2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00299

Large Pose Friendly Face Reenactment using subtle motions

Xiaomeng Fu[1,2],Xi Wang[1*],Jin Liu[1,2],Jiao Dai[1],Jizhong Han[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{fuxiaomeng,wangxi1,liujin,liuwantao,daijiao,hanjizhong}@iie.ac.cn

***Abstract*—Face reenactment aims to synthesis a photo-realistic video of the source face by imitating the motion and expression of the driving video while keeping the source appearance (i.e. identity). Although good results have achieved recently, most state-of-the-art methods remain vulnerable to extreme conditions, which greatly restricts the application in the real world. Among various extreme conditions, the large pose problem is the most common one. We clarify that the large pose problem is mainly caused by the severe motion change between the source image and the current driving frame. An intuitive solution is to divide the severe motion change into a sequence of subtle motions. Therefore, we propose a new scheme that exploring the temporal coherence between previous neighbor frame and current frame. The smaller motion change between consecutive frames help to solve the large pose problem. Furthermore, a calibration net is designed to eliminate the error accumulation of the previous step. Extensive experiments demonstrate that our method performs better on large pose face reenactment than the state-of-the-art in terms of large pose cases and visual quality.**

***Index Terms*—Face Reenactment, Integrator, Deep Learning**

I. Introduction

Face reenactment aims to synthesis a photo-realistic video of the source face by imitating the motion and expression of the driving video while keeping the source appearance (i.e. identity). It owns a wide range of applications such as digital human, video compression, virtual reality and etc.

Face reenactment is essentially a problem of feature decoupling. A typical reenactment process is as follows. Motion features, expression features and identity features of the source and driving are first decoupled. Then, identity features of the source, motion and expression features of the driving are fused together to synthesis the reenacted face.

Most state-of-the-art methods [1]–[3] follow the above reenactment process and achieved remarkable results. However, these methods are not robust enough to handle extreme conditions, which severely hindered the practical application of these methods. Among various extreme conditions, the large pose problem is the most common one.

A typical failure case caused by large pose problem is shown in Figure 1. We choose a self-reenacted video to show the common large pose problem. In self-reenacted process, the source person is one of the driving video frames. The self-reenactment case, compared with the cross-identity reenactment, is much simpler and the driving videos are also the ground truth. It can be noticed that severe performance decrease occurs on reenacted videos when the pose divergence between the source and the driving frames comes larger. First is the woman's ear become vague and then the whole area of neck and ear comes

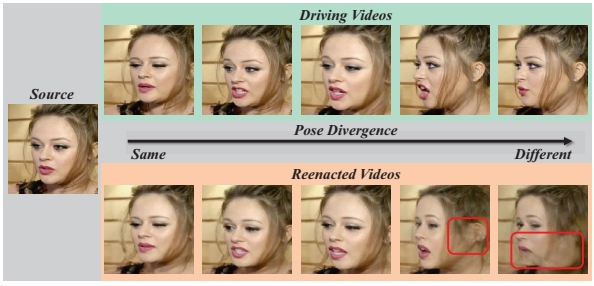


Fig. 1: **The large pose problem.** We reenact the source utilizing the same video as driving. It can be noticed that with the pose divergence getting larger, blurry regions gradually dominate the reenacted videos. We mark the blurry area with a red rectangle.

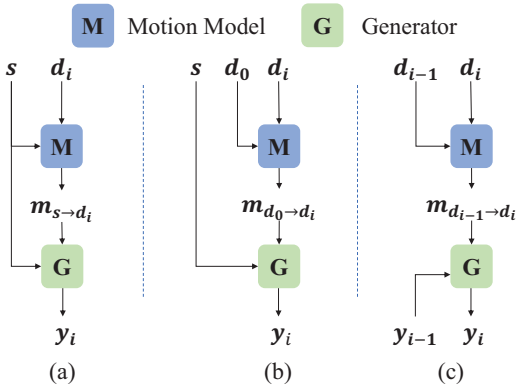


Fig. 2: Three types of inference schemes by how to use motion feature. Where d_i is the ith frame of the driving video. s represents the source face image. y_i is the current reenacted frame. M is a motion model to extract motion feature from x to y, denoted as $m_{x\to y}$. G is the generator to synthesize the final reenacted results.

