Any-to-one Face Reenactment Based on Conditional Generative Adversarial Network

Tianxiang Ma*, Bo Peng*, Wei Wang*, Jing Dong*
* National Laboratory of Pattern Recognition, CASIA
Center for Research on Intelligent Perception and Computing, CASIA
University of Chinese Academy of Sciences, Beijing, 100049, China
E-mail: tianxiang.ma@cripac.ia.ac.cn, bo.peng@nlpr.ia.ac.cn, wwang@nlpr.ia.ac.cn, jdong@nlpr.ia.ac.cn

***Abstract*—Face reenactment refers to the process of transferring the expressions and postures of a given face to the target face. We present a novel Any-to-one Face Reenactment Model based on Conditional Generative Adversarial Network, which has a simple dual converter structure: Any-to-one Face Landmarks Map Converter(AFLC) and Landmark-to-face Converter based on Conditional Generative Adversarial Network(LFC). The former transfers any source face into the landmarks map of the target face, and the map has the expression and posture attributes of the source face. The latter has a generator that transfers the landmarks map of the target face into the realistic and identity-preserving target facial image. The whole model is purely learning-based without any 3D model, and can generate high quality transferred face comparable to the state-of-the-art. What's more the model is highly robust to wild faces, including various faces of different complexions, ages, and genders. We performed an ablation study on our proposed AFLC to verify its importance for face reenactment of any object. AFLC helps the overall model to achieve an effective facial reenactment.**

I. Introduction

Face reenactment is a way to transfer the expressions and postures of the source face to the target face and realize facial expression reproduction. Face reenactment has a wide range of applications, such as CGI, virtual reality, video games, video conference and so on.

Face reenactment is essentially an image-to-image translation process, which transfers a source face to a target face. In this transfer process, it is necessary to keep the identity attribute of the target face unchanged, and the target face generated at the same time also needs to have facial attributes such as expressions and postures of source face. So this process is a very complex process that requires fine operations. Thanks to the development of 3D face models, many 3D model-based face reenactment methods [1], [2], [3], [4], [5] have emerged in recent years. These methods greatly improve the accuracy of the face reenactment, but they often require track and optimize the original video, fit complex parameters, and finally render output video. These methods are mostly closed-source and unfavorable for a wide range of applications.

The emergence of Generative Adversarial Network(GAN) provides a new direction for face reenactment. There have been many excellent work on image-to-image translation based on GAN, such as [6], [7], [8], [9]. Although the current GAN has been successful in image-to-image translation, there are still some problems in specific tasks such as face reenactment:

1) First, it is still very difficult to use the purely learning-based approach for end-to-end facial expression transformation. The reason is that limited training samples cannot cover the diversity of complex facial expressions and postures, as well as we can't get facial images that match in pairs. This problem makes the learning-based face reenactment difficult, but there are some successful research progress. For example, [10], [11] used CycleGAN of [7] to achieve one-to-one face reenactment. However, the use of advanced methods such as CycleGAN that do not require paired images often produce very unnatural or erroneous images, and can only be well applied in one-to-one transfer. This leads to the next problem.

2) Second, it is difficult to achieve face reenactment for any kind of faces. The method proposed by [12] implements many-to-one, but it transforms expressions and postures by training a transformer in a constructed faces set. so it doesn't perform good face reenactment on any wild face.

In this paper, we propose a novel Any-to-one Face Reenactment Model based on Conditional Generative Adversarial Network, which can effectively solve the above mentioned problems. Our main contributions are as follows:

- We propose a Landmark-to-face Converter based on Conditional Generative Adversarial Network(LFC). Its generator can transfer the landmarks map of the target face into the realistic and identity-preserving target face. Here the face landmarks map is used as an intermediate medium to connect the source face and the target face, which can solve the image pairing problem.
- We create an Any-to-one Face Landmarks Map Converter(AFLC), whose function is to transfer any source face into the target facial landmarks map while the landmarks map retain the source facial expressions and postures.
- Our model achieve purely learning-based facial expressions and postures transformation, which is able to implement face reenactment for various wild faces, including faces of different complexions, ages, and genders. The whole model can effectively achieve any-to-one face reenactment.

