Face Reenactment via Generative Landmark Guidance

Chen Hu, Xianghua Xie*, Lin Wu
Department of Computer Science, Swansea University, UK
**x.xie@swansea.ac.uk*

***Abstract*—The identity preserving problem is one of the major obstacles in face reenactment. The problem occurs when the model fails to preserve the detailed information of the source identity, and especially obvious when reenacting different identities. The underlying factors may include the leaking of driver identity, due to the identity mismatch, and unseen large head poses. In this paper, we propose a novel face reenactment approach via generative landmark coordinates. Specifically, a conditional generative adversarial network is developed to estimate reenacted landmark coordinates for the driving image, which successfully excludes its identity information. These generated coordinates are further injected into the subsequent inference style transferal module to increase the realism of face images. We evaluated our method on the VoxCeleb1 dataset for self-reenactment and the CelebV dataset for reenacting different identities. Extensive experiments demonstrate that our method can produce more realistic reenacted face images.**

***Index Terms*—face reenactment, GAN, style transfer, facial landmarks**

I. Introduction

Face reenactment is a conditional image generation task. The input, namely the condition, to a face reenactment model comprises two parts, the source and the driving. The source is one or a set of images of a specific person, serving to provide appearance features of the person. The driving image is another image with an arbitrary face. The goal of face reenactment is to transfer the head pose and expression from the driving image to the face in the source image. Real world application of face reenactment includes video conferencing and film production. In the scenario video conferencing, the speaker's face can be reenacted to match the face motion of a translator [1]. In film industry, face reenactment can be used in a similar fashion, creating more natural localized motion pictures in different languages. Film makers can also fine-tune actors' facial movements with the help of face reenactment models.

Given the fact that it is infeasible to collect image pairs of two different people with identical head pose and expression, the self-supervised training strategy proposed by authors of X2Face [6] greatly helped the evolution of face reenactment methods. During training, the self-supervised strategy constrains the source and driving image in an input pair to be taken from the same video clip of the same person, therefore the driving image is also the groundtruth for the generated image. Despite the ease of training, in the testing scenario where the source and driving image are taken from different people, models trained by this strategy is at the risk of mixing the driving's identity in the image generator, resulting in the identity preserving problem, that is, the reenacted image shares structural similarity with both people in the input, instead of being the exact person in the source image.

Location of facial landmarks is valuable for defining a person's identity and head pose. During self-supervised training, if the eyes, nose and mouth in the generated image are precisely aligned with their corresponding location in the driving image, the generated image is more likely to be a faithful reenactment. Landmark locations can then be used to guide the model in the self-reenactment scenario. However, when reenacting different people, landmark locations in the driving image do not lead to desired output, as the locations now reflects the facial structures of different people, which can only aggravate the identity preserving problem. To help face reenactment methods benefit from landmark location, landmark coordinates also need to be reenacted. If these coordinates reflect the source's identity while matching the driving's head pose and expression, they can guide the model to process the test sample as if it is a self-reenactment case.

To obtain more realistic landmark estimation, we model this problem from the perspective of face reenactment evaluation. Specifically, for a generated face to be considered as a good reenactment, the generated face should look like the same person in the source image. Meanwhile, the generated pose and expression should match the ones in the driving image. We then propose a conditional generative adversarial network (GAN) to generate facial landmark points based on a person's identity as well as the desired head pose and expression. We follow the convention of face reenactment evaluation to formulate the head pose as pitch, yaw and roll angles, the expressions are formulated as the combination of facial action units (AUs) [16], which are considered the fundamental components of human face expressions. The proposed landmark GAN is not only beneficial for improving the performance, but it may also be employed to semantically generate human faces with expressions. This is because expressions can be decomposed into combinations of AUs. Given a combination of AUs, our landmark GAN can estimate fiducial points that match the desired expression, which can be subsequently used to generate face images [38].

With generated landmark coordinates, we explicitly leverages them to guide the reenactment process. We align individual facial landmarks based on these coordinates, we also generate facial landmark heatmaps to yield style transfer parameters, which are used to improve the realism of generated faces. Our contributions of this paper can be summarized below:

1) We propose a face reenactment method explicitly guided by facial landmark coordinates. An optical flow is first estimated based on the input source and driving image. The optical flow is a grid of coordinates determining where each pixel in the input should be moved into. We use the estimated optical flow to warp the feature maps of the source image. Individual landmarks are then respectively reenacted and aligned to guide the warped result.

2) We introduce a conditional landmark GAN that generates landmark coordinates based on the input face's identity, desired head pose angles and facial action units. In addition, we estimate style transfer parameters based on our estimated landmark coordinates to obtain more realist images. No information on the driving's identity is involved in these two modules.

3) We evaluate our method on the VoxCeleb1 [34] dataset for self-reenactment and the CelebV [35] dataset for reenacting different identities.

II. Related Works

A. Face Reenactment

Approaches to face reenactment can be categorized by the operation they used to generate new images, specifically, early face reenactment studies [1]–[5] mainly focus on rendering desired images from estimated 3D face models, while recent research more relies on optical flows to warp input images. The pipeline of rendering-based methods generally involves fitting 3D faces from images, then morphing these 3D models and rendering the reenacted results. These methods often require a large quantity of video frames as inputs to extract texture features for rendering, they are also limited to reenacting specific people due to the availability of 3D models. Recent works [6]–[12] propose warping-based face reenactment methods which utilise optical flows to map pixels from the source image to the reenacted image. Image warping on convolutional neural networks (CNN) was first proposed in [13], where the model can estimate optical flows that warp skewed numerical digits back to the regular view, thus improving the classification accuracy. In the context of face reenactment, optical flows are estimated based on input images. Optical flows are used to warp the source images [6] or the feature maps of source images [7]–[9], [11]. Given an input image and an optical flow, the warping operation generates a new image by moving each pixel from the input image to the coordinate in the new image indicated by the optical flow.

As mentioned in Section I, obtaining images for different people with the exact same poses and expressions is infeasible in practice, a now widely adopted self-supervised learning paradigm was proposed in [6]. Given a source image sampled from a video sequence, a corresponding driving image of the same person is also randomly sampled from the same video, making supervised learning possible as the driving image is the expected reenactment result. The self-supervised strategy subsequently leads to the identity preserving problem described in [7]. To remedy this issue, facial landmark coordinates are widely used as the guidance in face reenactment. The authors of ReenactGAN [35] built person-specific models to estimate landmark boundaries of reenacted faces. This method is capable of predicting reasonable facial structure, however, it is required to train separate landmark boundary estimator for different faces. MeshGCN [11] employed 3D Morphable Models (3DMM) to estimate dense 3D points on the reenacted faces. The work of [11] explicitly excludes the identity information of driving images by constructing reenacted 3D faces using the identity parameters of the source. This method achieved good performance in identity preserving. However, its optical flow estimation module is rather computationally heavy, as it is a graph convolutional neural network [15] that runs on the source and the reenacted meshes each with 53,215 vertices. Inspired by 3DMM, authors of [7] proposed a landmark transformer, which breaks down sparse 3D facial landmark coordinates into a base 3D face, and principal components that controls face shapes and expressions. This method was also later used by authors of [10]. By estimating corresponding principal component coefficients, the landmark transformer modifies landmark coordinates of the driving image to be more fitting to the identity of the source image. However, the performance of [7] is limited by the expressiveness of chosen principal components. The authors FSGANv2 [37] took an iterative approach to gradually rotate landmark coordinates through multiple steps until 2D landmark points match the desired poses. For the reenactment on expressions,FSGANv2 directly swaps the mouth points in the source image with corresponding points from the driving image.

