Supplementary Material for Motion Transformer for Unsupervised Image Animation

Jiale Tao[1*], Biao Wang[2], Tiezheng Ge[2], Yuning Jiang[2], Wen Li[1†], and Lixin Duan[1]

School of Computer Science and Engineering & Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
{jialetao.std, liwenbnu, lxduan}@gmail.com
Alibaba Group
{eric.wb, tiezheng.gtz, mengzhu.jyn}@alibaba-inc.com

Fig. A1. Qualitative comparisons on the VoxCeleb, TaichiHD and TEDTalks datasets. We present results in a similar layout with that of the Figure 2 of our main paper, while different identities are showed here.

In this supplementary, we provide additional qualitative comparison between our proposed approach and the baseline methods on three benchmark datasets, and additional visualizations on the learned motion masks.

1 Qualitative Comparison with Videos

In addition to the qualitative results of the standard motion transfer in the main paper, we additionally provide video results on both the standard motion transfer and the relative motion transfer, to more intuitively compare our proposed motion transformer approach with baseline methods on three benchmark datasets. The video files are attached in the supplementary material package, named by TaichiHD.mp4, TEDTalks.mp4, VoxCeleb.mp4 and Relative.mp4 for the TaichiHD, TEDTalks, VoxCeleb datasets on the standard motion transfer and the three datasets on the relative motion transfer, respectively.
Standard motion transfer: We summarize part of the video frames in Fig. A1. For the VoxCeleb dataset, our method synthesizes the most realistic and detailed expression information, while FOMM [1] and MRAA [2] often fail to preserve

Fig. A2. Qualitative comparison of the relative motion transfer, we show examples on the TEDTalks, TaichiHD and VoxCeleb dataset. For clearness, we present the first frame and an another frame of the driving video, named as "First driving" and "driving" respectively, between them the relative motion is transferred to the source images.

the person identity information in the source image, and suffers in capturing the detailed expression information in the driving faces (such as eyes and mouths).

For the TEDTalks dataset, both FOMM and MRAA cannot properly synthesize the hands of speakers, especially in the situation of large motions, while our method can well estimate the hand motions between the driving video and source images.

For the TaichiHD dataset, it can be seen that videos generated by FOMM often shake at the area nearby human bodies. And compared to MRAA, our method generates more stable results and correspondingly with less artifacts.

Relative motion transfer: Our method shares a similar scheme with FOMM [1] on the relative motion transfer, where the motion between a driving image and the source image is not directly computed, instead the relative motion between the driving video and its first frame is transferred to the source image. For more details of the relative motion transfer, we refer readers to FOMM [1]. We here make a comparison on the relative motion transfer on the three datasets, as shown in Fig. A2.

We can take an intuitive understanding on the relative motion transfer from the TEDTalks dataset, where in the driving image, the motion of the man's two hands are different, *e.g.*, the man's left hand has a larger vertical translation than the man's right hand. As a result, the two hands in the driving image are at the same horizontal position, yet in the generated result they should be at different horizontal level since their motion relative to the first driving frame is different. This could be obviously observed in the last two-row results of our method, and the results generated from FOMM and MRAA are often with more artifacts (*e.g.*, the missing hands).

When the poses are similar between the first driving frame and the source image (*e.g.*, the TaichiHD dataset in Fig. A2), the relative motion transfer works

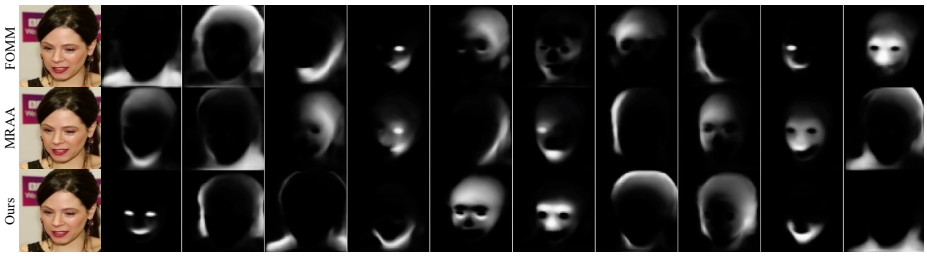

Fig. A3. Qualitative comparisons of learned motion masks on the VoxCeleb dataset. From top to bottom, we present the motion masks learned by FOMM, MRAA and our method. We further present the driving frame in the first column.

similar with the standard motion transfer while preserves the source shape well. It can be seen that our method also behaves better on this dataset.

On the VoxCeleb dataset, it is worth noting that, when there exists some-extent gaps between the initial poses of the first driving frame and the source image (*e.g.*, the first two examples), our method still performs well for the motion transfer and surpasses the existing methods [1,2].

In summary, the proposed motion transformer generally better handles global and local motions comparing to existing methods.

2 Additional Visualizations

Motion masks: To understand the work mechanism of the global motion patterns learned by our motion transformer, we visualize and compare the motion mask M^k in Equation (2) of the main paper. The motion mask M^k reflects the effective area of the k-th motion pattern. This means that, in the generated image, the masked area is obtained by warping the source image according to the part motion $\mathcal{T}^k_{S\leftarrow Z}(c)$ in Equation (1) of the main paper. We use the results on the VoxCeleb dataset to facilitate the analysis as shown in Fig. A3. We can observe that, our motion transformer generally attempts to learn a global symmetric motion pattern for all motion parts, even though some parts may overlap to some extent. While FOMM and MRAA are inclined to learn relatively irregular spatial parts. We argue that it is more reasonable to learn the facial motions in a global-aware fashion, since all motion parts should in fact share a global head rotation.

References

1. 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) 1, 2, 3
2. Siarohin, A., Woodford, O., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: CVPR (2021) 1, 3