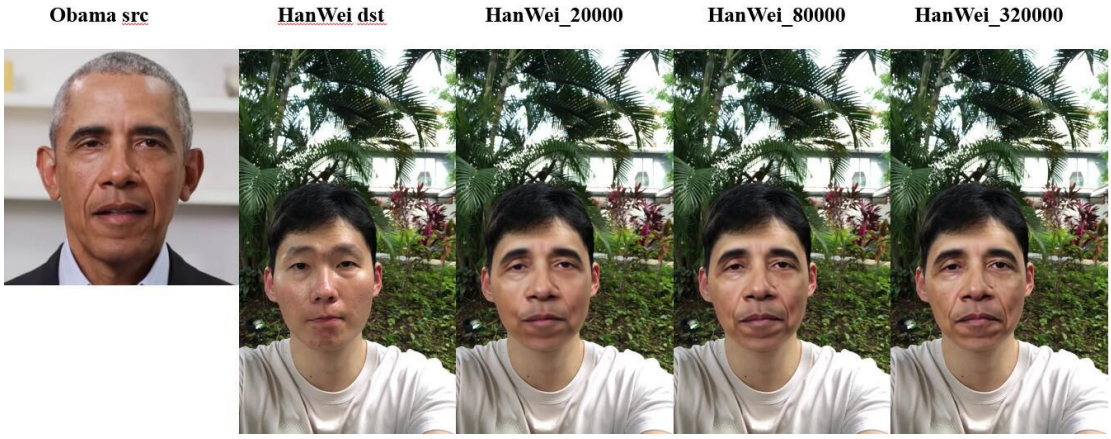

Figure 5

DeepFaceLab shows excellent performance under mini dataset. And we can conclude from the test of iterations that in most case the more iterations the model is trained, the clearer the result image is. However, we cannot exclude the possibility that model is over-fitted.

Figure 6

Figure 6 shows the results that are not that good. The first problem is it fails to imitate the movement of eyes, generated eyes always see straight forward. The second problem is when the emotion or posture turns to smile and side face, it becomes much more fuzzy. It is clear that the part near the mouth and teeth is fuzzy. And the edge of side face is also blurry.

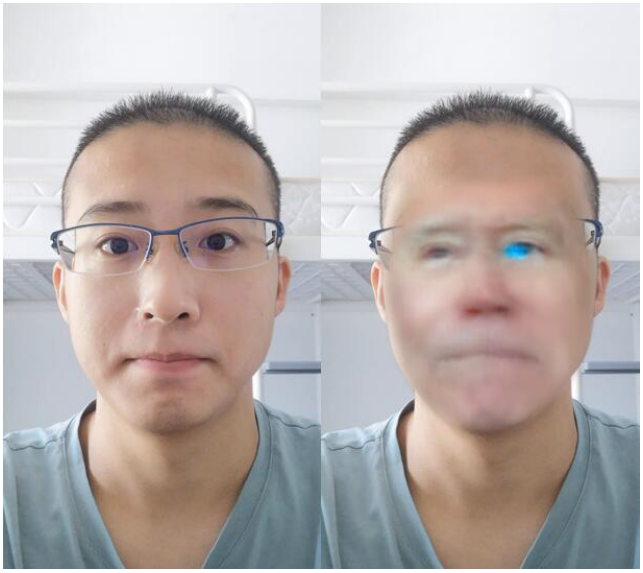

Figure 7

Figure 7 shows the result of no-training replacement from President Biden to my face. The image is very blur so a one-shot or zero-shot training cannot be achieved. What's more, we can find face-swapping cannot support glasses on my face.

3.3 Analyze and possible improve aspects

From figure 5, we can find DeepFaceLab is powerful enough to generate good result images as long as we have enough training iterations. From figure 6, we can find DeepFaceLab performs bad on the emotions and side face. The most possible reason is images that contain smile and side faces distributes too little in the whole training set.

So one of the enhance methods is to find more images which contain emotions and side faces. From figure 7, we can find DeepFaceLab doesn't support one-shot or zero-shot training and glasses is ignored. So change model to support them is a possible direction.

4 Enhancement

In this work, I choose the direction of data augmentation, which tries to use more training data to enhance the model. Data augmentation has two basic directions: find more real data or generate simulated data. Since it's very hard to find politicians video on the web that is clear enough, generate "fake" data to simulate is better.

I choose StyleGAN2 [11] to generate more Obama's faces. The result is as follow. The pictures which don't contain Obama's face are the failed ones.

Figure 8

Then I randomly generate 7000 images and add them into the training set. We have 43000 images in total now. I train the model of DeepFaceLab again but the result is not very good:

Figure 9

It is still somehow blurry. I try to use Fréchet Inception Distance (FID) to do some quantitative evaluation. The equation to compute FID is:

$$d^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + Trace(C + C_w - 2(CC_w)^{1/2})$$

where m and C are the mean value and covariance of the probability distribution of the result pictures. m_w and C_w are the mean value and covariance of the probability distribution of the original pictures.

The result indicates that although we don't do the enhancement fully, the pictures after enhancement get an improvement on the image quality:

Original picture	Result picture	FID
HanWei	Obama	539.48
HanWei	HanWei face swapping	72
HanWei	HanWei face swapping after enhancement	58

5 Future Work

First, I will continue to generate more src and dst data to test. I can also try to specifically generate side faces and smile faces.

Second, inspired by StyleGAN2-ADA [12], I think a smarter adoptive data augmentation (ADA) will work better. ADA is an algorithm to do augmentation like flip and rotation based on original data, which aims to promote model convergence. This direction falls on the usage of existing real data.

Third, we lack a good quantitative standard to score the fake video. DeepFaceLab use comparative result with other similar work but it's not convincible. So in the next stage, I prompt to find a useful quantitative score standard to help.

6 Reference

[1] Deep fakes and cheap fakes, The Manipulation of Audio and Visual Evidence. Britt Paris, Joan Donovan. SEPTEMBER 18 2019 . Page 1 https://datasociety.net/library/deepfakes-and-cheap-fakes/

[2] Perov I , Gao D , Chervoniy N , et al. DeepFaceLab: A simple, flexible and extensible face swapping framework[J]. 2020. https://github.com/iperov/DeepFaceLab

[3] Karras T , Laine S , Aittala M , et al. Analyzing and Improving the Image Quality of StyleGAN[C] 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020. https://github.com/NVlabs/stylegan2

[4] Perov I , Gao D , Chervoniy N , et al. DeepFaceLab: A simple, flexible and extensible face swapping framework[J]. 2020. Figure 2 Overview of extraction phase in DeepFaceLab (DFL for short). https://github.com/iperov/DeepFaceLab

[5] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 192–201, 2017. 3

[6] Adrian Bulat and Georgios 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*, pages 1021–1030, 2017. 3, 6

[7] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 3

[8] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746, 2018. 4, 8

[9] Perov I , Gao D , Chervoniy N , et al. DeepFaceLab: A simple, flexible and extensible face swapping framework[J]. 2020. Figure 3 Overview of extraction phase in DeepFaceLab (DFL for short)

[10] Youtube video source: https://youtu.be/mAFv55o47ok, https://youtu.be/NGEvASSaPyg, https://youtu.be/25GOnaY8ZCY

[11] Karras T , Laine S , Aittala M , et al. Analyzing and Improving the Image Quality of StyleGAN[C] 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[12] Karras T , Aittala M , Hellsten J , et al. Training Generative Adversarial Networks with Limited Data[J]. 2020.