This research is supported in part by the National Key Research and Development Program of China (2022YFC3302102) and the National Natural Science Foundation of China (61702502).

*Corresponding author: wangxi1@iie.ac.cn

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00299

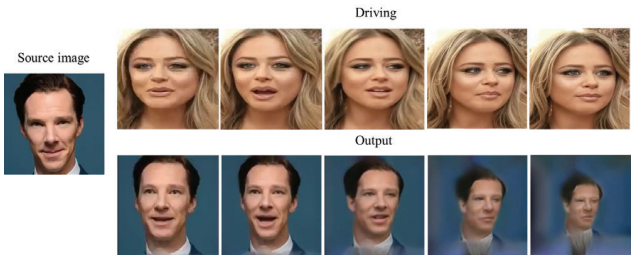


Fig. 3: Error propagation in (c). We reenact the source by the inference scheme (c). It can be seen that error propagate and accumulate over time, finally leading to corrupted outputs.

to a blur.

The mainly factor that leads to the large pose problem is the severe motion change between the source and the driving faces. Most of the existing methods directly use the motion change for generation, while neglecting the temporal coherence. We claim that severe motion change can be divided to a sequence of small motions between consecutive frames that can be easily captured. As illustrated in Figure 2, we display three types of inference schemes represented as (a), (b) and (c). (a) is the scheme adopted by most current methods, which is also the most intuitive one. (a) directly predicts the motion from the source S to the current driving frame d_i. Due to the face structure difference between the driving and the source, methods adopting (a) will suffer the identity mismatch problem when applied to cross identity reenactment. (b) predicts the motion between the d_0 and d_i, and apply the predicted motion to the source. (b) will not cause identity mismatch problem. However, Adopting these inference schemes, temporal consistency between frames is neglected and only the current driving frame will be taken into consideration. Great pose will lead to large motion change which make performance degradation.

In contrast, we propose the scheme (c). (c) receives the output y_{i-1} of the previous time step and predicts the motion between consecutive frames. Great pose changes can never occur in consecutive frames, for there are only microseconds between frames for the talking face to move. In this way, large pose problem is solved at ease. However, (c) has the problem of error propagation. Error in previous time step will be passed to the current time step. And the error will accumulate over time. As illustrated in Figure 3, the accumulated error enforces an inside motion to the talking face and leads to a corrupted output.

Therefore, we propose a method that integrates both schemes (b) and (c). We categorize (b) as a global integrator, and (c) is a local integrator. And then combine both (b) and (c) to get more accurate reenactment results. In our framework, the output of scheme (c) help to solve the large pose problem by using the subtle motion of consecutive frames instead of the severe motion. And the output of scheme (b) help to reduce the propagation of the truncation error of (c). Furthermore, we propose a calibration net to correct the mismatch in identities.

The main contributions of our work are:

- To reduce the large pose problem of face reenactment, we propose a novel scheme about using subtle motions of consecutive frames to limit the motion change between current generated frame and referent frame.
- We integrate both the local and global integrators to decrease the truncation error. In addition, a discrete latent code is introduced to limit the error propagation.
- Both qualitative and quantitative experimental results show that our method outperforms state-of-the-arts in large pose cases.

II. Related Works

As mentioned above, face reenactment tasks aims to transfer the driving motion to the source. The most intuitive idea is to decouple the motion and identity of the face and directly exchange the motion and the identity. Thus, a large part of works [4], [5] focus on how to accurately decouple the motion and identity. These methods often model identity as a specific face representation, such as facial landmark [6], keypoints or face-parsing segmentation maps. However, these face representations can only reconstruct coarse-grained face and fail to show personal characteristics, which gives rise to severe identity-mismatch problem.