Compared to previous works, our method estimates sparse 2D landmark coordinates in an end-to-end fashion through the landmark conditional GAN. The proposed method directly generates landmark points by considering the source's identity, desired head pose angles and expressions. Similar to 3DMM, landmark points estimated by our method exclude the driving's identity by only considering the source's facial structure, but 2D landmark points are more accessible than dense face vertices. In addition, our method adopts facial action units to formulate expressions instead of approximating by principal components. Evaluation results show that the proposed landmark GAN helps our model greatly improve the performance on identity preserving while maintaining a relatively low head pose error.

B. Generative Adversarial Network

Generative adversarial networks (GANs) [17] are a family of neural networks that learn to map input from certain distribution to a desired distribution. A GAN consists of a generator and a discriminator, these two networks play a zero-sum game such that when the training converges, the discriminator can no longer differentiate generated samples from real ones, namely the generator learns to produce realistic data samples. The input to the generator is not limited to random

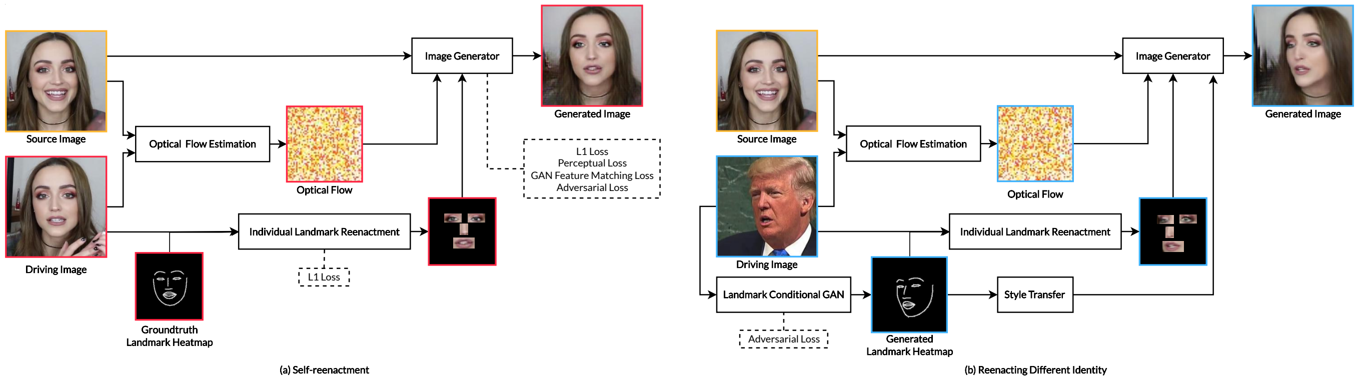

Fig. 1. Overview of proposed method for self-reenactment and reenacting different identities. Dashed boxes show loss functions that are responsible for the corresponding module.

variables sampled from a specified distribution, it can also be categorical labels [18], texts [19] or even images [20], granting more control over generated content. Face reenactment shares similarity with the image-to-image translation, a generative task that transfers an image from one domain to an image in another domain. Authors of Pix2Pix [20] have shown that supervision on pixel values combined with the use of GAN yields most realistic image-to-image translation results. This finding inspired many works on face reenactment, and GAN has become an indispensable component when generating realistic face images.

Vid2Vid [21] extended the work of Pix2Pix to generate video frames. For the pipeline of Vid2Vid, past frames and semantic segmentation maps are fed into the generator, an optical flow is estimated from the input and then applied to a past frame to generate the video frame for the current time step. The discriminator is responsible for the realness of generated video frames while generated frames are also supervised by the pixel values of real frames. Generative models such as Pix2Pix and Vid2Vid are supervised by groundtruth examples, and their input is also not non-stochastic, namely the input domain and the output domain are both well-defined, therefore these methods seldom consider the feature disentanglement problem presents in models with stochastic nature. Pipelines of optical flow based face reenactment methods are comparable to that of Vid2Vid, optical flows are estimated from the source and driving images, and then applied to the source image to synthesize reenacted faces. In terms of the use of GAN to generate face images, our method is also not an exception. GAN was deployed to ensure the realness of reenacted images, moreover, we also applied GAN to make sure that landmark coordinates generated by our method follows the distribution of real landmark coordinates. Details of GAN in our method are given in Section III.

III. Method

Figure 1 shows the overall framework of our face reenactment model. In general, we first estimate an optical flow based on input images. Then the eyes, nose and mouth in the source image are individually reenacted. Lastly, we use the estimated optical flow to warp the feature maps of the source image, and use reenacted facial landmarks to guide the subsequent image generation process. Specifically, when the identity of the driving image is different from the source, we can no longer leverage driving landmark coordinates for guidance because of identity mismatch. We therefore estimate landmark coordinates that match the source's identity and the driver's expression and head pose. These landmark coordinates are further adopted to yield style transfer parameters that improves the quality of generated images, as shown in Figure 1(b).

A. Optical Flow Estimation

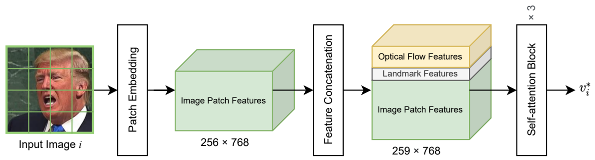

Fig. 2. Architecture of Vision Transformer as optical flow feature extractor.

Input images are sent to a neural network to extract features for optical flow estimation. Considering the fact that face reenactment requires the information on expression over an entire face, we adopt Vision Transformer as the feature extractor, as Vision Transformer directly learns the attention over the entire image while CNNs focus more on the locality of images. A ResNet feature extractor is also evaluated in Secction IV for comparison.

