Supplementary Materials: QS-Craft: Learning to Quantize, Scrabble and Craft for Conditional Human Motion Animation

Yuxin Hong[1], Xuelin Qian*[1], Simian Luo[1], Guodong Guo[3], Xiangyang Xue[1,2], and Yanwei Fu[1]

[1] School of Data Science, and MOE Frontiers Center for Brain Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
{20210980140,xlqian,18300180157,yanweifu}@fudan.edu.cn
[2] School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
xyxue@fudan.edu.cn
[3] Baidu
guoguodong01@baidu.com

Supplementary Materials

The supplementary document is organized as follows:

- Sec. A elaborates the architecture of the overall framework.
- Sec. B describes the formulations of loss objectives.
- Sec. C demonstrates face animation results on VoxCeleb dataset.
- Sec. D provides more animation results on Tai-Chi-HD dataset.
- Sec. E reports more experimental results on Penn Action dataset.
- Sec. F shows more visualizations of our proposed QS-Craft.
- Sec. G shows more qualitative and quantitative results of our proposed QS-Craft and other new baselines.

A. Network Architectures

Encoder & Decoder. The architecture of Encoder and Decoder used in our task follows VQGAN [1] with its ImageNet pre-trained parameters. The Encoder and Decoder have a symmetric structure with 9 layers. Take Encoder for example, we first use 2D Convolution layer to extract the feature map, then for the 2-5th layer, we use 2 ResBlock and one DownBlock to downsample the feature map by half in each layer. Then, we adopt AttnBlock and ResBlock in the 6-7th layer followed by GroupNorm layer and Swish activation. In detail, ResBlock consists of two stacked convolution layers along with a skip connection layer, while the AttnBlock has a multi-head self-attention to capture global information. At last, 2D Convolution layer is applied to obtain the final feature map.

* corresponding author

Table 1. The architecture of Encoder and Decoder

Encoder	Decoder
Conv2d ($256 \times 256 \times 128$)	Conv2d ($16 \times 16 \times 512$)
ResBlock×2 + DownBlock ($128 \times 128 \times 128$)	ResBlock + AttnBlock + ResBlock ($16 \times 16 \times 512$)
ResBlock×2 + DownBlock ($64 \times 64 \times 128$)	ResBlock×3 + AttnBlock×3 + UpBlock ($32 \times 32 \times 512$)
ResBlock×2 + DownBlock ($32 \times 32 \times 256$)	ResBlock×3 + UpBlock ($64 \times 64 \times 256$)
ResBlock×2 + DownBlock ($16 \times 16 \times 256$)	ResBlock×3 + UpBlock ($128 \times 128 \times 256$)
ResBlock×2 + AttnBlock×2 ($16 \times 16 \times 512$)	ResBlock×3 + UpBlock ($256 \times 256 \times 128$)
ResBlock + AttnBlock + ResBlock ($16 \times 16 \times 512$)	ResBlock×3 ($256 \times 256 \times 128$)
GroupNorm + Swish ($16 \times 16 \times 512$)	GroupNorm + Swish ($256 \times 256 \times 128$)
Conv2d ($16 \times 16 \times 256$)	Conv2d ($256 \times 256 \times 3$)

The detailed architectures of Encoder and Decoder are presented in Tab. 1. Additionally, in Quantize stage, the size of the codebook is set as 16384 with 256 channels.

Discriminator. Discriminator is used to help generate high-quality and authentic reconstructed images. The architecture of discriminator follows [3], which is a patch-based model.

Transformer. Transformer is used to learn Scrabble rule conditioned on motion information from driving videos. We adopt the decoder-only based transformer - GPT2 [9]. We use the transformer with 12 layers, and in each embedding layer, the dimension is kept as 768. For multi-head self-attention layers, 12 attention heads are adopted to capture the connection between image feature and conditional pose information. Dropout is attached to each multi-head self-attention layer with a rate of 0.1.

