One-shot Decoupled Face Reenactment with Vision Transformer

Chen Hu[1[0000−0001−5023−054X]], Xianghua Xie[1[0000−0002−2701−8660]]

Department of Computer Science, Swansea University, Swansea, UK

Abstract. Recent face reenactment paradigm involves estimating an optical flow to warp the source image or its feature maps such that pixel values can be sampled to generate the reenacted image. We propose a one-shot framework in which the reenactment of the overall face and individual landmarks are decoupled. We show that a shallow Vision Transformer can effectively estimate optical flow without much parameters and training data. When reenacting different identities, our method remedies previous conditional generator based method's inability to preserve identities in reenacted images. To address the identity preserving problem in face reenactment, we model landmark coordinate transformation as a style transfer problem, yielding further improvement on preserving the source image's identity in the reenacted image. Our method achieves the lower head pose error on the CelebV dataset while obtaining competitive results in identity preserving and expression accuracy.

Keywords: Face reenactment · Vision Transformer · optical flow · facial landmark.

1 Introduction

Face reenactment is an image generation task. In the one-shot setting, given a pair of human face images, called the source and the driving, respectively, the face in the generated image should not only have the same identity as the source image, but also share the same pose and expression in the driving image. Practical applications of face reenactment include video conferencing and film production. In video conferencing, the speaker's face can be reenacted to match the face motion of a translator [20]. For film production, substitute actors can dub an iconic character with mouth movements and expressions properly mapped to the original character's face.

Early studies [4, 12, 19, 20, 22] on face reenactment primarily focused on fitting faces from images to 3D models, then morphing 3D faces and rendering the reenacted results. These methods require a large quantity of video frames as inputs and are limited to reenacting specific identities. More recent studies [7, 17, 18, 24, 26, 27, 29] propose one-shot or few-shot face reenactment and utilise optical flow to map pixels from the source image to the reenacted image, image warping then becomes an essential operation for these methods. Image warping on convolutional neural networks (CNN) was first proposed in [10], where

the model can estimate an optical flow map that warps skewed digits back to the regular view, improving the classification accuracy. For face reenactment, image warping means estimating an optical flow that determines how pixel values should be sampled from the source image or its feature maps such that the desired reenacted image can be generated.

Since 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 [24]. Given a source image sampled from a video sequence, a corresponding driving image of the same person is randomly chosen from the same video, this makes 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]. The model is only supervised from optical flow estimation for the same person. When applied to reenacting faces with different identities, defected images may be generated, making the person in the reenacted image looks more similar to the one in the driving image. Inspired by 3DMM[3], authors of [7] approached this issue by proposing the landmark transformer, which breaks down 3D facial landmark coordinates into a base 3D face, and principal components that controls the person-specific shape and expression of the face. By estimating corresponding principal component coefficients, 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 derived principal components. The work of [27] estimates 3DMM parameters for source and driving images, the authors explicitly exclude the identity information of driving images by constructing reenacted 3D faces using only the identity parameters of source images. This method achieves the state-of-the-art performance in identity preserving, yet currently there is no 3DMM annotaion for face reenactment datasets, and the optical flow estimation module in [27] requires heavy computational resource, because it is a graph convolutional neural network [16] that runs on the source and the reenact mesh each with 53,215 vertices.

In a latest work [26], the authors also utilised landmark transformer to transform landmark coordinates of the driving image. They estimated an global optical flow for the source image based on landmark heatmaps [1] derived from transformed coordinates, while facial landmarks such as the nose and the mouth are separately reenacted. Inspired by this strategy, we also decoupled the reenactment of the entire face and facial landmarks. NeuralHead[28] is another reenactment method that relies on facial landmark coordinates. Compared to [7, 17, 18, 24, 26, 27], NeuralHead obtained competitive results on head pose and expression accuracy, however, the performance on identity preserving is significantly lower. Authors of [7] believe that this is an indirect evidence of the lack of identity-preserving capacity in methods based on adaptive instance normalization[9], we argue that this is the immediate effect of feeding a conditional GAN with inappropriate conditions. The image generator of NeuralHead directly takes face sketches generated by landmark coordinates of driving images as input. When reenacting different identities, no information on the identity of the source im-

age is encoded into NeuralHead's input. This suggests that when a reenactment method, e.g. NeuralHead, is conditioned on unmodified landmark coordinates of the driving image, such method can achieve great performance in reenacting poses and expressions without considering the identity of the source image. In contrast, a method that does not require any landmark coordinates, such as X2face, can easily outperform the conditioned method on identity preserving though poses and expressions are less accurate.