To overcome the identity-mismatch problem, motion-based methods [1]–[3] are proposed. Motion-based methods adopt a two-stage strategy. They first capture the motion from the source face to the driving face, and then warp the source face according to the captured motion. The most representative method is FOMM [7], which is also a pioneering work of these methods. FOMM first separately predicts the keypoints of the source and the driving and then derives the motion from the keypoints using First Order Taylor Expansion. As a first-order expansion, FOMM models the motion as a combination of local linear motions, which does not works for all the situations. Based on FOMM, a large number of works have been proposed to get more precise motions. To model the motion of articulated objects, MRAA [1] replaces the keypoints with the segmentation region. In this way, more accurate motions are conveyed. TPSMM [3] models the motion using thin-plate splines, which can better represent higher orders of the motion. To inject 3D information into model, DAGAN [2] first predicts the depth map of the source and driving and then fuses the depth map to motion, expecting to represent motion more precisely.

However, it should be noted that accurate estimated motion does not necessarily mean a photo realistic result. As mentioned before, changes in motion will leave large exposed area to inpaint, which causes the large pose problem. In contrast, we propose a method to autoregressively synthesis the reenacted video, which avoids the problem that may be caused by the large pose.

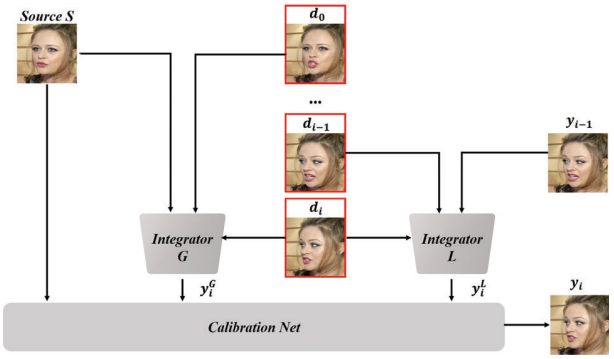


Fig. 4: The architecture of our proposed method. Our method contains three modules: the Global Integrator G, the Local Integrator L and the Calibration net C.

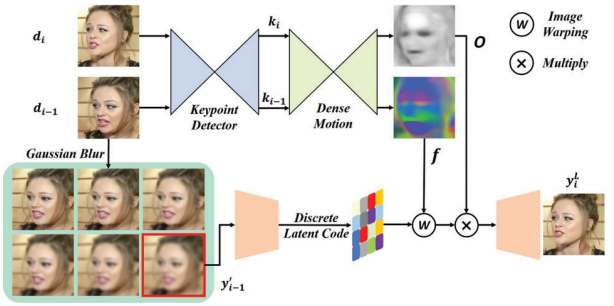


Fig. 5: The architecture of the Integrator L in the training phase.

III. Method

A. Overview

Given a source face S and a driving video $D = \{d_0, d_1, ..., d_n\}$, face reenactment aims to generate the reenacted video $Y = \{y_0, y_1, ..., y_n\}$. As illustrated in Figure 4, our method contains three main modules: the Global Integrator G, the Local Integrator L and the Calibration Net C.

