Let's All Dance: Enhancing Amateur Dance Motions (Supplementary Material)

Qiu Zhou, Manyi Li, Qiong Zeng, Andreas Aristidou, Xiaojing Zhang, Lin Chen, Changhe Tu

1 Experiment Details

We examine the impact of framework components in our approach by discarding or replacing the components with possible baselines, as described in Section 5.2 of the main paper. The network details of each experiment are described in this section.

1.1 Effect of the Temporal Alignment Stage

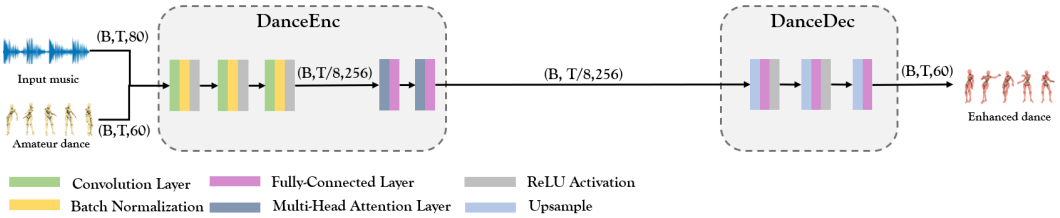

Figure 1: The network architecture of experiment *w/o alignment*.

Without the temporal alignment stage, the network of the experiment *w/o alignment* is an auto-encoder taking the concatenation of music and amateur dance as input and the enhanced dance as output. Figure 1 shows the detail of the network architecture. Note that there's no warping operation in between the encoder and decoder, comparing to the enhancement network of our approach.

1.2 Effect of the Dynamic Time Warping Component

In the experiment *w/o DTW*, the network contains the temporal alignment stage and dance enhancement stage connected as in our approach. The main difference is the absence of the dynamic time warping step, as well as the building and training of the affinity matrix in the temporal alignment stage. Specifically, we build the affinity matrix via the attention mechanism [3] by taking the music features as queries and motion features as keys, and mask out the values far from the diagonal items in the affinity matrix by multiplying with a mask matrix. The affinity matrix is defined as:

$$\begin{aligned} A(i,j) &= \frac{\exp\left(f_G(i)\cdot f_K(j)\right)}{\sum_k^T \exp\left(f_G(i)\cdot f_K(k)\right)} \cdot M(i,j), \\ M(i,j) &= \exp\left(\frac{-(i-j)^2}{\sigma^2+1}\right), \end{aligned} \tag{1}$$

where i and j are the frame index in the music feature sequence f_G and motion feature sequence f_K, σ is set as 50 in our implementation. The alignment network is trained with the loss between the affinity matrix and the ground-truth alignment path matrix. Apart from the above modifications, the dance enhancement stage uses the same network architecture with our approach.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

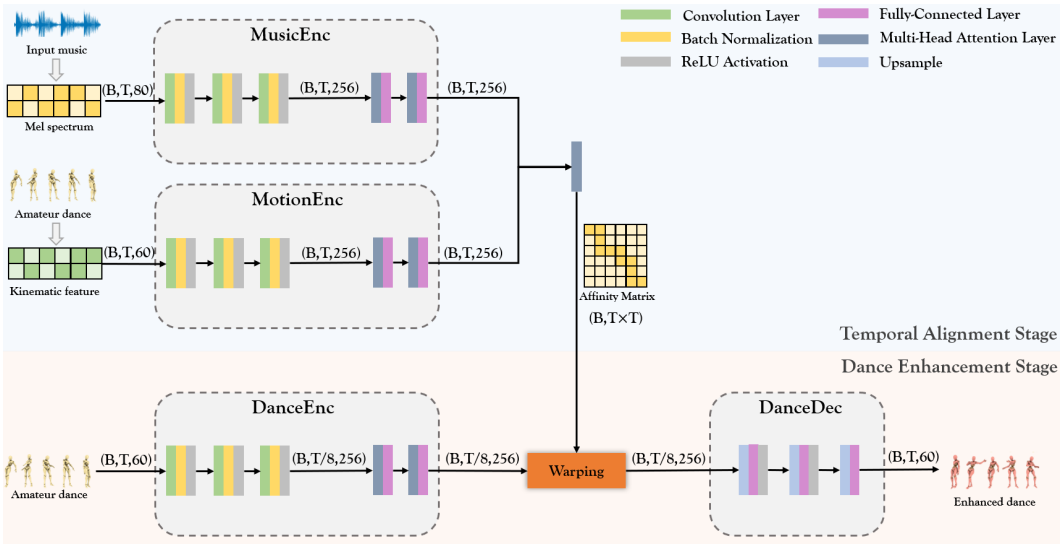

Figure 2: The network architecture of experiment *w/o DTW*.

