Supplementary Material: Thin-Plate Spline Motion Model for Image Animation

Jian Zhao Hui Zhang
School of Software, BNRist, Tsinghua University, Beijing, China
zhaojian20@mails.tsinghua.edu.cn huizhang@tsinghua.edu.cn

1. Ablation study on the number of keypoints.

In the paper, we predict $K \times N$ keypoints $(N = 5)$ for both $\mathbf{S}$ and $\mathbf{D}$ to generate K TPS transformations. We did additional ablation experiments on the TaiChiHD [6] dataset with video reconstruction task on the number of keypoints to demonstrate that $K \times 5$ pairs of keypoints achieve the best motion transfer performance. For comparison, We predict $K \times 3$ and $K \times 8$ pairs of keypoints respectively to generate K TPS transformations. Tab. 1 shows the results. Because the more keypoints are predicted, the more difficult it is for the Keypoint Detector to predict them accurately, motion transfer using $K \times 8$ pairs keypoints does not perform well. On the other hand, TPS transformations generated by fewer keypoints are not flexible enough for representing motions. Therefore, $K \times 5$ is an appropriate number of predicted keypoints.

	$\mathcal{L}_1$	(AKD, MKR)	AED
$K \times 3$	0.0459	(5.00, 0.021)	0.1577
$K \times 5$	**0.0452**	(**4.57**, **0.018**)	**0.1507**
$K \times 8$	0.0457	(4.92, 0.023)	0.1538

Table 1. Ablation study on the number of keypoints with $K = 10$. (Lower is better, best result in bold)

2. Implementation details

We modify and extend the architecture of MRAA [7]. For the Keypoint Detector and the BG Motion Predictor, we employ the architecture of ResNet18 [1] and modify the number of neurons in the fully connected layer to $K \times N \times 2$ for the Keypoint Detector and 6 for the BG Motion Predictor. We use the hourglass [4] architecture for the Dense Motion Network and the Inpainting Network, and their encoders have five and three "Convolution - InstanceNorm [8] - ReLU - AvgPooling" blocks, respectively. The decoder of the Dense Motion Network consists of five "Upsampling - Convolution - InstanceNorm - ReLU" blocks and the decoder of the Inpainting Network is described in the paper. Our method is trained using Adam [2] optimizer with learning rate $2e-4$, $\beta = (0.5, 0.999)$ and batch size 28 for VoxCeleb [3], TaiChiHD [6], MGif [5], 12 for TED-talks [7]. We trained 100 epochs on each dataset and decayed the learning rate to 0.1 times once the number of epochs reached 70 and 90.

The architecture of the shape-pose disentanglement network is the same as that in MRAA [7]. Both shape and pose encoders consist of three "Linear - BatchNorm1D - ReLU" blocks and a linear layer with 64 neurons. The decoder receives the concatenation of the two latent feature maps, which consists of three "Linear - BatchNorm1D - ReLU" blocks and a linear layer with $K \times N \times 2$ neurons. The network is trained using Adam [2] optimizer with learning rate $1e-3$, $\beta = (0.5, 0.999)$ and batch size 256. We trained 60 epochs and decayed the learning rate to 0.1 times once the number of epochs reached 40 and 50.

3. Bad cases

Both MRAA [7] and our approach cannot perform well when an extreme identity mismatch occurs. Fig. 1 shows an example on MGif [5] dataset.

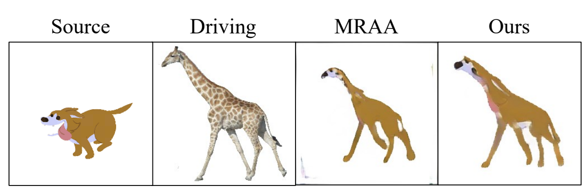

Figure 1. A bad case on MGif dataset.

4. Additional qualitative comparisons

Figs. 2 to 9 show additional qualitative comparisons on TaiChiHD [6], TED-talks [7], VoxCeleb [3] and MGif [5]. Both MRAA [7] and our method use the *avd* mode to generate image animation.

Figure 2. Qualitative comparisons on TaiChiHD.

Figure 3. Qualitative comparisons on TaiChiHD.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 1

[3] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 1

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 1

[5] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[6] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019. 1

[7] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[8] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016. 1

Figure 4. Qualitative comparisons on TED-talks.