B. Loss Functions

Considering both the reconstruction quality of images and patchworks, we design two losses in the first stage,

$$\mathcal{L}_{Recon} = \mathcal{L}_G + \mathcal{L}_S \tag{1}$$

Specifically, $\mathcal{L}_G$ is a generative adversarial loss, which aims to play a min-max game with the generator and discriminator. The formulation can be written as,

$$\begin{aligned} \mathcal{L}_G = \ & logD\left(x_s\right) + log\left(1 - D\left(\hat{x_s}\right)\right) \\ & + logD\left(x_t\right) + log\left(1 - D\left(\hat{x_t}\right)\right) \end{aligned} \tag{2}$$

where $\hat{x_s}$ and $\hat{x_t}$ represent the reconstructed source and target images, respectively. For the second term $\mathcal{L}_S$, it is designed to supervise the quality of reordered patchwork,

$$\begin{aligned} \mathcal{L}_S = \ & \beta||sg\left[z_s\right] - \hat{z_s}||_2^2 + ||z_s - sg\left[\hat{z_s}\right]||_2^2 \\ & + \beta||sg\left[z_t\right] - \hat{z_t}||_2^2 + ||z_t - sg\left[\hat{z_t}\right]||_2^2 \end{aligned} \tag{3}$$

where $sg\left[\cdot\right]$ denotes the stop-gradient operation and $\beta = 0.25$ is a weighting factor [1]. $\mathcal{L}_{codebook}$ means a codebook learning loss, we use the same one as [7].

Fig. 1. Face animation results on VoxCeleb. Best viewed in color and zoom in.

To train the transformer model, we adopt the NLL loss for learning the rule of Scrabble. Denote the predicted likelihood of each target index as $p\left(t|s,c\right)$, the loss function in the second stage can be defined as,

$$\mathcal{L}_{Trans} = -\mathbb{E}_{t\sim p(t|s,c)} \log p\left(t|s,c\right) \tag{4}$$

where s and c mean the source embedding and the condition embedding, respectively. Besides, Mask R-CNN [2] is applied to extract the foreground masks for the strategy of RoI weight. Masks are used as guidance to re-weight the NLL loss for each predicted index. We empirically set the weight of the foreground as 1.6 and the background as 0.4.

C. Face Animation Results on VoxCeleb

In this section, we show some animation results of our QS-Craft on the face dataset. We emphasize that the experiment here is just to further demonstrate the effectiveness and generalizability of our proposed method. Due to face privacy issue, VoxCeleb thus has less priority as the benchmark in the main paper.

Dataset. VoxCeleb dataset [6] is a face dataset of $22,496$ videos extracted from YouTube videos. Following [12], we do pre-processing to obtain $12,331$ training videos and 444 test videos, with lengths varying from 64 to 1024 frames and the resolution of 256×256.

Results. We compare the qualitative results with FOMM [12] and MRAA [13] in Fig. 1. As observed, our animated results are more realistic and can mimic subtle expressions in driving videos. For example, in the top left subgraph of Fig. 1 our QS-Craft can generate mouth movements that are closer to the driving compared to other two baselines. In general, our method can maintain the identity information well and catch more detailed facial expressions.

Fig. 2. More animation results on Tai-Chi-HD. Best viewed in color and zoom in.

D. More Animation Results on Tai-Chi-HD

In this section, more qualitative animation results on Tai-Chi-HD dataset are demonstrated in Fig. 2, 3 and 4. We compare with three outstanding competitors, including Monkey-Net [11], FOMM [12], and MRAA [13].

As we can see, **(1)** Monkey-Net fails to complete the animation due to high resolution (256×256). For example, in the right part of Fig. 2, human poses in the generated results from Monkey-Net remain unchanged. **(2)** For FOMM, it can capture the change of poses between the source image and driving frames, however, when encountering the situation of large pose changes, FOMM is inclined to generate twisted and blurry animated images. **(3)** MRAA can animate the source image according to driving video in general. Though its proposed module captures precise motion of each local part, the deformation from the sparse flow to dense makes it tend to ignore fine details after warping. For example, in the left part of Fig. 3, MRAA loses many hand movements, while our QS-Craft preserves these details. **(4)** Compared with these baselines, our proposed method can capture the large pose changes as well as maintain image details. More animation results on Taichi-HD are shown in Fig. 4.

E. More Pose Guided Results on Penn Action

In this section, we report more qualitative pose-guided results on Penn Action(PA) Dataset. We compare with competitors containing PG^2 [5], PATN [18], PN-GAN [8] and MR-Net [16], as shown in Fig. 10.

For PATN, the generated images tend to be unrealistic and unreasonable compared to the targets as shown in the third column of Fig. 10. For PN-GAN, it just copies source pose directly for all the cases (in the sixth column). Additionally, it is apparent that PG^2 in most cases generates blurred images, failing

Fig. 3. More animation results on Tai-Chi-HD. Best viewed in color and zoom in.

to capture the change of pose between the source image and target image. Finally, for MR-Net, the area around the generated human is blurred and it tends to fail in some difficult situations(*e.g.*, barbell not lifted in the third row). Compared to these models, our QS-Craft demonstrates its efficiency in generating the accurate human poses guided by conditions and keeping images realistic.