Considering the implication of performance and limitations of [28], our method directly estimate the global optical flow from the source and driving image without any prior. Additionally, we use Vision Transformer [6] for optical flow estimation. Vision Transformer is an extension of the attention-based neural network [21] to computer vision tasks. Unlike CNNs that are characterised by weight sharing and locality, Vision Transformer has less inductive bias [13], attention weights are dynamically computed depending on the input and features are aggregated from all elements in the input sequence instead of an neighbouring area. Although experiments show that Vision Transformer would require much more parameters and tens of millions training examples to reach the same level of performance as CNN[6], we show that a shallow Vision Transformer is also a good optical flow estimator for face reenactment.

As for individual landmark components, we use face sketches generated by landmark coordinates of the driving face for reenactment. Our approach to the identity preserving problem is aligning the mean and variance of the driving face's landmark coordinates with those of the source face. Experiments show that this approach effectively improve the performance of our models on identity preserving.

2 Methods

Figure 1 shows the overall framework of our proposed one-shot face reenactment method. The source and driving image are first fed into the facial feature extractor, extracted features for the source and driving image are concatenated and sent to a multi-layer perceptron (MLP) and transpose convolutional layers to estimate an optical flow map, which later warps the feature map of the source image in the face reenactment module. The landmark reenactment module is responsible for individually reenacting the left eye, the right eye, the nose and the mouth of the source image. As shown in Figure 1, only cropped parts in the source image and the face sketch are sent to this module. The face reenactment module takes the source image, the estimated optical flow map and reenacted landmark parts as input and generates the reenacted face image.

2.1 Facial Feature Extractor and Optical Flow Estimation

Facial feature extractor is responsible for extracting and aggregate image features for optical flow estimation. The extractor is comprised of a Vision Transformer with three layers. The architecture of this module is shown in Figure 2. An

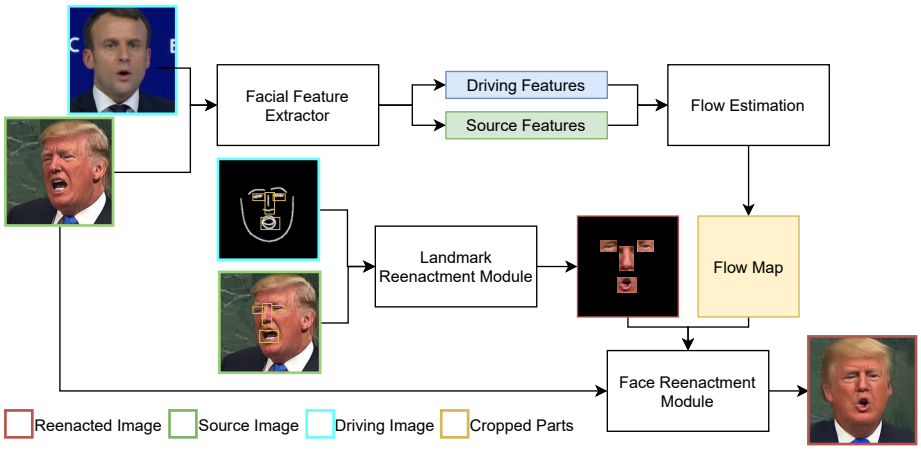

Fig. 1. The overall architecture of proposed method.

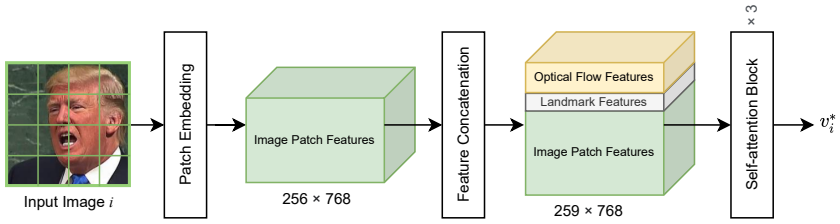

Fig. 2. Architecture of Facial Feature Extraction Module

input image i with size 224×224 is divided into 256 patches with size 14×14. 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, $d_q = d_k = d_v = 768$, $\alpha \in \mathbb{R}^{259\times259}$ 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.