II. Related works

The research work on face manipulation began with the pioneering work of [13], which used a 3D manipulation model to fit a single image, and map the texture of the face to the

parameters space of the facial expression and posture, and then the facial expression of the face is manipulated. Afterwards, there have also been a number of facial manipulation works. For example, [14] introduces a method to manipulate the camera viewpoint from a single input facial image, [15] transfers lip motions to an existing target video, [16] focuses on the refinement of mouth movements. And more and more 3D model-based face capture and track models [17], [18], [19], [20] appear.

These studies of face manipulation have the face reenactment to begin to develop. [5] performs facial reenactment by tracking a face template, [21] proposes an image-based automatic method to replace the entire face, [22] presents a real-time avatar animation system from a single image. Most of those work needs to track the face of the original video or build a 2D or 3D parametric model.

In recent years, due to the extensive application of deep learning in the field of computer vision, some researches have been carried out using deep neural networks for facial manipulation and expression synthesis. [23] Constructs a facial transformation network based on convolutional neural networks, [24] introduces a variational autoencoder to learn the expression flow maps. In addition, the generation model is used to process fine-scale details. Such as the pioneer work in [25].

There is an exciting generation model in the field of deep learning that is inherently suitable for image generation and transformation tasks. This is Generative Adversarial Network(GAN), which proposed in [26]. This type of model has evolved in recent years to produce high-quality images that are realistic and indistinguishable to the naked eye. What's more, GNA is widely used in image style transfer, image super-resolution, image editing, and other application scenarios. The use of the GAN for expression manipulation and generation is also active. [27] trains a deep generative network that can infer realistic per-frame texture deformations, including the mouth interior, [28] proposes an expression generative adversarial network for photo-realistic facial expression editing with controllable expression intensity, [29], [30] uses geometric facial landmarks to guide the network to control facial detail synthesis. For face reenactment, many work use CycleGAN or other image-to-image algorithms for end-to-end transfer such as [10], [11], they only implement one-to-one face reenactment. Recently, some outstanding GAN-based face reenactment work has also appeared. [12] designs a boundary-based transfer to achieve many-to-one face reenactment. [31] introduces a warp-guided generative model for realtime portrait animation.

III. Any-to-one face reenactment based on conditional generative adversarial Network

The proposed framework is depicted in Fig. 1 , this model is a dual converter structure: Any-to-one Face Landmarks Map Converter(AFLC) for transferring any source face into the target facial landmarks map, and Landmark-to-face Converter based on Conditional Generative Adversarial Network(LFC) for generating target face from target facial landmarks map. The forward propagation of the model is a simple procedure that transfers the source face into target face through AFLC and LFC. Below we introduce LFC and then introduce AFLC.

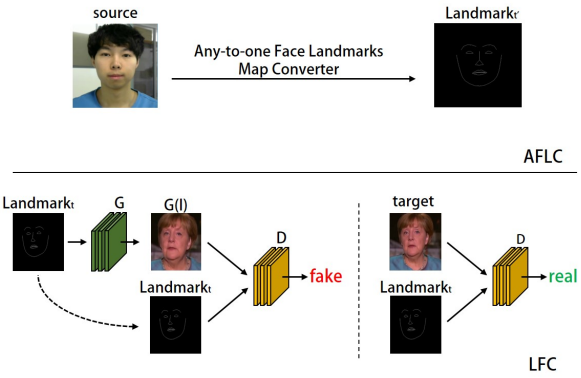

Fig. 1. Any-to-one Face Landmarks Map Converter(AFLC) transfer any source face into the target facial landmarks map. The detailed structure of AFLC is shown in Fig. 2 . Landmark-to-face Converter based on Conditional Generative Adversarial Network(LFC) transfer target facial landmarks map to the target face.