For Vision Transformer in our method, its architecture is shown in Figure 2. An input image i with size 224 $\times$ 224 is divided into 256 patches with size 14 $\times$ 14. Each image patch is embedded into a 768-dimensional vector, resulting in a 256×768 tensor v_i for an input image. In addition, a tensor $t \in \mathbb{R}^{3\times768}$ with learn-able initial values are concatenated to v_i, the first two rows of t store features for the optical flow estimation, and the third row of t contains features for landmark coordinate regression, which acts as an auxiliary task that helps the model perceive human faces. After an input image being embedded into $v_i \in \mathbb{R}^{259\times768}$, it further goes through three self-attention layers. The self-attention process is given as follows.

$$Q = v_i W_q,\ K = v_i W_k,\ V = v_i W_v \tag{1}$$

$$\alpha = softmax(QK^T/\sqrt{d_k}),\ v_i^* = \alpha V \tag{2}$$

where $W_q \in \mathbb{R}^{768\times d_q}$, $W_k \in \mathbb{R}^{768\times d_k}$ and $W_v \in \mathbb{R}^{768\times d_v}$ are learn-able parameters, we set $d_q = d_k = d_v = 768$. $\alpha \in \mathbb{R}^{259\times259}$ is the attention score given the input tensor v_i, and $v_i^* \in \mathbb{R}^{259\times768}$ is the output of the self-attention operation, it further goes through an MLP layer to yield the final result of a transformer block. Note that we only take the first two rows of v_i^* as the feature for optical flow estimation. In the case of ResNet, features obtained from the final global average pooling layer are projected to the dimension of $\mathbb{R}^{2\times768}$ that matches Vision Transformer's output.

Optical flow features for the source and the driving image are denoted by $u_s, u_d \in \mathbb{R}^{2\times768}$ respectively. u_s and u_d are first compressed to $\mathbb{R}^{2\times128}$ then reshaped to $\mathbb{R}^{1\times256}$, next, these two features are concatenated and sent to an multi-layer perceptron, resulting in $f \in \mathbb{R}^{1\times6272}$, f is reshaped to $\mathbb{R}^{7\times7\times128}$ and after going through a series of transpose convolutional layers, the estimated optical flow $f^* \in \mathbb{R}^{2\times224\times224}$ is obtained.

B. Individual Landmark Reenactment

Our motivation for individually reenacting facial landmarks is to help the model correctly perceive human faces when reenacting different identities. We observed that without explicitly specifying individual facial landmarks in the image generator, the model tends to synthesize more mouths or eyes than they should be on a single face when the source and the driving have different identities. This problem is more likely to happen when the model is trained on a dataset with fewer identities, such as CelebV. Another benefit of individually reenacting facial landmarks is that these landmarks can be aligned based on landmark coordinates to explicitly guide the reenactment process. We use four convolutional neural networks with an identical architecture, and each of them is dedicated to reenacting a different part of the face, namely the left eye, the right eye, the nose, and the mouth. Figure 3(a) shows we concurrently reenact selected landmarks; Figure 3(b) gives an example of the crop of the mouth from the source image, along with its counterpart from the landmark heatmap of the driving image are first sent to convolution layers, with the size of feature maps reduced by max pooling, then feature maps of the RGB mouth crop and that of the landmark heatmap are added element-wise and sent to transpose convolution layers to generate reenacted landmarks. All crops are fixed-sized and they are cropped around the centre point of corresponding landmark coordinates. The size of a landmark crop takes the value of the average size of corresponding landmark in the dataset. The landmark heatmap is obtained by first drawing 68 facial landmark points on a 224×224 image with black background, then points are connected by fitting B-spline curves, drawing the outlines of the face, eyes, eye brows, nose and mouth. When all landmarks are reenacted, they are directly placed on another blank 224×224 image I_p, and their centre point all align with the centre point of corresponding parts in the landmark heatmap.

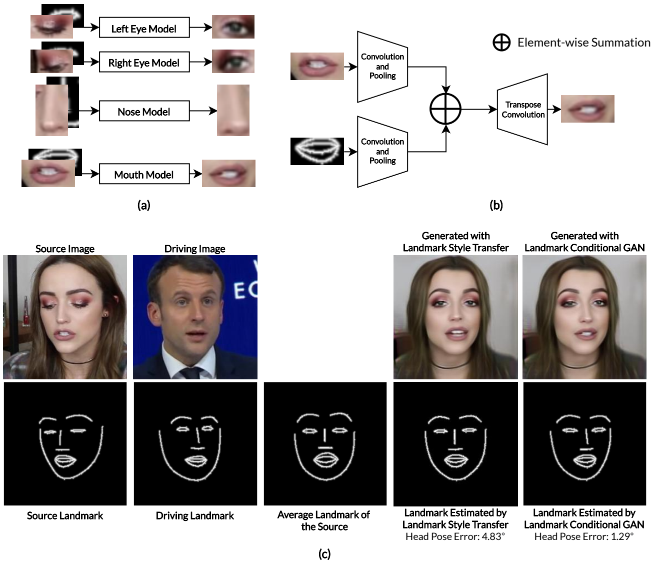

Fig. 3. (a) Individual landmarks are concurrently reenacted with models that share the same architecture. (b) Example of mouth reenactment showing the architecture of the landmark reenactment model. (c) Landmark coordinates estimated by style transfer significantly sacrifice head pose accuracy for identity preserving, whereas landmark conditional GAN better balances this trade off.

1) Landmark Coordinate Style Transfer: Since the driving landmark coordinates need to be modified to fit the source's identity, and inspired by [33]; we first propose a naive method for estimating the landmark coordinates, i.e., aligning the mean and variance of the driving coordinates $lmk_{driving}$ and those of the source coordinates lmk_{source}. The alignment is defined as,

$$lmk_{reenact} = \frac{lmk_{driving} - \mu_{driving}}{\sigma_{driving}} \times \sigma_{source} + \mu_{source} \tag{3}$$

$\mu_{source}, \sigma_{source}, \mu_{driving}, \sigma_{driving}$ can be obtained by computing the mean and variance of each person's landmark coordinates in the dataset, no learning is involved in this process. We also shift $lmk_{reenact}$ such that its centre point is at the same location as $lmk_{driving}$. Figure 3(b) shows an example the driving landmark heatmap generated by the original landmark coordinates and the one generated by style-transferred coordinates.

2) Landmark Conditional GAN: One major problem with the above landmark style transfer is that Equation 3 pushes landmark coordinates towards the average head pose in the dataset instead of truthfully acting as the desired pose. As shown in Figure 3(b), landmark coordinates modified by style transfer To remedy this problem, we propose the landmark conditional GAN as a more reliable estimator.

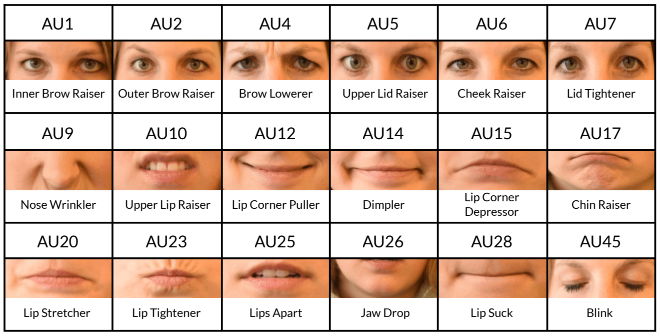

Fig. 4. Facial action units (AUs) in this study.