Please note that although the three encoders, *MusicEnc*, *MotionEnc*, *DanceEnc* share the same architecture, the convolution layers have different stride settings and thus output the features with different dimensions. The convolution layers in *MusicEnc* and *MotionEnc* use stride$=1$ to preserve the length of feature sequences, in order to obtain the frame-to-frame alignment. In *DanceEnc*, we use stride$=2$ to compress the feature temporally and down-sample the alignment path to complete the warping. We experimentally found that the compressed feature sequence helps to generate more smooth and natural enhanced dance movements with less jitter. Plus, the attention layers in the encoders are implemented as multi-head attention layer with head number being 4, while the attention layer to compute the affinity matrix is implemented differently following Eq 1. For the convenience to understand the modification, we present the network of our approach in Figure 3 for a comparison.

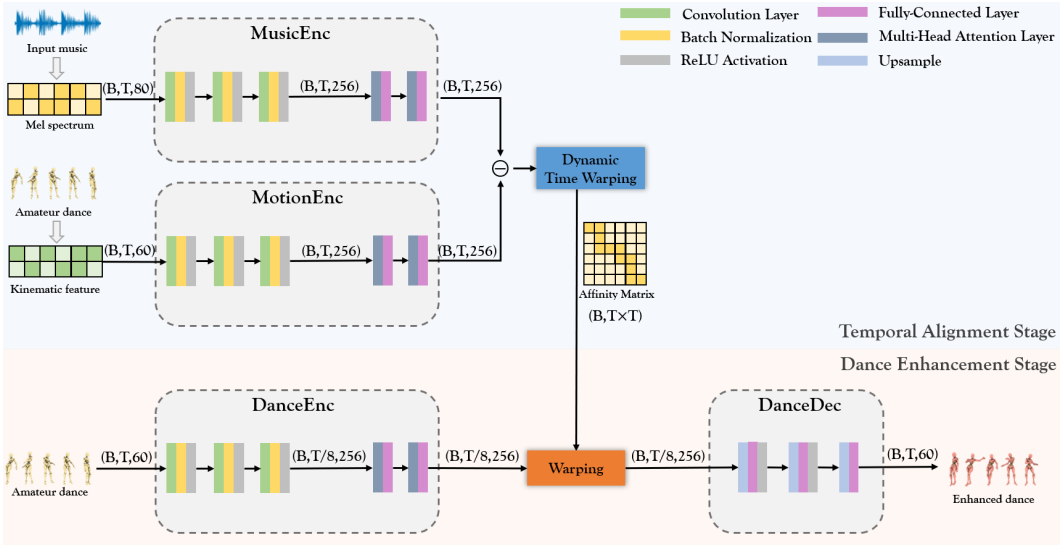

Figure 3: The network architecture of our approach.

1.3 Effect of Dance Enhancement Stage

We evaluate the performance of different network architectures for the dance enhancement stage. In the experiments, i.e. *ConvNet*, *Transformer*, *Conv+Trans*, they share the same pre-trained alignment network in the first stage and differ in the dance enhancement stage. Therefore, we only show their enhancement network in Figure 4.

All the networks follow the encoder-decoder architecture and have a warping operation in the middle to temporally modify the latent feature sequence based on the alignment matrix, which is estimated from the temporal alignment stage. They share the same warping operation and the MLP implementation for the decoder. As for the different encoder implementation, *ConvNet* uses a three conv-relu-bn network structure, *Transformer* contains two blocks each has a multi-head attention layer and a feed-forward linear projection layer, while the *Conv+Trans* is a concatenation of them and is adopted in our approach.