A. landmark-to-face converter based on conditional generative adversarial network

We construct the converter through the conditional GAN, and generate the real image of the target face by using the landmarks map of the target face as the condition information. The specific structure is shown in Fig. 1 LFC, where G is the generator, $Landmark_t$ is the Landmarks feature map of the target face, $G(l)$ is the generated target face, and D represents the discriminator to identify whether $G(l)$ is true—target face. The generator we use is a U-net structure of [32], and the discriminator is a patchGAN network in [6]. The loss function of the model consists of 3 parts. The first is conditional generative adversarial loss in Eq. 1 , the second is the L1 distance loss in Eq. 2 , where G represents the generator, D represents the discriminator, l represents $Landmark_t$ as input to the generator, z is random noise vector, t represents the real target face, $G(l,z)$ represents the generated target face. The third loss is the identity loss in Eq. 3 , where $V^{(i)}$ represents the i-th output of the Vgg19 network in [33], and $W^{(i)}$ is the corresponding weighting factor. The combination of these three losses is widely used for image reconstruction to generate sharp and realistic outputs.

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{l,t}[\log D(l,t)] + \mathbb{E}_{x,z}[\log(1 - D(l,G(l,z))] \tag{1}$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{l,t,z}\left[\|t - G(l,z)\|_1\right] \tag{2}$$

$$\mathcal{L}_{\text{id}}(G) = \mathbb{E}_{l,t,z}\left\{\sum_{i=1}^{T} W^{(i)}\left[\left\|V^{(i)}(t) - V^{(i)}(G(l,z))\right\|_1\right]\right\} \tag{3}$$

The total objective function of LFC is in Eq. 4 . Through the training of the network we will get a good generator G, which can transfer the landmarks map of the target face into the realistic target face.

$$\mathcal{G} = \arg\min_{G}\max_{D} \mathcal{L}_{cGAN}(G,D) + \lambda_1 \mathcal{L}_{L1}(G) + \lambda_2 \mathcal{L}_{\mathrm{id}}(G) \tag{4}$$

B. Any-to-one face landmarks map converter

If the source facial landmarks map is directly transmitted to LFC as the input, the generated target face has the expressions and postures of the input face. But it is very different from the texture and shape features of the real target face, including the size of the face and mouth, the distance between the eyes and so on. These differences make the generated face neither like source nor target. In order to solve this problem, we propose an Any-to-one Face Landmarks Map Converter(AFLC) module, which can effectively transfer the source face into the landmarks map of target face. AFLC can maintain the source facial expression and posture attributes while maintaining target facial shape and texture characteristics. What's more this transformation is universal and can be applied to various faces. In other words, AFLC can effectively perform any-to-one face landmarks map transformation and help the whole model achieve any-to-one face reenactment.

AFLC is shown in Fig. 2 . It includes the following parts: facial anchor points similarity transformation, weighted facial landmarks merge, image transformation based on Delaunay triangulation, weighted face fusion and local landmarks detection.

a) Facial anchor points similarity transformation: This step is the process of face alignment. We define two points in the facial landmarks(the left corner of left eye and the right corner of right eye) as the anchor points of the face. Then the anchor points of the source face is taken as the target position, and the anchor points of the target face are transferred to the target position. This is a process of similar transformation, and the following similar transformation matrix is obtained:

$$S = \begin{bmatrix} s_x\cos(\theta) & \sin(\theta) & t_x \\ -\sin(\theta) & s_y\cos(\theta) & t_y \end{bmatrix} \tag{5}$$

Where s_x and s_y represent scaling in the x, y direction, θ represents the angle of rotation, t_x and t_y represent the displacement transformation. Now transfer the landmarks of the target face to the corresponding position according to this similar transformation matrix. For each landmark, we have the following transformation formula:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} s_x\cos(\theta) & \sin(\theta) \\ -\sin(\theta) & s_y\cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{6}$$

b) Weighted facial landmarks merge: The purpose of this step is to merge the transferred target facial landmarks with the corresponding source facial landmarks, and the combination is weighted, represented by the weighting parameter α. The formula is Eq. 7 , where (x_t, y_t) is the coordinate of the target facial landmarks after transferring, and (x_s, y_s) is the coordinate of the source facial landmarks. After this, we can get a merged landmarks map which has compound information of source face and target face. Statistically, this can superimpose the distribution of source face and target face. Change the value of α to adjust the correlation between the target facial landmarks and the source facial landmarks.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \alpha \begin{bmatrix} x_t \\ y_t \end{bmatrix} + (1-\alpha) \begin{bmatrix} x_s \\ y_s \end{bmatrix} \tag{7}$$

c) Image transformation based on Delaunay triangulation: After obtaining the merged facial landmarks, the target face and the source face need to be transferred to the position corresponding to the merged facial landmarks, for making the face fusion without dislocation on the facial textures. We use Delaunay triangulation to divide the face into a series of disjoint triangles according to the distribution of landmarks, like Fig. 3 . It construct a unique mapping of the target face and the source face to the merged face corresponding to the merged landmarks map. Then pixel transform is performed on each facial triangle slice according to affine transformation.