For the global integrator G, we utilize d_0, d_i and S to compute y_i^G:

$$y_i^G = G(d_0, di, S) \tag{1}$$

For the local integrator L, we utilize d_{i-1}, d_i, y_{i-1} to compute y_i^L:

$$y_i^L = L(d_{i-1}, d_i, y_{i-1}) \tag{2}$$

Output Y^G suffers from the large pose problem while output Y^L are not sharp enough due to error propagation. Thus, a calibration net C is designed to integrate these two outputs. In addition, the source S is also taken as one of the inputs to provide the identity information:

$$y_i = C(y_i^G, y_i^L, S) \tag{3}$$

B. Integrators

The Integrator G and L are designed to generate coarse results. Thus, the inputs and outputs are the same as current reenactment methods. We utilize the widely used architecture which contains a keypoint detector, a dense motion network and a generator. We first introduce the architecture of Integrator L, for Integrator G can be easily derived from L by removing the Gaussian Blur and Discrete Latent modules. The architecture of Integrator L is illustrated in Figure 5. The $i-1th$ and ith of driving frames are first fed into a keypoint detector to obtain corresponding keypoints k_{i-1} and k_i. Then the keypoints k_{i-1} and k_i are fed into a dense motion network to predict the dense motion flow f and occlusion map O.

At training, the d_{i-1} also plays the role as y_{i-1} in the inference. It is first fed to a down sampling convolution module then is warped by the predicted motion flow f. Finally, under the guidance of occlusion map, exposed areas are inpainted by the up sampling module to obtain the final output y_i^L.

The inputs of Integrator L include the result of previous step which is not so clear as the source image. However, in training, we only have clear d_{i-1} to replace the blurry y_{i-1}. To ensure that the visual quality of the inputs in training and inference are almost the same, a random Gaussian Blur is added in training stage to simulate the blurry generated images y_{i-1},

$$y'_{i-1} = \text{GaussianBlur}(d_{i-1}, \sigma) \tag{4}$$

where σ is the variance randomly selected from 0 to 10.

Furthermore, we limit the error propagation in L by discrete latent coding. Simply mapping the y'_{i-1} to a continual latent space does no help to limit the error propagation. On the contrary, we map the y'_{i-1} to a discrete latent space utilizing gumbel softmax.

There is no error propagation for the single step Integrator G. Thus, we remove the discrete latent code module in Integrator G. The image distribution gap between training and inference does not exist in G as well. So we remove the Gaussian Blur in Integrator G. The other part of G keeps the same as L.

C. Calibration Net

The outputs of Integrator G and L (represented as y_i^G and y_i^L separately) have imperfections in different aspects. When handling large pose cases, the y_i^G have a good performance on global visual quality while it fails to generate good results in large pose area. The y_i^L, on the contrary, owns lower visual quality but can generate reasonable results in large pose area. Thus, we design a calibration net to decide which output (y_i^G or y_i^L) should be used for a certain area. We also hope the final output y_i keep the identity unchanged. Thus, the source face S is also added as one of the inputs for the calibration net. The detail architecture of the calibration net can be found in supplementary materials.

D. Loss Functions

We constrain the network with a reconstruction loss $\mathcal{L}_w$. During the training stage, we adopt the self-supervised training scheme, which means the driving video and the source are the

Model	CSIM(↑)	SSIM(↑)	PSNR(↑)	PRMSE(↓)	AUCON(↑)
X2face [8]	0.689	0.719	22.537	3.26	0.813
NeuralHead [9]	0.229	0.635	20.818	3.76	0.719
MarioNETte [4]	0.755	0.744	23.244	3.13	0.825
FOMM [7]	0.813	0.723	30.394	3.20	0.886
MeshG [10]	0.822	0.739	30.394	3.20	0.887
OSFV [11]	0.895	0.761	30.695	1.64	0.921
DaGAN [2]	**0.899**	0.804	31.220	1.22	**0.939**
TPSMM [3]	0.834	0.736	29.547	1.27	0.896
ours	0.892	**0.821**	**31.989**	**0.87**	0.936

TABLE I: Quantitative results on the self reenactment on VoxCeleb1 dataset.