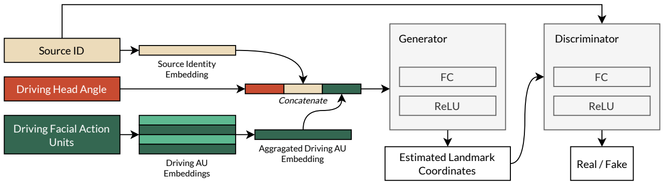
Source ID
Source Identity Embedding
Driving Head Angle
Concatenate
Driving Facial Action Units
Driving AU Embeddings
Aggregated Driving AU Embedding
Generator
FC
ReLU
Estimated Landmark Coordinates
Discriminator
FC
ReLU
Real / Fake

Fig. 5. The architecture of landmark conditional GAN.

The input to our conditional GAN is inspired by the evaluation metrics of face reenactment methods, specifically, we feed the source's identity, the driving's head pose, and facial action units appeared on the driving's face into the generator to obtain 68 2-D landmark coordinates. Facial action units (AUs) are predefined basic muscle movements on human faces. Figure 4 shows selected AUs in our method, these AUs are also used for face reenactment evaluation.

The convention of AU study is that a complex expression can be expressed by the addition of many different facial action units. For instance, an unhappy mouth can be expressed as AU15+AU17. We then took a similar approach to process AUs in the conditional GAN. Embedding vectors of facial action units appeared in the driving image are first selected, then these vectors are summed up to yield the overall expression feature for the input. The overall architecture of landmark conditional GAN is shown in Figure 5.

C. Image Generator

The face reenactment module is a U-Net-like convolutional neural network with only one skip-connection in the middle, Figure 5 shows its overall architecture. The source image is first sent to three convolutional layers with the size of its feature map r being reduced to 58×58, then the estimated optical flow map f^* (Section 2.1) with size 224×224 is resized to match the size of r and warps r, yielding the warped feature map r^*. The image I_p with reenacted landmark parts from the landmark reenactment module (Section 2.2) is also resized to 58×58 and concatenated to r^*. The concatenated feature map $r^*_{cat.}$ continues to go through intermediate convolutional layers with no change in feature map size, then r^* is concatenated to $r^*_{cat.}$ through the skip connection, the resulting feature map is further upsampled through bilinear interpolation and processed by convolution layers to generate the final reenacted image. The use of bilinear upsampling is aiming for alleviating the checkerboard artifact in images generated by convolutional neural networks [36].

1) Style Transfer Branch: In the case of reenacting different identities, although our landmark estimation methods greatly alleviated the identity preserving problem, unnatural face deformations exist as a result of inaccurate optical flow estimation. To rectify this issue, we further introduce a style transfer branch to the generator. The architecture of the style transfer branch is inspired by StyleGAN2 [26]. Instead of estimating style transfer parameters from random inputs, our model takes 1-channel landmark heatmaps as input. These landmark heatmaps are generated by first estimating the landmark coordinates using the conditional GAN in Section III.B.(2), then b-spline curves are fitted between adjacent landmark points that belong to the same facial landmark, namely drawing out the contours of the face, eyes, eyebrows, nose, and mouth. The use of heatmaps avoids the identity leak which is destined to happen if RGB driving images were used. Furthermore, since the heatmaps are generated based on coordinates estimated by our landmark conditional GAN, the identity information of the driving person is excluded as much as possible. The architecture of the style transfer branch is show in Figure 5.

D. Loss Function

Overall we use the weighted sum of four types of loss function to train our face reenactment model.

- L1 Loss: L1 loss is responsible for supervising the pixel value in generated images. During training, driving images are also the groundtruths for generated images, L1 loss is computed between these images. The weight on this loss is set to 20 for the entire image, and 5 for individually reenacted landmarks. We find that putting more weight on the L1 loss prevents the model from generating unexpected artifacts.
- Adversarial Loss: The adversarial loss we used for training is the same as [28]. Driving images are treated as ”real samples” while reenacted images are labeled as ”fake”.We set the weight for this loss to 1.
- GAN Feature Matching Loss [27]: GAN feature matching requires the discriminator to return intermediate features of real and generative samples, then forcing these features to be the same. For generative tasks with groundtruth samples, GAN feature matching loss makes the training more stable and converge faster. The weight for this loss is set to 1.
- Perceptual Loss [29]: Perceptual loss relies on a pre-trained VGG model to extract shallow visual features for real and generative samples. Pushing these features to be close ensures that low level features in the generated image, such as the shape of the face and shoulder, to be more realistic. The weight for this loss is set to 10.

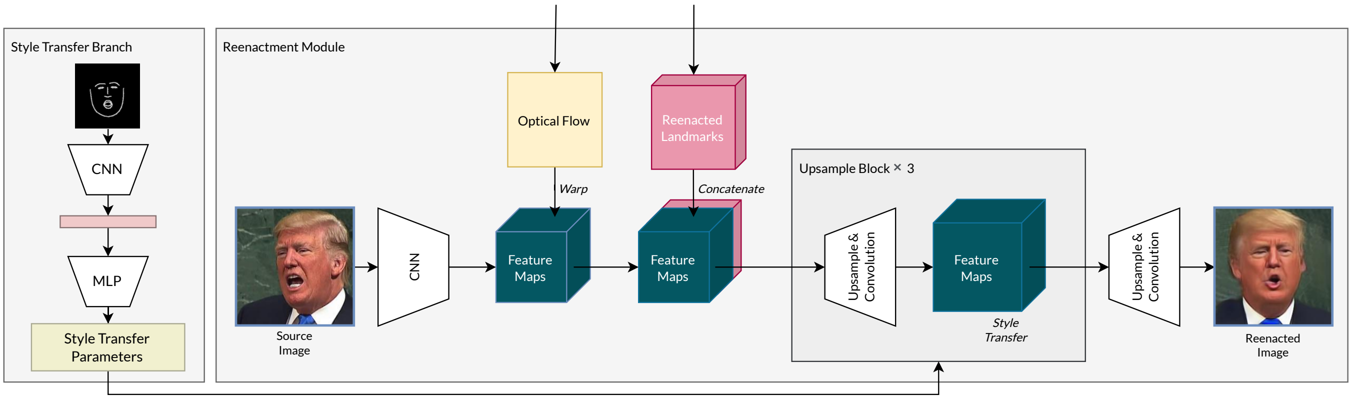

Fig. 6. The style transfer branch and the image generator.

IV. Experiments

A. Datasets and Experimental Settings

We evaluated our methods on the VoxCeleb1 dataset for self-reenactment, and the CelebV dataset for reenacting different identities. VoxCeleb1 is a dataset with 22,496 video clips extracted from YouTube. It contains 1,251 identities, and people's faces have been cropped into 256×256 images. CelebV has around 40,000 images for each of the five people in the dataset. For each person, their images are sampled from the same video and has also been cropped into 256×256 images.