F. More Visualizations of Our QS-Craft

In this section, we visualize the conditioned pose sequences and animation results in Fig. 7. The first row shows the source image and generated video sequences given the conditioned pose sequences. The second row shows the change of one pose sequence. As expected, QS-Craft model can capture the conditioned motion information and generate reasonable images with just these simple key points sequences while other baselines need to generate more complex flow maps as guidance. We also present qualitative results in two different cases: (1) *misalignment*: human poses in the source image and the first driving frame are NOT aligned. As shown in Fig. 8, our QS-Craft outperforms FOMM which cannot generate accurate and reasonable animation results. (2) *alignment*: human poses in the source image and the first driving frame are aligned. Visualizations are illustrated in Fig. 9. We observe that FOMM can indeed capture driving pose information but suffer from lower image quality and loss of motion details compared to our results.

G. More Qualitative and Quantitative Results

In this section, we add more qualitative and quantitative results of our QS-Craft and other new baselines including DAM [14], TPS [17] and NTED [10]. Moreover, new metrics like FVD [15] and WarpError [4] are also involved to

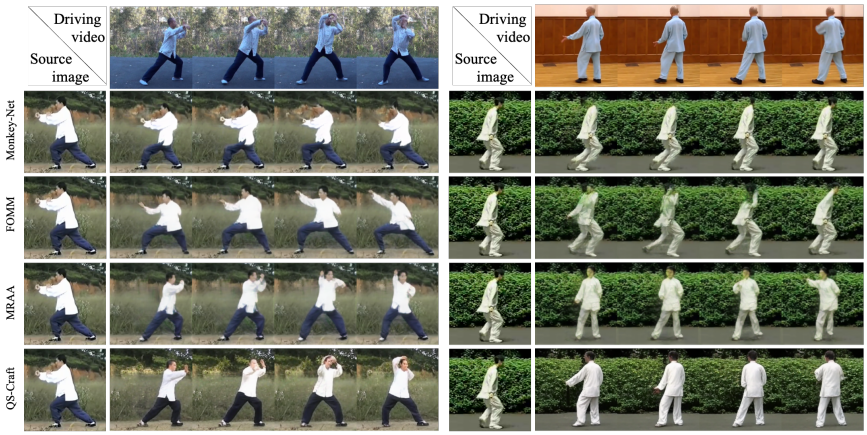

Fig. 4. More animation results on Tai-Chi-HD. Best viewed in color and zoom in.

Table 2. Results of animation on TaiChi dataset and pose-guided generation on PA dataset.

Method	Tai-Chi-HD			Penn Action		
	AKD	MKR	FID	AKD	MKR	FID
DAM [14]	5.95	0.025	49.77	-	-	-
TPS [17]	4.80	0.019	37.28	-	-	-
NTED [10]	-	-	-	32.92	0.477	63.41
Ours	**4.61**	**0.017**	**25.06**	**16.36**	**0.121**	**30.14**

Table 3. Results of animation on TaiChi dataset with video metrics.

Method	FVD	WarpError
MRAA [13]	2056.32	0.0011
TPS [17]	1154.06	**0.0010**
Ours	**1113.18**	0.0015

further demonstrate our superiority. Finally, we show some animation results under some extreme cases, especially the different size of the person in the source and another person in driving images. All the figures and tables can be found in the following two tables (Tab. 2 and Tab. 3) and two figures (Fig. 5 and Fig. 6).

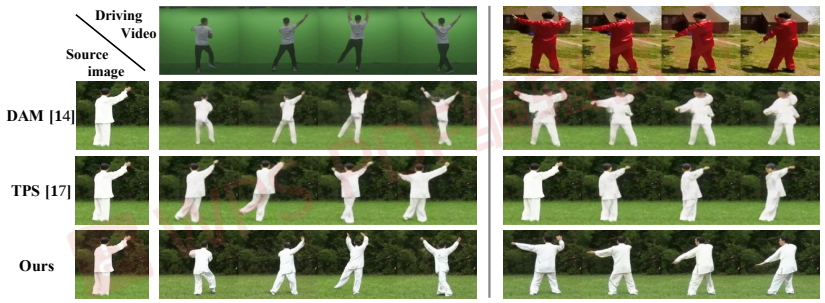

Fig. 5. Animation with online (left) and TaiChi images (right) as driving video. Please zoom in.