same person. The driving video is also the ground truth. We calculate the reconstruction loss according to the ground truth:

$$\mathcal{L}_r = ||I - \hat{I}|| \tag{5}$$

where $\hat{I}$ are the predicted images and I are the ground truth image. We also utilize the perceptual loss [12] that is widely used in image synthesis. The perceptual loss is defined as:

$$\mathcal{L}_{per} = \sum_i ||\phi_i(I) - \phi_i(\hat{I})|| \tag{6}$$

where ϕ_i is the output of the i-th layer of the VGG19 network. It should be noted that both the integrator and the calibration net are expected to obtain as good synthesized results as possible. Thus, the outputs of integrators and calibration net leverage these loss functions:

$$\mathcal{L}_{net} = \lambda_{r,net}\mathcal{L}_{r,net} + \lambda_{per,net}\mathcal{L}_{per,net} \tag{7}$$

where net=L, G, C representing global integrator, local integrator and calibration net separately. λ is the corresponding weight. The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_L + \mathcal{L}_G + \mathcal{L}_C \tag{8}$$

IV. Experiment

We conduct experiments on VoxCeleb1 [13] and CelebV [14] dataset. These experiments can be divided into two categories: self reenactment and cross-identity reenactment. Self reenactment first chooses a video clip as driving. Then the first frame of the video clip is selected as the source. Therefore, self reenactment is a task of video reconstruction. In self-reenactment, the driving video also plays the role as ground truth. Different from self reenactment, cross-identity reenactment chooses a different identity as the source, and thus there is no ground truth.

A. Experimental Setups

We follow the metrics that are used in DaGAN. We utilize the structure similarity (**SSIM**) and peak signal to noise ratio (**PSNR**) to evaluate the visual quality of the generated images. We utilize the **CSIM** metric [15] to evaluate the identity preserving ability of the model. The head Pose Root Mean Square Error (**PRMSE**) and Action Unit Consistency (**AUCON**) [16] are used to evaluate the head pose and expression accuracy of the reenacted results.

Comparison Methods. We compare the proposed method with DaGAN [2] and TPSMM [3], which are the most-of-the-art methods. We also compare with both motion-based and decouple-based methods. The motion-based methods are X2face [8], FOMM [7], OSFV [11]. The decoupling-based methods are NeuralHead [9], MarioNETte [4] and MeshG [10].

Implementation Details. All models are implemented by Pytorch. We train the whole model in an adversarial way [17]. A patch discriminator [18] is used to help to get high fidelity results. Details of our network and the training methods can be found in the supplementary materials.

B. Quantitative Comparisons.

We train the models on VoxCeleb1 datasets and conduct both self reenactment and cross-identity reenactment to evaluate the effectiveness of our proposed method. The quantitative results of self reenactment on VoxCeleb1 are shown in Table I. Our method gets the highest scores in SSIM, PSNR metrics, outperforming both motion-based and decouple-based methods, which demonstrates the effectiveness of our proposed method. Though DaGAN outperforms our method in CSIM and AUCON metrics, the difference is very small. It should be noted that our method gets great progress in PRMSE metrics, which demonstrates that our method can perfectly track the head motion of the driving.

We directly test these models on CelebV without finetuning to evaluate the performance of the models reenacting more complex sources. The quantitative results of cross-identity reenactment on CelebV are shown in Table II. We get the highest scores on PRMSE and comparable results on AUCON metric. It should be noted that our method are weak to preserve the source identity in CelebV dataset. (only 0.676 in CSIM) We explain that CelebV dataset has poor temporal coherence. More specifically, the image quality of the CelebV dataset varies greatly with time.

Ablation analysis. In this paper, our mainly contribution is to propose a new framework to utilize subtle motions to solve the large pose problem. In our method, the FOMM [7] is used as our Integrator G. It can be found that our subtle motions framework can boost the model over all matrices especially for the PRMSE which reflects the head pose accuracy.