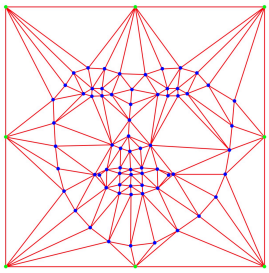

Fig. 3. Face segmentation with Delaunay triangulation on facial landmarks map.

d) Weighted face fusion and local landmarks detection: In the previous step, we get the transferred images of the target face and the source face. Now we need to fuse them together to get a merged face. This fusion process is also weighted and expressed by the β weighting factor. The formula is Eq. 8 . Where I_t represents the transferred target face, and I_s represents the transferred source face. Applying this formula, we can get the merged face, as shown in the merged face1/2 of the Fig. 2 .

$$I_{merge} = \beta I_t + (1-\beta) I_s \tag{8}$$

This step further fuses the characteristics of source face and target face, and can freely control the proportion of the two facial characteristics. The purpose is to enable us to obtain the attribute features we want in source face and target face. In this paper we use two different sets of (α, β). In the merged face1 generated by the first set, we extract the local facial landmarks of the cheek. In the merged face2 generated by the second set, we extract the local face landmarks except the cheek. And then we combine the two to generate a complete face landmarks

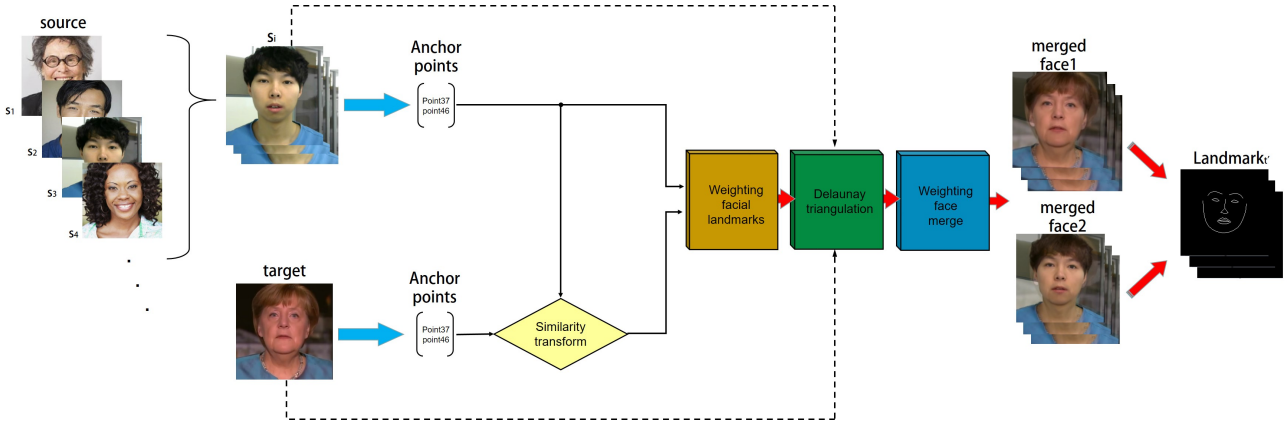

Fig. 2. Any-to-one Face Landmarks Map Converter(AFLC). Where s_i represents any input face in wild source dataset, target represents the target face, Anchor points are the two points selected in landmarks as the anchors for the calibration landmark positions, the four color modules are the geometry processing modules for landmarks and images, the two merged faces are weighted merged faces by different weights of α and β, and the $Landmark_{t'}$ is combined by local landmarks map extracted from the merged faces.

map, as shown by $Landmark_{t'}$ in Fig. 2 . In the experiments of LFC, we found that the cheek part of landmarks map mainly determines the shape and quality of the generated face, while other facial features mainly determine the expression properties of the generated face. We can combine the features we want in this way, and this is suitable for any source face.