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 MLP, 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.

2.2 Landmark Reenactment Module

Landmark reenactment modules reenacts individual facial landmarks, it contains four convolutional neural networks that share the same architecture, however, 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. The architecture of each neural network is similar to an autoencoder. Figure 3(a) shows the crop of the mouth from the source image, along with its counterpart from the face sketch of the driving image are first sent to three convolution layers, with the size of feature maps reduced by half through max pooling, then feature maps of the RGB mouth crop and that of the face sketch are element-wise added and sent to transpose convolution layers to generate the reenacted parts. All crops are fixed-sized and they are cropped around the centre point of corresponding landmark coordinates. The sketch of a face is obtained by first drawing 68 facial landmark points on a 224×224 image with pure black background, then points are connected by B-spline curves, drawing the outlines of the face, eyes, eye brows, nose and mouth.

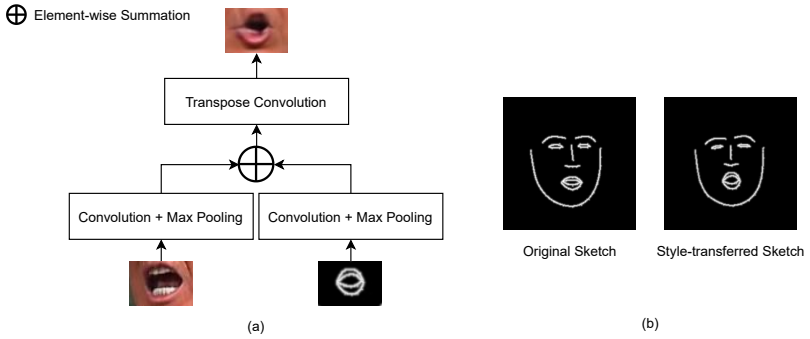

Fig. 3. (a) The architecture of landmark reenactment module. (b) An example of landmark style transfer.

When all parts 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 driving sketch.

Landmark Style Transfer Although our landmark reenactment module relies on the face sketch generated by driving landmark coordinates, no modification on landmark coordinates is needed during training as source images and driving images share the same identity. When we reenact faces with different identities, this leads to the identity preserving problems described in Section 1 due to the identity mismatch between the source image and the driving sketch. To remedy this, we modify driving landmark coordinates by treating it as a style transfer problem. Inspired by [9], to adapt the driving person's landmark coordinates to the landmark style of the source person, we align the mean and variance of the driving coordinates $lmk_{driving}$ with those of the source coordinates lmk_{source},

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

$\mu_{source}, \sigma_{source}, \mu_{driving}, \sigma_{driving}$ 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 face sketch generated by the original landmark coordinates and the one generated by style-transferred coordinates.

2.3 Face Reenactment Module

The face reenactment module is a U-Net-like convolutional neural network with only one skip-connection in the middle, Figure 4 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 [14].

2.4 Loss Function

All modules of our method are jointly trained in the adversarial and self-supervised fashion. Adversarial [15] loss is essential for image generation tasks, and since driving images are groundtruths for training face reenactment models, L1 loss on pixel values and the perceptual loss [11] are adopted. We also use the GAN feature matching loss [23], as it can stabilize and speed up the training of image generation tasks when groundtruth images are available. Lastly, we also consider