For self-reenactment, we followed the protocol in [7], [11] and trained our model on the VoxCeleb1 dataset. The test set for evaluation consists of 100 videos from the test split given by the authors of VoxCeleb1, 2,083 source-driving image pairs are sampled from these videos for evaluation. For reenacting different identities, since our landmark conditional GAN and style transfer module require the information of known identities, we evaluated our method in two different scenarios. The first scenario follows [7], [11] and aims at reenacting unseen identities. Models are only trained on the VoxCeleb1 dataset, however, the test set are image pairs sampled from the CelebV dataset. For each person in CelebV, 2,000 source-driving image pairs are randomly sampled. In the second scenario, models are only trained on the CelebV dataset. Test set in this case also comprises 2,000 source-driving image pairs randomly sampled for each person.

B. Model Variants

We evaluated two model variants, denoted by their backbone network for optical flow estimation, namely ViT and ResNet. Further ablation studies were also conducted to validate the proposed landmark GAN and style transfer branch. The ViT model has three Vision Transformer layers for optical flow estimation, for the ResNet variant, transformer layers are replaced by ResNet-34. In the work of [22], a modified ResNet-50 (25 million parameters) outperforms the base 12-layer Vision Transformer (86 million parameters) on ImageNet top-1 accuracy by 10% with a pre-training dataset of 10M images. Given that there are three Vision Transformer layers (19M parameters) in our baseline model, we hence choose ResNet-34 (21M parameters) for comparison, which is shallower than ResNet-50. Additionally, we applied landmark style transfer described in Section 2.2 to both models and evaluated their performance accordingly. Models with landmark style transfer are denoted by ViT+LSt and ResNet-34+LSt.

C. Metrics

Performance on self-reenactment was evaluated through the following metrics, cosine similarity (CSIM), structural similarity (SSIM), peak signal-to-noise ratio (PSNR), root mean square error of head pose angles (PRMSE), and the ratio of correct facial action units (AUCON). CSIM measures the model's capability on identity preserving. It is derived from the cosine similarity between embedding vectors of the source and generated image, these vectors are extracted by a pretrained face recognition model Arcface [32]. SSIM and PSNR are exclusive to self-reenactment evaluation as they both require ground-truth images to compute, which is not possible for reenacting different identities. PSNR evaluates low-level similarity between generated images and ground-truths, while SSIM jointly evaluates the contrast, luminance, and structural between images. Head pose angels and facial action units are detected by OpenFace [30]. PRMSE is computed by calculating the root mean square error of head pose angles angels of the generated image compared against those of the driving image. For AUCON, both the driving and generated image are sent to OpenFace, the returned results show if facial action units in Figure 4 appear or not in the given image. Given the AU recognition results of the driving image, the ratio of AUs that correctly appear or do not appear in the generated image yields the AUCON.

D. Experimental Results and Analysis

Our experiments show that landmark coordinates of the driving image is a helpful heuristics for preserving the source's identity and achieving accurate head poses. By directly using driving landmark coordinates to guide the alignment of individual landmarks in the generated image, our model achieved

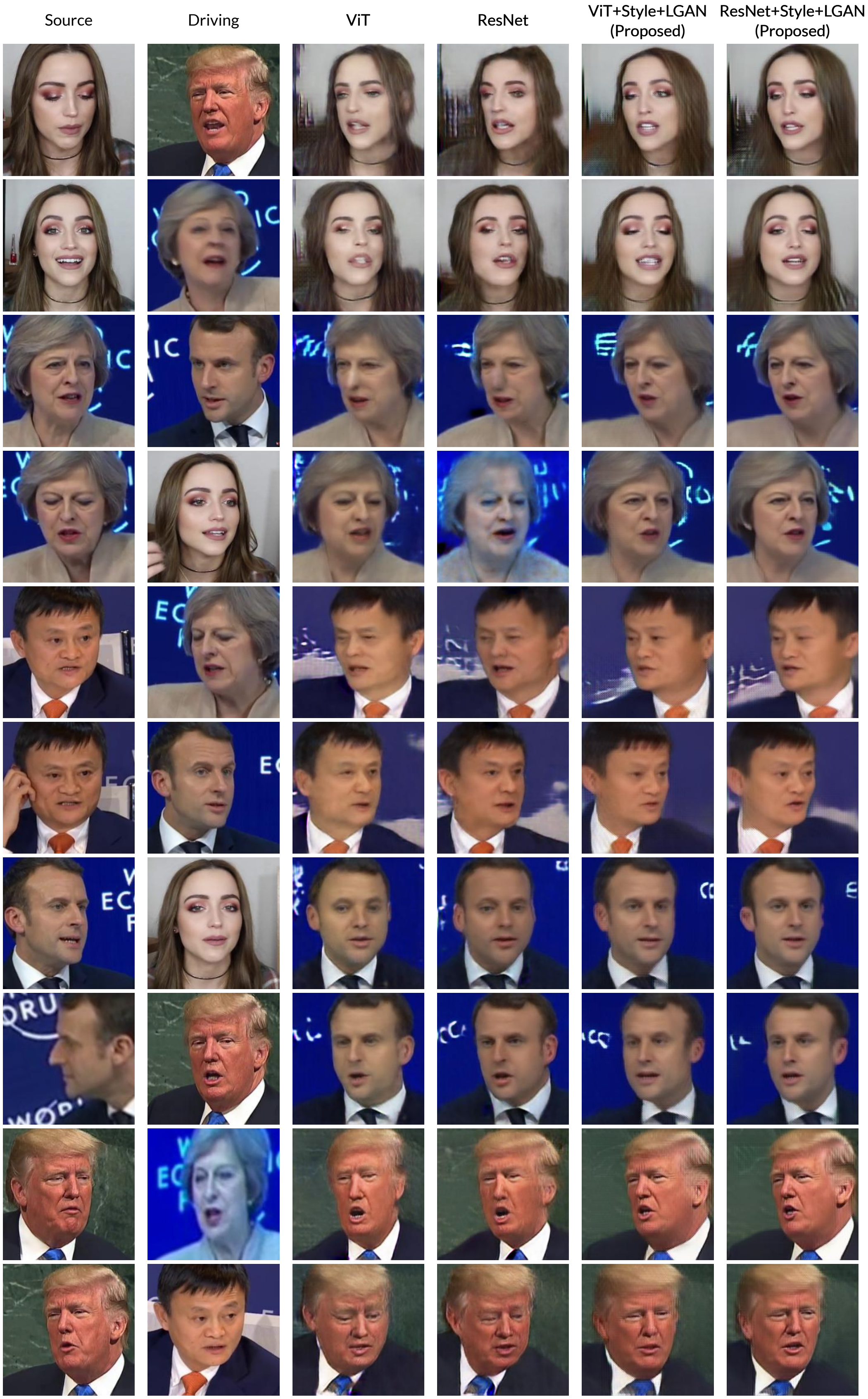

Fig. 7. Qualitative results of proposed models on CelebV dataset.