Through the above steps of AFLC, the obtained facial landmarks map have the expression and posture attributes of the source face, and can well preserve the textures and shape of the target face as well. AFLC makes the model a good implementation of any-to-one face reenactment.

C. Model training and testing methods

The model proposed in this paper is divided into two converters, AFLC and LFC. LFC needs to be trained, and training process is described in Algorithm 1 . AFLC is a geometric transformation model whose algorithm is shown in Algorithm 2

Algorithm 1 Training algorithm of LFC

Input: Target facial landmarks map $\boldsymbol{landmark_t}$, target facial image **target**
Output: Trained generator **G**

Initialize W, λ_1, λ_2; **G**, **D**
$i \leftarrow 0$
while $i < iter$ **do**
  Sample training data
  Model forward propagation
  Calculate G loss: $\mathcal{L}_{cGAN}$, $\mathcal{L}_{L1}$ and $\mathcal{L}_{\text{id}}$; D loss: $\mathcal{L}_{real}$, $\mathcal{L}_{fake}$
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{cGAN} + \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{\text{id}}$
  Optimize **D** by minimizing $(\mathcal{L}_{real} + \mathcal{L}_{fake})/2$
  Optimize **G** by minimizing $\mathcal{L}_{total}$
  $i \leftarrow i + 1$
end while

Algorithm 2 Algorithm of AFLC

Input: Any source facial image **s**, one target facial image **t**
Output: Target facial landmarks map $\boldsymbol{l_t}$

$i \leftarrow 1$
while $i <= 2$ **do**
  Initialize α and β
  Extract the facial anchor points coordinates of **s** and **t**
  Transfer the **t** anchor points to the positions of **s** anchor points, and get the corresponding similar transformation matrix S
  Get landmarks of **t** transferred by S
  for j in length of landmarks **do**
    $landmarks^j_{merge} = \alpha * landmarks^j_t + (1-\alpha) * landmarks^j_s$
  end for
  Get $Landmarks_{merge}$
  Map **s** and transferred **t** to $Landmark_{merge}$ by Delaunay triangulation, and get warped image I_t and I_s
  for every pixel in image **do**
    $I^i_{merge} = \beta I_t + (1-\beta) I_s$
  end for
  $i \leftarrow i + 1$
end while
Extract the landmarks of the cheek in I^1_{merge} and the remaining landmarks in I^2_{merge} except the cheek, and combine them as $\boldsymbol{l_t}$

The forward propagation of the model is very simple, we only need to transfer the source face through AFLC and then generate the target face through the generator of LFC, and then we can get the transferred face with source facial expressions and postures as well as target facial textures and shape.

IV. Experiments

We use the python programming language, the pytorch deep learning framework, the OpenCV image processing library,