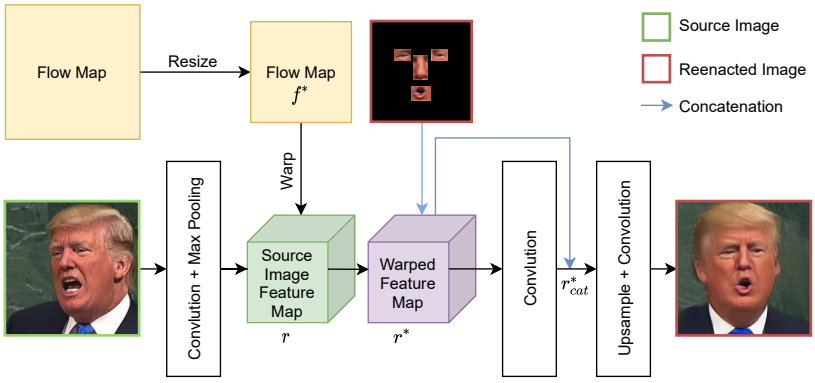

Fig. 4. The Overall architecture of Face Reenactment Transformer

the L1 loss for landmark coordinate regression for the auxiliary task described in Section 2.1. A linear combination of above losses is suffice to train our proposed method, the loss function J is given by,

$$J = \sum_{k=0}^{5} \lambda_i J_i \tag{4}$$

where J_0 is the L1 loss of regressed landmark coordinates, J_1 is the GAN feature matching loss, J_2 is the adversarial loss, we let $\lambda_0 = \lambda_1 = \lambda_2 = 1$; J_3 is an L1 loss for pixel values of reenacted landmark parts in Section 2.2, and $\lambda_3 = 5$; J_4 is the perceptual loss of reenacted images, $\lambda_4 = 10$; J_5 is the L1 loss for pixel values of the entire reenacted image, $\lambda_5 = 20$. We find that putting more weight on the L1 loss of pixel values prevents the model from generating unexpected artifacts, and the emphasis on perceptual loss helps obtaining faces and shoulders with more realistic shapes.

3 Experiments

We evaluated our methods on the CelebV[25] dataset following protocols in [7]. CelebV is a dataset with video frames for five celebrities, each of them has around 40k images. The evaluation focuses reenactment with different identities, namely the person in the source image is different from the one in the driving image.

3.1 Model Variants

We tested two model variants: the baseline model, denoted by **ViT**, has three Vision Transfomrer layers for facial feature extraction, and the **ResNet-34** model with all Vision Transformer layers in the baseline model replaced by a ResNet-34 [8] backbone for feature extraction. In [6], a modified ResNet-50 (25M parameters) outperforms the base 12-layer Vision Transformer (86M 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.

3.2 Metrics

Cosine similarity (CSIM) measures the quality of identity preserving by comparing the distance between face embedding vectors of source images and reenacted images, where embedding vectors are estimated by the pre-trained face recognition model ArcFace[5]; the root mean square error of the head pose angles (PRMSE), and the ratio of identical facial action unit values (AUCON) compares driving images and reenacted images, PRMSE tells how accurately the head pose is reenacted, and AUCON represents how close the reenacted expression is to that of the driving image. Both head pose angles and action unit values are estimated by OpenFace[2].

Table 1. Evaluation results of reenactment with different identities on CelebV following protocols in [7]. Values in bold stands for the best results, underlined values are the second best ones. The upward arrow indicates the larger the value, the better the performance, the downward arrow means a smaller value is better.

Model	CSIM↑	PRMSE↓	AUCON↑	Source of Optical Flow
Mesh Guided GCN[27]	**0.635**	3.41	0.709	3D faces
MarioNETte[7]	0.520	3.41	<u>0.710</u>	raw coordinates
MarioNETte+LT[7]	0.568	3.70	0.684	transformed coordinates
NeuralHead-FF[28]	0.108	3.30	**0.722**	no optical flow
X2face[24]	0.450	3.62	0.679	raw images
ResNet-34	0.570	**2.57**	0.695	raw images
ResNet-34+LSt	0.616	3.78	0.650	raw images
ViT	0.568	<u>2.77</u>	0.692	raw images
ViT+LSt	<u>0.620</u>	3.87	0.646	raw images

3.3 Analysis

Table 1 shows the metrics of our methods compared to other methods evaluated under the same protocol, including types of input that the optical flow estimation is based on. Figure 5 shows the qualitative results of our model variants as well as typical failure cases. Among our proposed methods, the model with ResNet-34 as the backbone of optical flow estimator shows the best performance,

achieving lower head pose error than previous work. The use of landmark style transfer significantly boosts the identity preserving capability of our methods while decreasing the head pose and expression accuracy. Distorted input and large head pose are two challenging cases for our method, a distorted driving image results in a face which is more similar to the distorted shape, and a large head pose induces misaligned facial landmarks in the reenacted image. Detailed analysis is presented in following sections.