Model	CSIM(↑)	PRMSE(↓)	AUCON(↑)
X2face [8]	0.450	3.62	0.679
NeuralHead [9]	0.108	3.30	0.722
marioNETte [4]	0.520	3.41	0.710
FOMM [7]	0.462	3.90	0.667
MeshG [10]	0.635	3.41	0.709
OSFV [11]	**0.791**	3.15	0.805
DaGAN [2]	0.723	2.33	**0.873**
TPSMM [3]	0.703	2.17	0.843
ours	0.676	**1.85**	0.843

TABLE II: Quantitative results of cross-identity reenactment on CelebV dataset.

Fig. 6: Qualitative comparisons of self reenactment on the VoxCeleb1. We specially select results for large pose cases.

C. Qualitative Comparisons.

We conduct both self reenactment and cross-identity reenactment in VoxCeleb1 dataset. The self reenactment results are shown in Figure 6. It can be seen that our method generates more realistic images. The face geometry is well kept even the head pose changes drastically.

The cross-identity reenactment results are shown in Figure 7(a). It can be seen that methods that only take one frame into account are vulnerable to large pose changes. Our method takes consecutive frames and is robust to large pose changes. More results can be found in the supplementary materials. We also conduct cross-identity reenactment in CelebV dataset to evaluate the performance of the models reenacting more complex sources. The results are shown in Figure 7(b). Previous methods fail in extreme head pose while our method can keep the face geometry undamaged.

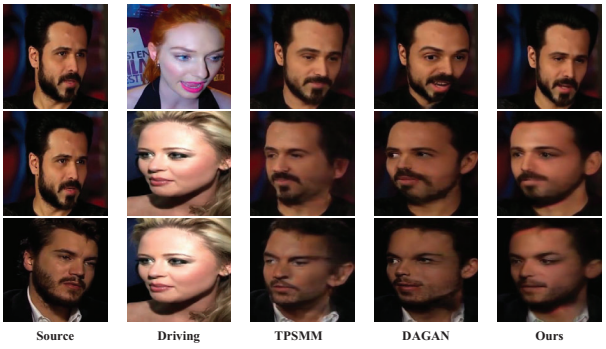
	CSIM	PRMSE	AUCON	PRMSE	AUCON
Intergrator G only	0.813	0.723	30.394	3.20	0.886
ours	0.892	0.821	31.989	0.87	0.936

TABLE III: Ablation studies on VoxCeleb1 dataset.

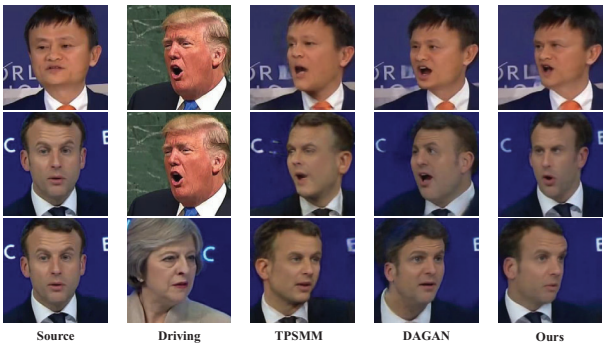
D. Ablation Study.

We conduct ablation study on VoxCeleb1 dataset. The quantitative results of self reenactment are shown in Table III. It can be found that our framework can boost the model over all matrices especially for the PRMSE which reflects the head pose accuracy. We do not conduct ablations without Global integrators because accumulated errors will lead to corrupted results if we only use integrator L (as shown in Figure 3).

Furthermore, as our main contribution is to propose a new framework to utilize subtle motions to solve the large pose problem. Actually, our framework can also improve the identity preservation ability as the framework is easy to combined with other reenactment methods. As shown in Table IV, in the original Integrator G (Integrator G only (w/o our L&C)), we do not use any facial structure priors or deep 3D features, which leads worse CSIM. By using our subtle motions framework, the CSIM and AUCON also got significant improvement. And



