1. Ablation study on the number of keypoints.

In the paper, we predict $K \times N$ keypoints ($N = 5$) for both $S$ and $D$ 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.

<table>
<thead>
<tr>
<th>$K \times N$</th>
<th>$\mathcal{L}_1$ (AKD, MKR)</th>
<th>AED</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K \times 3$</td>
<td>0.0459 (5.00, 0.021)</td>
<td>0.1577</td>
</tr>
<tr>
<td>$K \times 5$</td>
<td><strong>0.0452</strong> (4.57, 0.018)</td>
<td><strong>0.1507</strong></td>
</tr>
<tr>
<td>$K \times 8$</td>
<td>0.0457 (4.92, 0.023)</td>
<td>0.1538</td>
</tr>
</tbody>
</table>

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.

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


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.