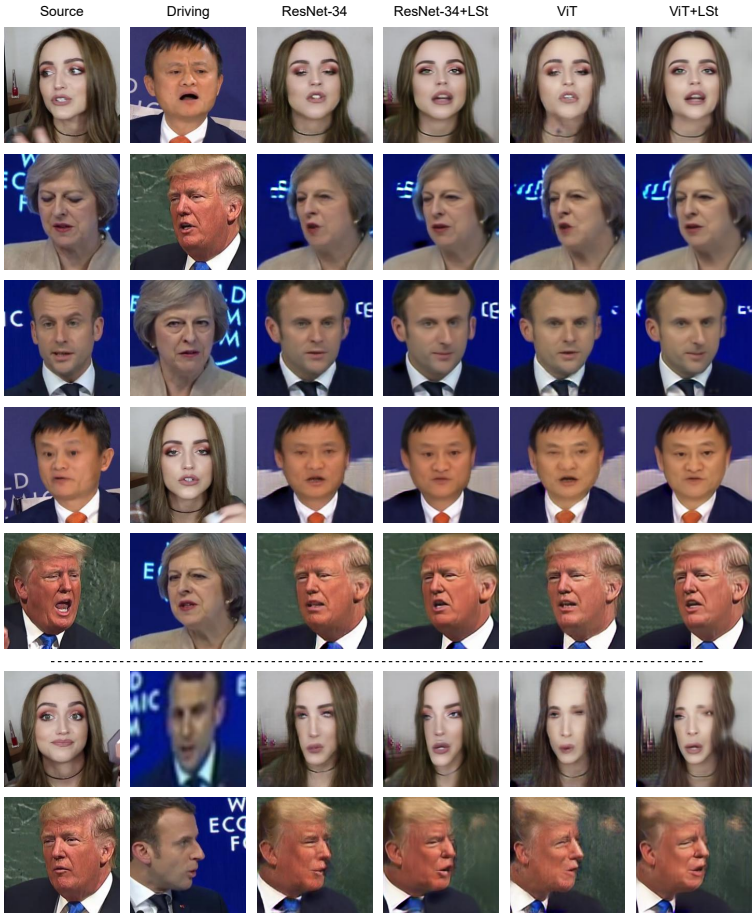

Fig. 5. Qualitative results of proposed models. The last two rows are typical failure cases, our method is sensitive to distortion in images and struggles with very large poses.

Comparison of Methods Unlike other methods that involve image warping, optical flows estimated by X2face are directly applied to images instead of fea-

ture maps, no module is responsible for refining the warped results, hence X2face generally performs poorly in recent literature. NeuralHead-FF was implemented by authors of [7] without meta-learning in the original paper [28]. As described in Section 1, NeuralHead-FF generates faces from face sketches of driving images, it does not estimate any optical flow. Mesh Guided GCN benefits from 3D models in which the driving identity is completely discarded. We believe the 3DMM parameter estimator and mesh down-sampling are the main sources of error for [27], as there is no 3DMM annotation for face reenactment datasets and loss-less down-sampling is not feasible for general surfaces [16]. MarioNETte estimates optical flows from feature maps of face sketches generated by original landmark coordinates of source and driving images, these optical flows are exerted on feature maps instead of raw images. Face sketches in MarioNETte+LT are generated from landmark coordiantes transformed by the landmark transformer. Without modification on landmark coordinates, our ResNet-34 and ViT achieve lower PRMSE than previous methods. The use of landmark style transfer boosts CSIM for both ResNet-34 and ViT, though ViT benefits more from it. Compared to the landmark transformer in [7], our landmark style transfer better promotes the quality of identity preserving but suffers more on the accuracy of pose and expression, this is because our method is more closely related to how images in the dataset are captured, and each person's preferred poses and expression intensity, landmark coordinates transformed by our method are more similar to how the person behaves in the recorded video.