(a) On the VoxCeleb1 dataset

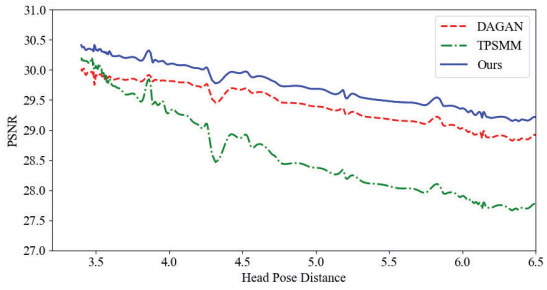


(b) On the CelebV dataset

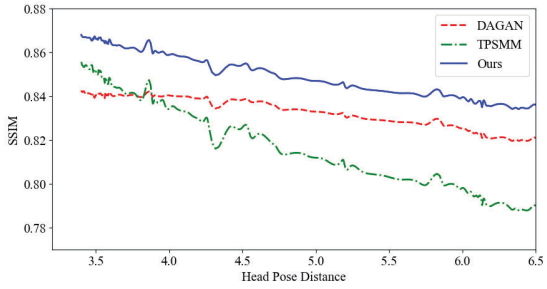
Fig. 7: Qualitative comparisons of cross-identity reenactment on two datasets. We select results for large pose cases.

Model	CSIM	PRMSE	AUCON
DaGAN	0.723	2.33	0.873
TPSMM	0.703	2.17	0.843
Integrator G only	0.462	3.90	0.667
ours	0.676	1.85	0.843
ours+DaGAN	0.735	1.97	0.888

TABLE IV: Further experiments on effectiveness of extra priors on VoxCeleb1 dataset.



(a)



(b)

Fig. 8: Performance for large pose cases. We self reenact the first frame of the driving video and compute the PRMSE as the x-axis. The PSNR/SSIM metric is y-axis in (a)/(b).

the integrator G can be easy replaced with other methods. We inject the face priors by simply replacing the Integrator G with DaGAN, the results show that all score is getting better, especially PRMSE.

E. Performance in Large Pose Cases.

In Figure 8, we show how large pose influences the performance of reenactment methods. We choose the PRMSE as the head pose distance and take it as the x-axis. We choose PSNR and SSIM as the y-axis. Then we plot out the relation between head pose distance and metrics of visual quality. The x-axis starts with 3.3 to focus on large pose cases. It can be seen that both of these methods perform worse as the head pose distance increases (in both PSNR and SSIM metrixs). Comparing Figure 8(a) and 8(b), it can be concluded that TPSMM performs well in common cases but are especially vulnerable to large pose cases. DAGAN performs better than TPSMM but is inferior to our method, which demonstrates the effectiveness of our proposed method.

V. Conclusion

In this paper, we discuss the large pose problem in face reenactment. Instead of using severe motion between the source frame and the current driving frame, we divide the severe motion into a sequence of subtle motions to achieve better reenactment results. Both quantitative and qualitative results demonstrate that our method is more robust to large pose cases.

References

[1] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 653–13 662.

[2] F.-T. Hong, L. Zhang, L. Shen, and D. Xu, "Depth-aware generative adversarial network for talking head video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3397–3406.

[3] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3657–3666.

[4] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 10 893–10 900.

[5] M. Meshry, S. Suri, L. S. Davis, and A. Shrivastava, "Learned spatial representations for few-shot talking-head synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 829–13 838.

[6] G.-S. Hsu, C.-H. Tsai, and H.-Y. Wu, "Dual-generator face reenactment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 642–650.

[7] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] O. Wiles, A. Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686.

[9] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9459–9468.

[10] G. Yao, Y. Yuan, T. Shao, and K. Zhou, "Mesh guided one-shot face reenactment using graph convolutional networks," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1773–1781.

[11] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 039–10 049.

[12] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[14] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, "Reenactgan: Learning to reenact faces via boundary transfer," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 603–619.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[16] B. Amos, B. Ludwiczuk, M. Satyanarayanan *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, p. 20, 2016.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.