TABLE I
EVALUATION OF SELF-REENACTMENT ON VOXCELEB1

Model	CSIM↑	SSIM↑	PSNR↑	PRMSE↓	AUCON↑
Mesh Guided GCN [11]	0.822	0.739	**30.394**	3.20	**0.887**
MarioNETte [7]	0.755	**0.744**	23.244	3.13	0.825
Monkey-Net [9]	0.697	0.734	23.472	3.46	0.770
FirstOrder [8]	0.813	0.723	30.182	3.79	0.886
NeuralHead-FF [31]	0.229	0.635	20.818	3.76	0.791
X2face [6]	0.689	0.719	22.537	3.26	0.813
ViT	**0.879**	0.608	29.297	1.97	0.767
ResNet	0.878	0.650	29.606	**1.58**	0.793

Bold shows the best results, second bests are underlined. ↑ indicates the larger the value, the better the performance, ↓ means otherwise.

TABLE II
EVALUATION OF REENACTING DIFFERENT IDENTITIES WITH UNSEEN DATA ON CELEBV

Model	CSIM↑	PRMSE↓	AUCON↑
MarioNETte [7]	0.520	3.41	0.710
Mesh Guided GCN [11]	**0.635**	3.41	0.709
Monkey-Net [9]	0.451	4.81	0.584
FirstOrder [8]	0.462	3.90	0.667
NeuralHead-FF [31]	0.108	3.30	**0.722**
X2face [6]	0.450	3.62	0.679
ViT	0.525	2.95	0.694
ResNet	0.515	**2.35**	0.708

TABLE III
EVALUATION OF REENACTING DIFFERENT IDENTITIES WITH MODELS TRAINED ON CELEBV

Model	CSIM↑	PRMSE↓	AUCON↑
X2Face [6]	0.467	8.12	0.611
ViT	0.568	2.77	0.692
ViT+LGAN+Style	0.653	2.66	0.675
ResNet	0.570	**2.57**	**0.695**
ResNet+LGAN+Style	**0.661**	2.68	0.672

TABLE IV
EVALUATION OF LANDMARK ESTIMATION FOR REENACTING DIFFERENT IDENTITES ON CELEBV

Model	CSIM↑	PRMSE↓	AUCON↑
ViT	0.568	2.77	0.692
ViT+LSt	**0.620**	3.87	0.646
ViT+LGAN	0.619	2.60	0.682
ResNet	0.570	2.57	**0.695**
ResNet+LSt	0.616	3.78	0.650
ResNet+LGAN	0.614	**2.49**	0.687

TABLE V
EVALUATION OF STYLE TRANSFER FOR REENACTING DIFFERENT IDENTITES ON CELEBV

Model	CSIM↑	PRMSE↓	AUCON↑
Style	**0.647**	4.75	0.646
ViT	0.568	2.77	0.692
ViT+Style	0.587	3.22	0.668
ResNet	0.570	**2.57**	**0.695**
ResNet+Style	0.606	2.97	0.670

better performance on identity preserving and head pose accuracy on the VoxCeleb1 dataset, shown in Table I.

Following evaluation protocols in [7], [11], we evaluate our baseline models trained on VoxCeleb1 for reenacting unseen people from the CelebV test set. Shown in Table II, our methods still have lower head pose error, however, the identity preserving capability is less ideal compared to Mesh Guided GCN [11]. The main reason is that the driving's identity information was dismissed in the optical flow estimation stage of [11]. The landmark GAN and style transfer module we proposed are aiming at alleviating the identity preserving problem. These methods are designed to leverage training data to improve the quality of generated images, hence their evaluation are shown in separate tables, viz. Table III. In general, both landmark GAN and style transfer can improve the model's identity preserving capability. When combined together, our method achieves better identity preserving capability while maintaining a lower head pose error.

1) Self-reenactment: Table I shows models' performance on the VoxCeleb1 dataset. Our method better preserves identities (higher CSIM) and shows lower error on head pose angels (lower PRMSE). This illustrates that coordinates of driving landmarks are a strong prior that can help models perform better on these two metrics. SSIM takes the structural similarities into consideration, which includes both the face and background of the image. Our method pays more attention on the face region, backgrounds in reenacted images are often distorted, resulting in a low score in SSIM. We believe the expression accuracy (AUCON) of our method is related to the presumption made in terms of reenacting individual facial landmarks. In the preprocessing stage, eyes and mouths for all people in the dataset are cropped into fixed sizes to ensure that the landmark reenactment model can handle varying landmark and camera movement in images. However, this also limits the model's capability as there are cases where landmarks cannot fit in the cropped region. For instance, a wide open mouth or a close-up camera can lead to a larger mouth region, the model may still try to fit the entire mouth into region we cropped, resulting in less accurate expression reenactment. This phenomenon is also observed when reenacting different identities.

2) Reenacting Different Identities: Table II shows the overall performance on the CelebV dataset for models trained only on the VoxCeleb1 dataset. As mentioned above, Mesh Guided GCN [11] excluded the driving's identity when reconstructing 3D face models, the optical flows are then estimated based on these 3D models, leading to better identity preserving in generated images. With the guidance of landmark locations, our methods show more accurate head poses, since these landmark locations do not reflect the identity of the source image, our methods performs poorer than the self-reenactment scenario. In Figure 8, we cited image from [7] to compare images generated by different models.

The proposed landmark estimation and style transfer methods rely on learning from training samples to assist the image generation process, we then trained these models on the CelebV dataset. X2Face [6] is also trained from scratch on CelebV for comparison. Shown in Table III, X2Face shows slightly better identity preserving compared to results in Table II, however, the head pose error significantly increased. We found that X2Face has difficulty in converging when trained on a smaller dataset such as CelebV. Our methods achieve better identity preserving and lower head pose error thanks to the proposed landmark GAN and style transfer module.

3) ViT vs ResNet: In general, the ResNet variant of our method performs slightly better than its ViT counterpart. We believe this is because estimated optical flows are used to warp CNN features regardless of the network's backbone. ResNet is more compatible with this feature representation as it is also a CNN based model. However, Vision Transformer is still promising for face reenactment. When evaluated on ImageNet [22] with 10 million images for training, a Vision Transformer with 86 million parameters is outperformed by ResNet-50 with only 25 million parameters. In our case, the ViT head for optical flow estimation has 19 million parameters while the ResNet head has 21 million parameters. Our results show that the performance difference between these two models is negligible, a future study on Vision Transformer based image generator for face reenactment is worth investigating.

4) Landmark Estimation and Style Transfer: Table IV shows the ablation study on proposed landmark estimation methods. Landmark Style Transfer, denoted by LSt, is a crude way of estimating landmark coordinates, it achieves the best identity preserving among evaluated models, but it also significantly hinders the pose and expression accuracy. Landmark conditional GAN (LGAN), on the other hand, better balances these metrics.