Performance of Vision Transformer ResNet-34 performs slightly better than ViT, nonetheless, the shallow Vision Transformer in our proposed method obtains satisfactory results, which is on par with pure CNN methods such as [7]. Since the estimated optical flow operates on feature maps extracted by CNN, ResNet, which is also a CNN, can perceive optical flows that are more compatible with these feature maps. Current architectures of Vision Transformer make them natural feature extractors, even so, due to the need for image warping in face reenactment, CNN is still the dominant and more direct way of image synthesis, reenactment methods that adopt Vision Transformer for image generation need further study.

Effects of Landmark Reenactment Module Regarding the use of landmark style transfer, we notice the same pattern that presents in MarioNETte and MarioNETte+LT: the improvement on the quality of identity preserving comes at the cost of less accurate head poses and expressions. This is because both our method and [7] leverage face sketches generated by landmark coordinates for reenactment. When those coordinates are modified, reenacted images are subsequently altered. We notice the alteration brought by landmark style transfer favors faces in frontal view, but performs poorer when faces have large poses. For instance, in the first row of Figure 5, with the landmark coordinates of the driving image being transferred to the style of the source image, the eyes and mouth in the reenacted image are properly opened, yet in the second row of

Figure 5, the style-transferred coordinates make the mouth region less truthful to the expression in the driving image. As mentioned above, we believe the effectiveness of landmark style transfer is closely related to how training images are captured and the preferred expression intensity of each person. In terms of the strategy of placing reenacted landmarks at the same location as in the driving image, the difference in CSIM and PRMSE before and after the use of landmark style transfer indicates that this is a strong prior for lowering head pose error, but it is leaking the driving identity to the reenacted image.

4 Conclusions

In this paper, we propose a one-shot face reenactment framework in which the overall face and individual landmarks are reenacted separately. Vision Transformer is used to estimate the optical flow that warps the entire source image while landmarks are individually reenacted by corresponding sketches through CNN. We further propose landmark style transfer to alleviate the identity mismatching problem. Compared to other methods, we achieved more accurate head poses and the proposed landmark style transfer better preserves identities that other methods that also rely on facial landmark coordinates. One possible future work is to investigate the use of Vision Transformer for the image synthesis stage in face reenactment.

References

1. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science (2016)
2. 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)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
4. Cheng, Y.T., Tzeng, V., Liang, Y., Wang, C.C., Chen, B.Y., Chuang, Y.Y., Ouhyoung, M.: 3d-model-based face replacement in video. In: SIGGRAPH (2009)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: arXiv: 2010.11929 (2020)
7. Ha, S., Kersner, M., Kim, B., Seo, S., Kim, D.: Marionette: Few-shot face reenactment preserving identity of unseen targets. In: AAAI (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: arXiv: 1512.03385 (2015)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)

10. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k.: Spatial transformer networks. In: NIPS (2015)
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
12. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (2018)
13. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z.: A survey of visual transformers. In: arXiv: 2111.06091 (2021)
14. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016)
15. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: arXiv: 1511.06434 (2016)
16. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: ECCV (2018)
17. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Advances in Neural Information Processing Systems (2019)
18. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: CVPR (2019)
19. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: What makes tom hanks look like tom hanks. In: ICCV (2015)
20. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
22. Vlasic, D., Brand, M., Pfister, H., Popović, J.: Face transfer with multilinear models. ACM Trans. Graph. (2005)
23. 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)
24. Wiles, O., Koepke, A.S., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV (2018)
25. Wu, W., Zhang, Y., Li, C., Qian, C., Loy, C.C.: Reenactgan: Learning to reenact faces via boundary transfer. In: ECCV (2018)
26. Yao, G., Yuan, Y., Shao, T., Li, S., Liu, S., Liu, Y., Wang, M., Zhou, K.: One-shot face reenactment using appearance adaptive normalization. In: arXiv: 2102.03984 (2021)
27. Yao, G., Yuan, Y., Shao, T., Zhou, K.: Mesh guided one-shot face reenactment using graph convolutional networks. In: 28th ACM International Conference on Multimedia (2020)
28. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: ICCV (2019)
29. Zeng, X., Pan, Y., Wang, M., Zhang, J., Liu, Y.: Realistic face reenactment via self-supervised disentangling of identity and pose. In: AAAI (2020)