Figure 5. Qualitative comparisons on TED-talks.

Figure 6. Qualitative comparisons on VoxCeleb.

Figure 7. Qualitative comparisons on VoxCeleb.

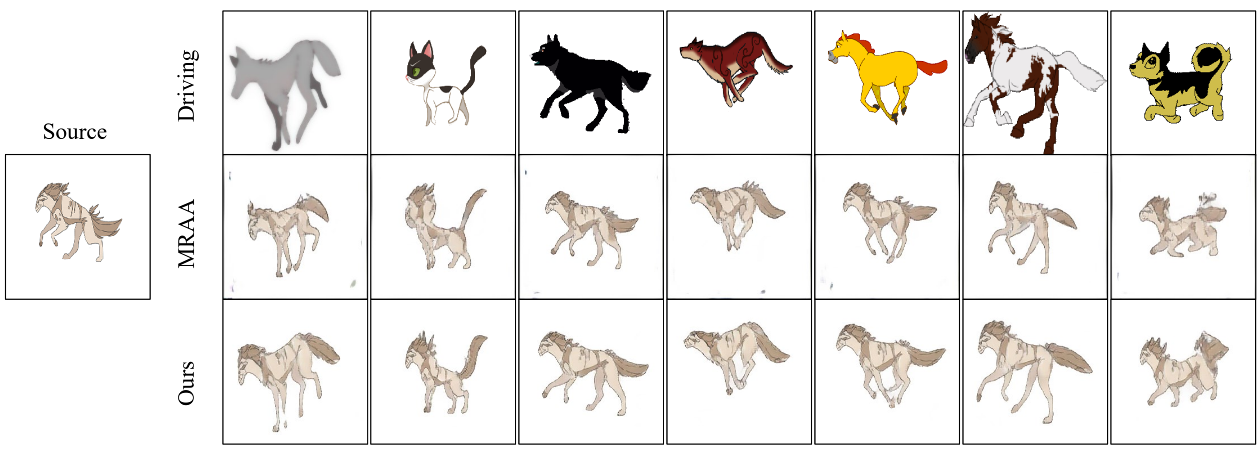

Figure 8. Qualitative comparisons on MGif.

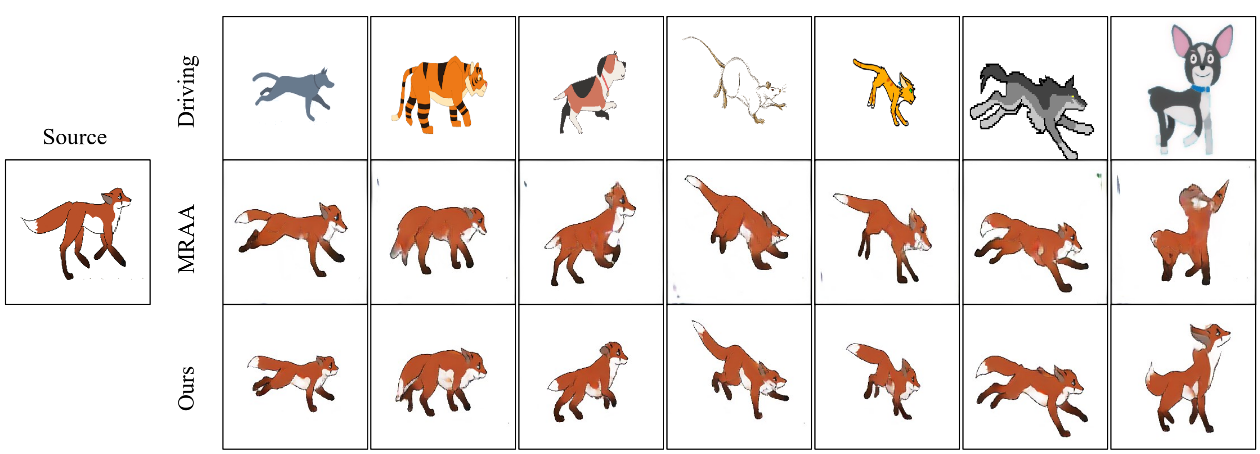

Figure 9. Qualitative comparisons on MGif.