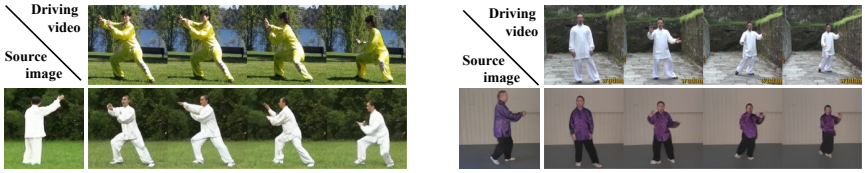

Fig. 6. Animation of extreme cases (left: back-to-front turns, right: scale size changes). Please zoom in.

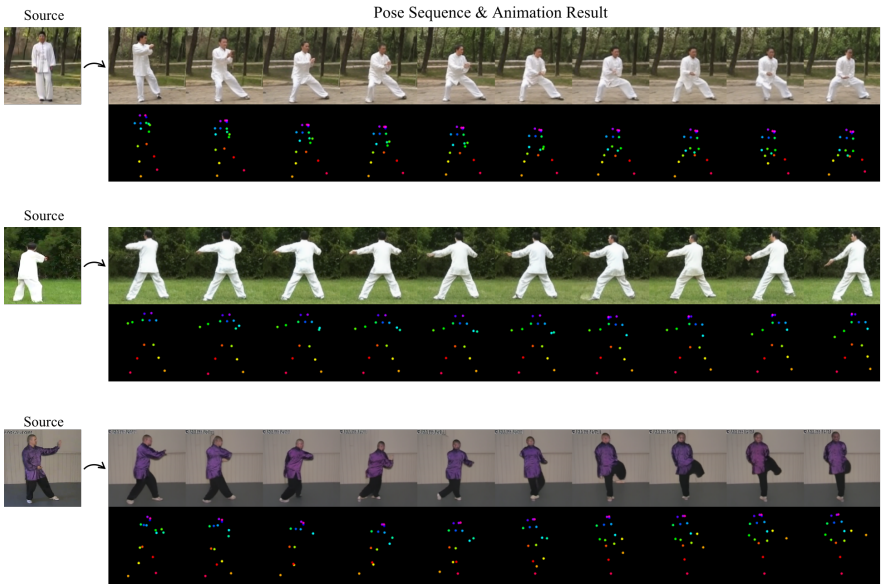

Fig. 7. Visualization results of animation on Tai-Chi-HD. In each example, the first row shows the source image and generated animation sequence. The second row is the given target pose sequence. Notice that our method can capture the conditioned motion information well.

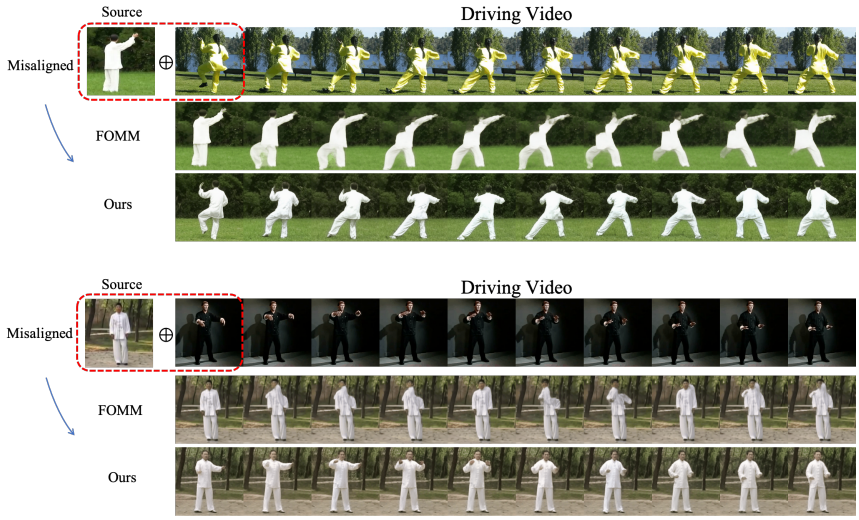

Fig. 8. Visualization results of animation on Tai-Chi-HD in the case of *misalignment*. In each example, human poses in source image and the first driving frame are misaligned. Notice that our method can capture the conditioned motion information well and generate better quality images.