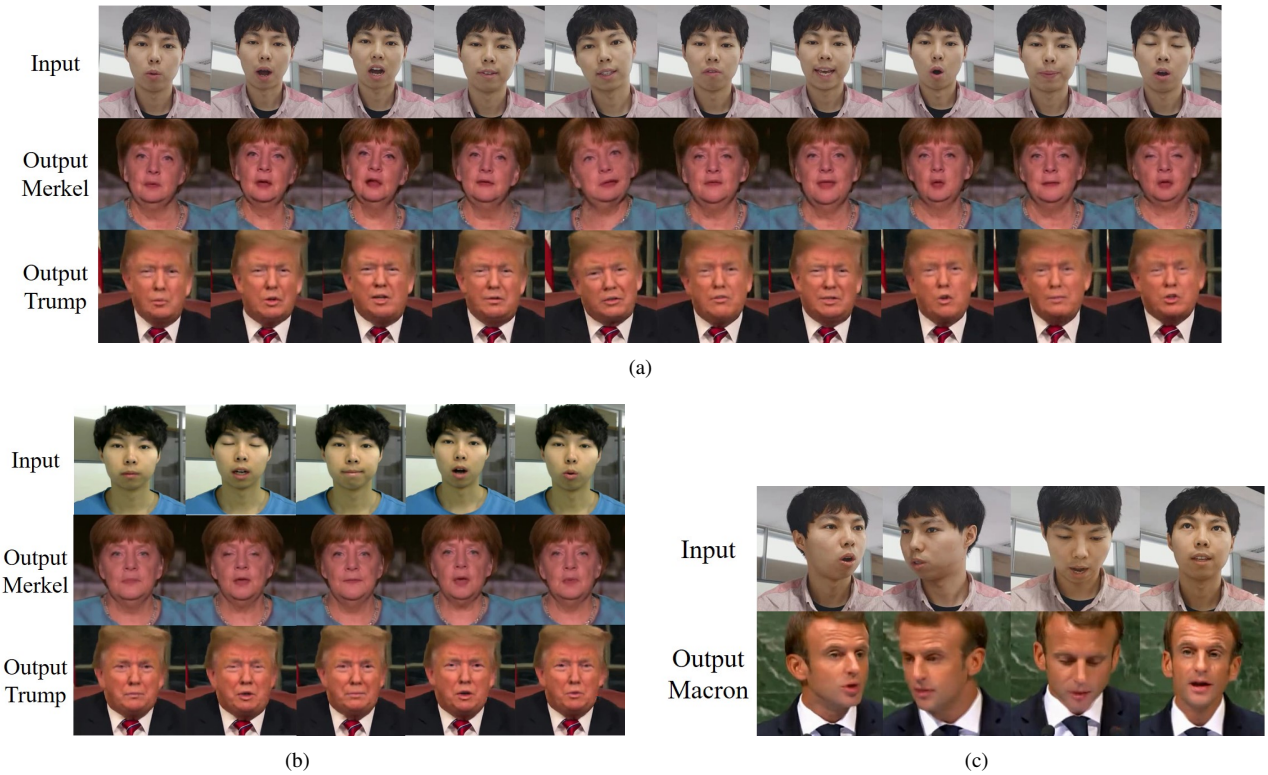

Fig. 4. Facial expressions and postures transferred results: (a) expressions transformation on 256*256 resolution, (b) expressions transformation on 512*512 resolution, (c) postures transformation on 256*256 resolution.

dlib face landmarks detection library and Cuda to implement the algorithms proposed in this paper. The computing platform used is CPU: Intel Xeon E5-2650(2.2GHz), GPU: 1*NVIDIA TITAN X(Pascal)(12GB). The training datasets we used was created from YouTube videos. The testing datasets includes facial images and videos collected from the Internet, and we record our own facial test videos. When training LFC, we use parameters λ_1=100.0, λ_2=10.0. $W^{(i)}$ is set to [1.0/32, 1.0/16, 1.0/8, 1.0/4, 1.0] according to experience. We use Adam optimizer with a learning rate of 2e-4 and β_1=0.5, β_2=0.999. The parameter (α, β) of AFLC is set to (0.6, 0.6) for merged face1 and (0.4, 0.4) for merged face2.

A. Evaluation of facial expressions and postures transformation

We use the videos of Angela Merkel[1], Donald Trump[2] and Emmanuel Macron[3] downloaded from YouTube for testing the transformation effect of the proposed model on facial expressions and postures. We cut each video to get average of 5 minutes of video to create a training dataset, and take some facial videos as the testing dataset. Some experimental results are shown in Fig. 4 . The first line of each image represents the input source face, and the remaining lines represent the output image of the target face. The faces generated in the figure maintain the textures of the target face, as well as the facial expressions (action of mouth, eyes, head, etc.) and face's postures (rotating head, etc.) are well-transferred. And the model is suitable for different resolution translation tasks.

B. Any-to-one transfer evaluation

We collect many facial images from Google Images to create a property-difference faces dataset, including different complexions(Black, White, Asian), different ages(Children, Adults, Seniors) and different genders(Woman, Man). This dataset has high individual difference in facial attribute characteristics and can represent universal facial images in wild. Use our model on this dataset, some results are shown in Fig. 5 . The experimental results show that the proposed model has good performance for the faces of different genders, ages and complexions, and has the generalization ability of any-to-one face reenactment.