Table V shows the ablation study on style transfer. The model named ”Style” is a baseline model without optical flow estimation, its reenactment process solely relies on the image generator and style transfer branch in Figure 6. Although this model shows relatively good identity preserving capability, it generates images with the poorest quality. Facial textures in generated images are often a mixture of the source and driving image. Although we explicitly excluded the RGB information from the style transfer input by using one-channel landmark heatmaps instead, the model ”memorizes” the connection between landmark heatmaps and their corresponding color images due to the self-supervised nature of the training stage. Evaluation metrics also show that style transfer promotes our models' the identity preserving capability at the cost of head pose and expression accuracy. However, this does not reflect the real contribution of style transfer. As shown in Figure 9, faces generated by non-style-transfer methods are distorted because of the warp operation, style transfer can help our model revert unnecessary distortion on faces, generating more realistic images.

5) Comparison with FSGANv2: Unlike most methods in Table I, the face reenactment of FSGANv2 does not rely on the optical-flow-based warping operation to generate images. We cited generated images from [37] to compare FSGANv2 with our method. Figure 10 shows the comparison between our ViT model and FSGANv2 on self-reenactment. The image at the top left corner is served as the the source image, while all images in the first row are driving images. In this scenario, when the difference between head pose angles are relatively small, both methods can generate realistic faces, while our method preserves sharper details in the source's face. As the head pose angles increase, the quality of generated images decreases for both methods. Figure 11 shows the comparison on reenacting different identities. Both the source and driving images are in-the-wild samples for our model. In addition, our landmark GAN and the style transfer branch is limited to working on known faces, therefore we also opted to the ViT model in Table I for this comparison. As mentioned in Section I, FSGANv2 iteratively rotates facial landmarks then synthesizes images based on rotated landmark points, enabling FSGANv2 to better preserve facial structures in generated images.

V. Conclusions and Future Work

We propose a face reenactment method guided by generative landmark coordinates. We evaluated our method in the following scenarios:

- **Self-reenactment**. In this scenario, the source and the driving image are taken from the same video clip of the same person, which allows us to directly use landmark coordinates in the driving image to guide the reenactment. We evaluated our method on the VoxCeleb1 dataset and compared against exiting methods following the same protocol. We show that images generated by our method are more similar to the input image's identity, and our method has lower head pose error compared to others. Our result show that landmark coordinates in the driving image are informative and helpful for identity preserving and accurately reenacting head pose angles.
- **Reenacting Different Unknown Identities**. In this scenario, the identities of the source and the driving image are different and they are not included in the training set. We used our self-reenactment model trained on VoxCeleb1 for evaluation on the CelebV dataset. Our method still managed to achieve lower head errors, indicating that the heuristic of using driving landmark coordinates to guide face reenactment is beneficial for accurately reenacting head movement. However, landmark coordinates require further adjustment to match the source face's identity.
- **Reenacting Different Known Identities**. In this scenario, the identities of the source and the driving image

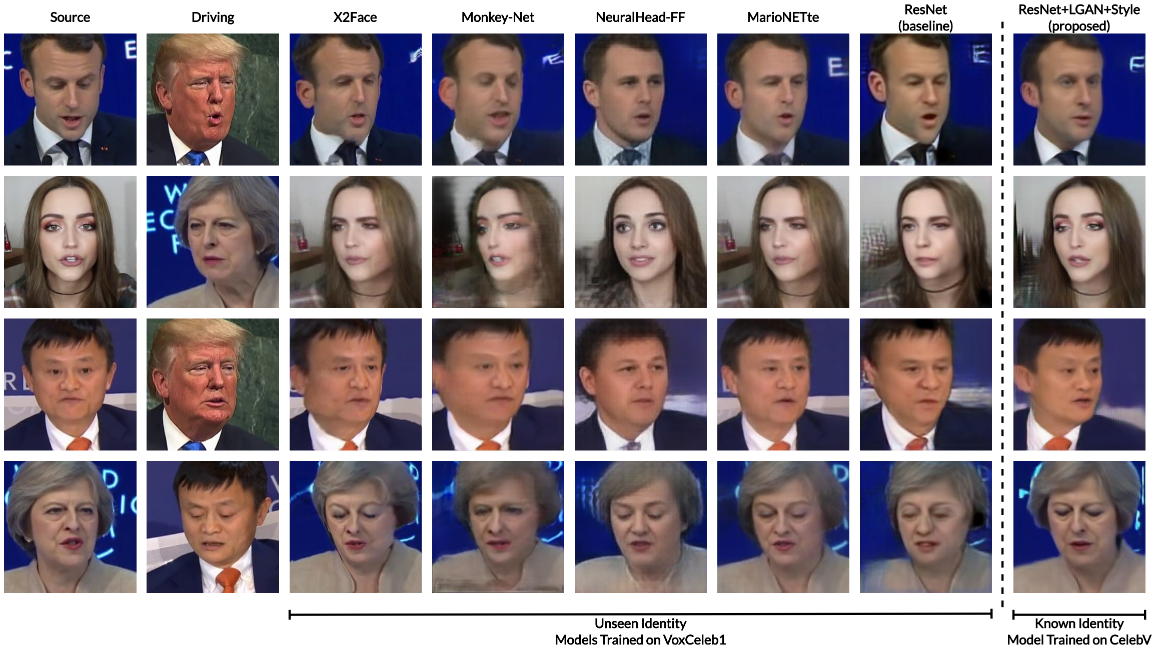

Fig. 8. Comparison of Reenacting Different Identities on CelebV.

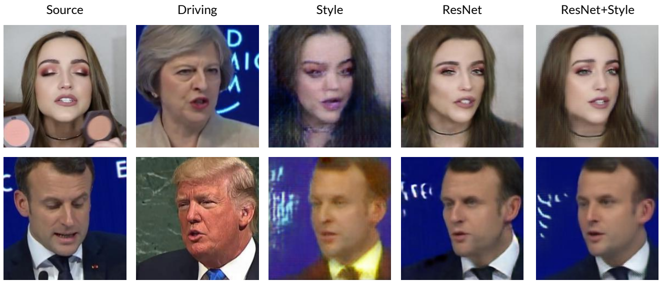

Fig. 9. Optical flow combined with style transfer improved the quality of generated images.