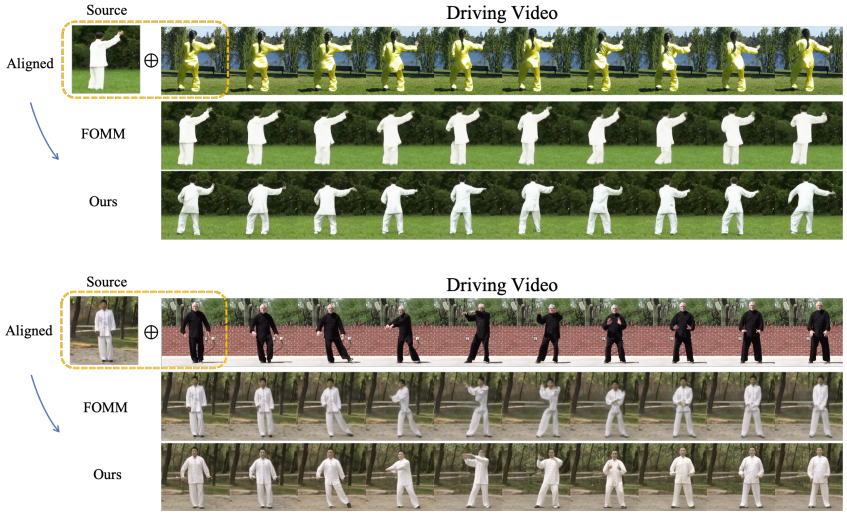

Fig. 9. Visualization results of animation on Tai-Chi-HD in the case of *alignment*. In each example, human poses in source image and the first driving frame are aligned. Notice that our method can capture more detailed motion information and generate better quality images.

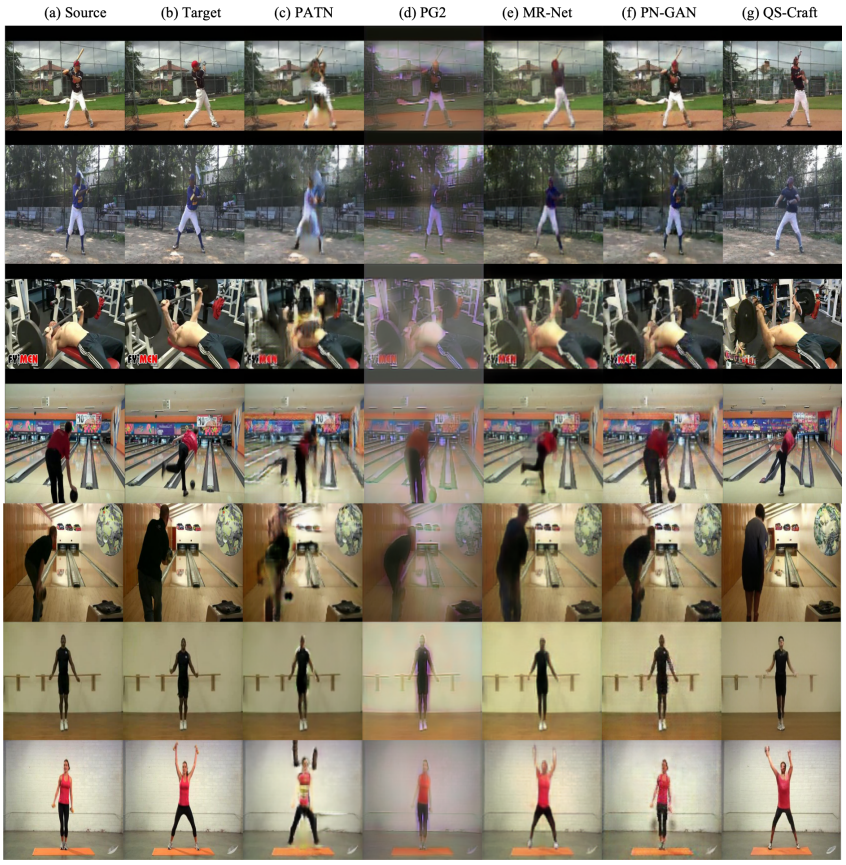

Fig. 10. More pose-guided results on PA dataset. Best viewed in color and zoom in.

References

1. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
2. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
4. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)
5. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. arXiv preprint arXiv:1705.09368 (2017)
6. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)
7. Oord, A.v.d., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. arXiv preprint arXiv:1711.00937 (2017)
8. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV). pp. 650–667 (2018)
9. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
10. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13535–13544 (2022)
11. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2377–2386 (2019)
12. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32**, 7137–7147 (2019)
13. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13653–13662 (2021)
14. Tao, J., Wang, B., Xu, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Structure-aware motion transfer with deformable anchor model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3637–3646 (2022)
15. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
16. Xu, C., Fu, Y., Wen, C., Pan, Y., Jiang, Y.G., Xue, X.: Pose-guided person image synthesis in the non-iconic views. IEEE Transactions on Image Processing **29**, 9060–9072 (2020)
17. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3657–3666 (2022)
18. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2347–2356 (2019)