C. Comparison with state-of-the-art

We compare the our model with some of the best performing models in face reenactment, which are model-based algorithm Face2Face in [3] and learning-based algorithm ReenactGAN in [12]. Because neither has open source code, and we have not yet obtained code from the authors of the two papers. We can only compare our results with their demos and the results presented in their papers. The training dataset used in the experiment is consistent with one used in the Face2Face and ReenactGAN papers, the Donald Trump[4] video downloaded

[1]https://www.youtube.com/watch?v=mJEKql2QV48
[2]https://www.youtube.com/watch?v=KWcmZ8hozvU
[3]https://www.youtube.com/watch?v=pfrXzsfIEW0
[4]https://www.youtube.com/watch?v=rTGMu8BrzS8

TABLE I
THE USER STUDY RESULTS. THE RANKINGS (1-5) SIGNIFY LOW(VERY UNLIKE, ETC.) TO HIGH(VERY LIKE, ETC.) SCORES. THE WEIGHTING FACTOR IS THE CORRESPONDING SCORES

	Face2Face					ReenactGAN					Ours				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
facial expression and posture transformation	0.21	0.30	0.28	0.20	0.01	0.07	0.30	0.45	0.15	0.03	0.10	0.27	0.34	0.24	0.05
Weighted average			2.5					2.77					2.87		
maintenance of facial identity attribute	x	x	x	x	x	0.15	0.27	0.40	0.16	0.02	0.00	0.18	0.31	0.33	0.18
Weighted average			x					2.63					3.51		

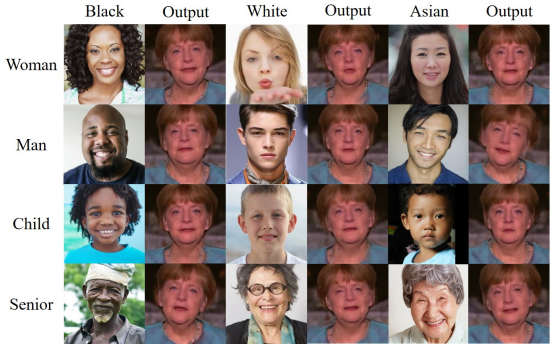

Fig. 5. Any-to-one expressions transferred results. From top to bottom, they represent woman, man, child, and the Senior. They are divided into three groups according to columns: black, white, and Asian. Each group is an input image and output image of target face(Merkel).

Fig. 6. Comparison with state-of-the-art. From top to bottom are the source inputs, Face2Face results, ReenactGAN results and our results.

from YouTube. The comparison results are shown in the Fig. 6 . Comparing the results of Face2Face, our model is better to capture the pose information of the input face and perform refined facial expressions and head postures transformation. Comparing the results of ReenactGAN, our model can better maintain the shape and textures information of the target face. The face generated by ReenactGAN is elongated under the influence of the input face, but in our results the generated faces still maintain the shape of the target face.

In addition to images contrast, we also perform a user study since human observation is more direct and reasonable for the validation of perceptual realism. We conduct two studies to assess the state-of-the-art and our method in facial expressions and postures transformation and the maintenance of facial identity attribute. We recruit 36 participants(17 females) to compare the generated faces of SOTA and ours, and rate them based on how similar to the expressions and postures of input faces and how similar to the real target face in the identity attribute. We use the 5-point Likert range of scores: 1-5 points represent Very unlike, unlike, a little like, like and very like. Our summary statistics are shown in Table I . In terms of facial expressions and postures transformation, our score is similar to ReenactGAN, and higher than that of Face2Face. This is because the results of Face2Face do not transfer the postures of the face, and this also illustrate our model can better transfer facial expressions and postures than Face2Face and ReenactGAN. In the maintenance of facial identity attribute, our score is higher than that of ReenactGAN(Face2Face does not have this item). This shows that our model maintains the shape and textures of the target face better than ReenactGAN. The above experiments show that our model can ensure the facial shape and textures are consistent with the target face while transferring expressions and postures. This makes our model more robust to the input face.

D. Ablation study on AFLC

We perform an ablation study on the key of model, AFLC. In this experiment, we use the dataset that mentioned above to test the impact of using and not using AFLC on the overall model. Our results are shown in Fig. 7 . The experimental results illustrate that when the AFLC is not used, the output faces will over-fitting the facial features of the input faces, including face shape, mouth size, distance between eyes, textures and so on. This makes the generated facial expressions exaggerated and fake, and we can easily distinguish that the textures of the generated faces do not match the real target face. When using the AFLC, the generated faces match the real target faces in appearance. Output facial expressions and textures are natural, meanwhile consistent with that of input faces. Therefore, the AFLC proposed in this paper is important and effective for the overall face reenactment model.