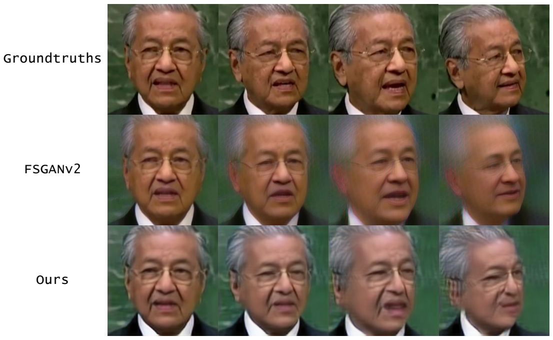

Fig. 10. Comparison with FSGANv2 on Self-reenactment

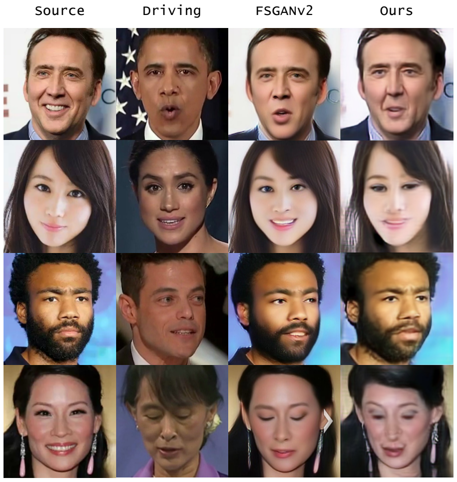

Fig. 11. Comparison with FSGANv2 on Reenacting Different Identites.

are different but they all appear in the training set. We proposed the landmark conditional GAN and the style transfer branch to assist our baseline model. Ablation experiments show that the landmark conditional GAN mainly contributes to identity preserving while the style transfer branch fixes shape distortions in generated images. When these two modules are combined together, the performance on identity preserving is greatly improved while the head pose errors remain relatively low. This proves that using head pose angles and facial action units can effectively estimate landmark coordinates of desired faces.

One noticeable limitation of our method is that it does not generate well to reenacting different and unknown identities. Our method heavily relies on landmark coordinates, but proposed landmark estimation method is only applicable to known faces. Another limitation of our method involves the proposed style transfer branch. Because the input to this module is a facial landmark heatmap, the modules struggles with appearance variations in the training data. For instance, in the VoxCeleb1 dataset, there are multiple video clips for the same person. The person may have beard in one video but the beard may be shaved in another video. In this case, the style transfer branch tends to remove the beard in generated images because the beard is not represented in the input.

Based on the above limitations, we suggest that one possible future work is to enhance the proposed landmark estimation method's generalisation capability. More generalised identity representations, such as identity embeddings from a face recognition model, can be used. Another topic worth investigating is to develop the style transfer branch that can handle varying appearances. Instead of modifying the entire feature maps, we could only transfer the styles to facial regions that needs changing. Lastly, we argue that it is also possible to design Vision Transformer based image generator. The performance of our model with the Vision Transformer backbone is lower than its ResNet counterpart. We believe this is because our estimated optical flows are used to warp feature maps estimated by a CNN, therefore ResNet is more compatible with this task. A Vision Transformer based image generator may be more suitable to the Vision Transformer backbone, however, how to define the warping operation on intermediate features of the Vision Transformer is still an open problem.

References

[1] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in CVPR 2016.

[2] Y.-T. Cheng, et al., “3D-model-based face replacement in video,” in SIGGRAPH, 2009.

[3] H. Kim, et al., “Deep video portraits,” in ACM Transactions on Graphics, 2018.

[4] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “What makes tom hanks look like tom hanks,” in ICCV, 2015.

[5] D. Vlasic, M. Brand, H. Pfister, and J. Popović, “Face transfer with multilinear models,” ACM Trans. Graph., 2005.

[6] O. Wiles, A. S. Koepke, and A. Zisserman, “X2face: A network for controlling face generation using images, audio, and pose codes,” in ECCV, 2018

[7] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, “Marionette: Few-shot face reenactment preserving identity of unseen targets,” in AAAI, 2020.

[8] A. Siarohin, S. Lathuili‘ere, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” in Advances in Neural Information Processing Systems, 2019

[9] A. Siarohin, S. Lathuili‘ere, S. Tulyakov, E. Ricci, and N. Sebe, “Animating arbitrary objects via deep motion transfer,” in CVPR, 2019

[10] G. Yao, Y. Yuan, T. Shao, S. Li, S. Liu, Y. Liu, M. Wang, and K. Zhou, “One-shot face reenactment using appearance adaptive normalization,” in AAAI, 2021

[11] G. Yao, Y. Yuan, T. Shao, and K. Zhou, “Mesh guided one-shot face reenactment using graph convolutional networks,” in 28th ACM International Conference on Multimedia, 2020.

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

[13] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, “Spatial transformer networks,” in NIPS, 2015.

[14] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3d faces,” in SIG-GRAPH, 1999.

[15] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, “Generating 3D faces using convolutional mesh autoencoders,” in ECCV, 2018.

[16] W. V. F. Ekman, P., “The facial action coding system: A technique for measurement of facial movement,” 1978.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in NIPS, 2014.

[18] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” in arxiv: 1411.1784, 2014.

[19] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in ICML, 2016.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with condi-tional adversarial networks,” in CVPR, 2017.

[21] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, “Video-to-video synthesis,” in NeurIPS, 2018.

[22] A. Dosovitskiy,et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in ICLR, 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in NIPS, 2017.

[24] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, “A survey of visual transformers,” in arXiv: 2111.06091, 2021.

[25] F. Zhu, Y. Zhu, L. Zhang, C. Wu, Y. Fu, and M. Li, “A unified efficient pyramid transformer for semantic segmentation,” in ICCV Workshops, 2021.

[26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in CVPR, 2020.

[27] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., ”High- resolution image synthesis and semantic manipulation with conditional gans,” in CVPR, 2018

[28] Radford, A., Metz, L., Chintala, S., ”Unsupervised representation learning with deep convolutional generative adversarial networks,” in ICLR, 2016

[29] Johnson, J., Alahi, A., Fei-Fei, L., ”Perceptual losses for real-time style transfer and super-resolution,” in EVVC 2016

[30] Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P., ”Openface 2.0: Facial behavior analysis toolkit,” in 13th IEEE International Conference on Automatic Face Gesture Recognition, 2018

[31] Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V., ”Few-shot adversarial learning of realistic neural talking head models,” in ICCV, 2019

[32] Deng, J., Guo, J., Xue, N., Zafeiriou, S., ”Arcface: Additive angular margin loss for deep face recognition,” in CVPR, 2019

[33] Huang, X., Belongie, S., ”Arbitrary style transfer in real-time with adaptive instance normalization,” in ICCV, 2017

[34] Nagrani, A., Chung, J., Xie, W., Zisserman, A., ”Voxceleb: Large-scale speaker verification in the wild,” in Computer Science and Language, 2019

[35] Wu, W., Zhang, Y., Li, C., Qian, C., Loy, C., ”ReenactGAN: Learning to reenact faces via boundary transfer,” in EVVC 2018

[36] Odena, A., Dumoulin, V., Olah, C., ”Deconvolution and checkerboard artifacts,” in Distill 2016

[37] Nirkin, Y., Keller, Y., Hassner., Tal, ”FSGANv2: Improved Subject Agnostic Face Swapping and Reenactment,” in PAMI, 2022

[38] Sun P., Li Y., Qi H., Lyu S., ”LandmarkGAN: Synthesizing faces from landmarks,” in Pattern Recognition Letters, 2022