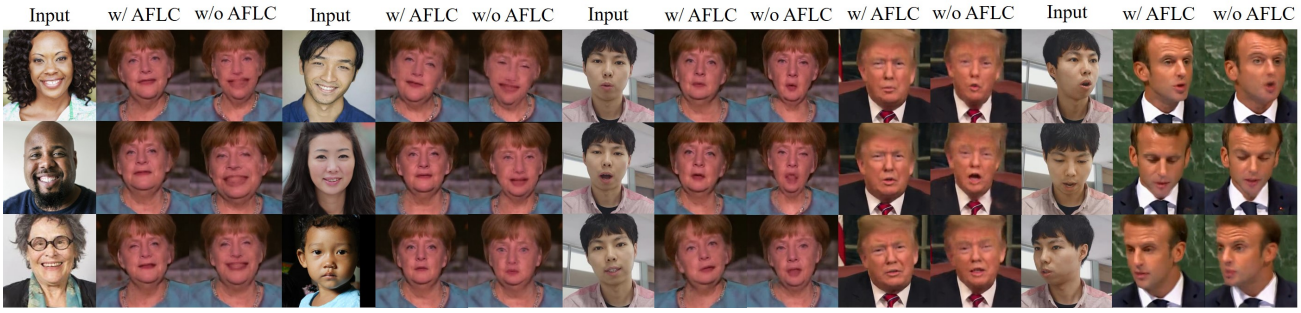

Fig. 7. Ablation study on AFLC. Compare the results of model with and without the Any-to-one Face Landmarks Map Converter.

V. Conclusion

This paper has proposed an Any-to-one Face Reenactment Model based on Conditional Generative Adversarial Network. The overall model structure is simple and lightweight, which has a dual converter structure: Any-to-one Face Landmarks Map converter(AFLC) and Landmark-to-face Converter based on Conditional Generative Adversarial Network(LFC). LFC we proposed can generate realistic and identity-preserving facial image from landmarks map. We create AFLC that guarantees to generate robust faces, and transfers facial expressions and postures while preserving the shape and textures of the target face. We verify that our proposed model enables pure learning-based any-to-one face reenactment. The state-of-the-art papers' source code is not available for comparison, hence we compare our results with those demos and results showed in their papers and perform a user study. From these comparison results we confirm the effect of our proposed model is comparable to or better than state-of-the-art.

Acknowledgment

This work was partly supported by NSFC (No. 61772529, U1636201 and U1536120) and the National Key Research and Development Program of China (No. 2016YFB1001003).

References

[1] Ira Kemelmacher-Shlizerman, Eli Shechtman, Rahul Garg, and Steven M Seitz. Exploring photobios. In *ACM Transactions on Graphics (TOG)*, volume 30, page 61. ACM, 2011.

[2] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3952–3960, 2015.

[3] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[4] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

[5] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM transactions on graphics (TOG)*, volume 24, pages 426–433. ACM, 2005.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[9] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[10] Runze Xu, Zhiming Zhou, Weinan Zhang, and Yong Yu. Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*, 2017.

[11] Xiaohan Jin, Ye Qi, and Shangxuan Wu. Cyclegan face-off. *arXiv preprint arXiv:1712.03451*, 2017.

[12] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018.

[13] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[14] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics (TOG)*, 35(4):128, 2016.

[15] Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer graphics forum*, volume 34, pages 193–204. Wiley Online Library, 2015.

[16] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library, 2003.

[17] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013.

[18] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):46, 2015.

[19] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42–1, 2013.

[20] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1675–1683, 2015.

[21] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. Automatic face reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4217–4224, 2014.

[22] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Real-time avatar animation from a single image. In *Face and Gesture 2011*, pages 117–124. IEEE, 2011.

[23] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017.

[24] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016.

[25] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. In *Affective Computing*. IntechOpen, 2008.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[27] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. Realistic dynamic facial textures from a single image using gans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5429–5438, 2017.

[28] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[29] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018.

[30] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 627–635. ACM, 2018.

[31] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. In *SIGGRAPH Asia 2018 Technical Papers*, page 231. ACM, 2018.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.