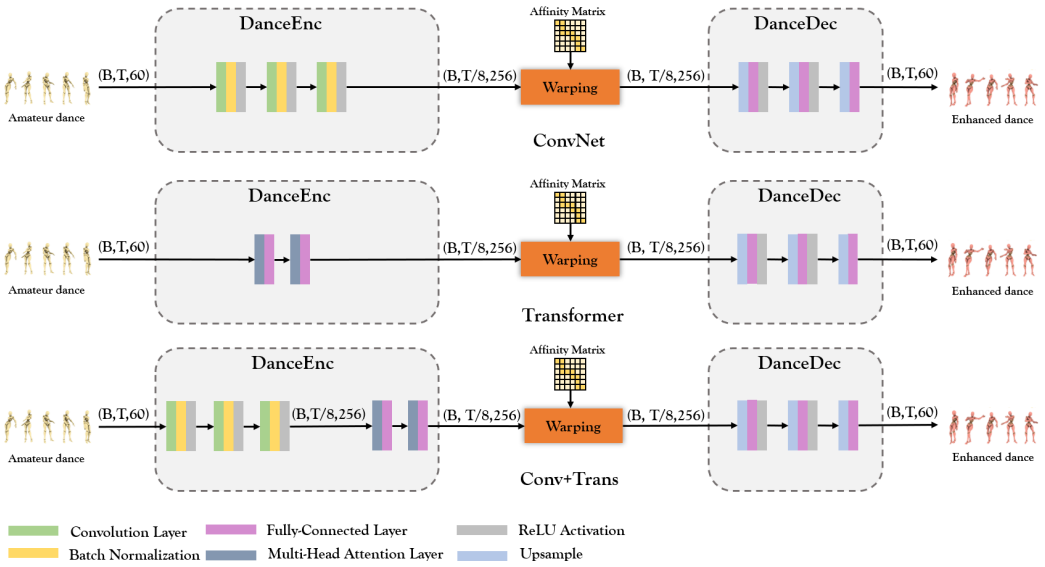

Figure 4: The network architectures to analyze the dance enhancement stage.

2 Perceptual Study: Synthetic Dataset Evaluation

We performed an online perceptual study to examine the quality and realism of our synthetic amateur motions. In this study, each participant was required to answer a 5-likert scale questionnaire about whether the motion the presented motion has been captured from an amateur dancer, or algorithmically generated by a computer. For this task, we recruited, in total, 20 participants, out of which 11 were females and 9 males. Figure 5 shows the task page of the questionnaire, and Table 1 shows the original scores of the participants.

3 Perceptual Study: Professionalism Evaluation

We performed two online perceptual studies to examine the quality and realism of our experimental results in enhancing professionalism on amateur movements, using our synthetically generated amateur dataset and real motion-captured amateur dances. Below we show more details about the experimental results.

Professionalism Evaluation on Synthetic Data. We first conducted a perceptual study to examine the three professionalism aspects of our results, compared to the two baselines [2, 1], the input, and the ground-truth. A total of 20 participants have participated in our study. Each participant first read a brief introduction and then did the 28 tasks. In each task, they were required to watch two dance

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Q1(S)	3	3	1	3	4	3	4	1	2	1
Q2(A)	1	1	3	3	3	3	0	4	3	3
Q3(S)	1	0	2	4	4	1	3	3	2	1
Q4(S)	1	1	3	4	3	0	0	2	2	0
Q5(A)	0	0	3	3	3	3	3	4	0	0
Q6(A)	0	0	1	4	4	1	3	4	3	3
Q7(S)	0	3	3	4	3	3	1	4	3	4
Q8(S)	3	3	3	4	4	4	0	4	3	4
Q9(S)	3	1	2	4	3	3	0	4	4	1
Q10(A)	3	0	3	4	4	3	0	3	2	4
Q11(S)	3	4	2	4	3	2	0	1	1	0
Q12(A)	3	1	3	4	4	3	1	4	2	4
Q13(S)	3	0	3	3	4	1	3	4	3	3
Q14(S)	4	4	2	4	3	1	3	3	2	3
Q15(S)	1	0	2	4	4	1	3	3	3	1
Q16(A)	1	1	2	3	4	4	0	3	1	4
Q17(S)	1	0	2	3	3	0	3	1	0	4
Q18(A)	2	2	1	4	4	4	2	3	1	4
Q19(A)	3	2	1	4	3	3	3	4	2	3
Q20(S)	1	1	2	3	4	2	3	4	2	1
Q21(A)	1	1	2	3	3	2	4	3	1	4
Q22(S)	3	0	2	3	4	1	3	2	2	4
Q23(S)	2	2	2	4	3	2	3	3	3	2
Q24(A)	3	0	2	4	4	3	4	4	2	4
Q25(S)	3	2	2	4	4	3	3	4	1	4
Q26(S)	3	4	2	4	3	3	0	3	2	4
Q27(A)	2	2	2	4	4	3	4	3	1	3
Q28(A)	2	0	2	4	3	3	4	3	3	4

	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Q1(S)	0	4	1	0	1	1	1	3	0	2
Q2(A)	2	1	3	4	2	4	2	3	4	1
Q3(S)	0	1	0	0	1	3	0	3	4	0
Q4(S)	1	2	3	4	2	2	3	3	0	0
Q5(A)	3	1	4	4	0	3	4	3	4	1
Q6(A)	3	0	1	0	1	4	0	0	4	1
Q7(S)	2	0	3	4	1	3	0	3	4	3
Q8(S)	4	4	3	4	1	3	4	3	3	2
Q9(S)	1	4	2	4	1	1	4	3	0	1
Q10(A)	4	4	4	4	1	4	4	3	4	1
Q11(S)	2	1	1	0	2	0	0	3	3	1
Q12(A)	3	3	3	4	2	1	4	3	4	2
Q13(S)	4	3	3	1	3	1	1	1	3	3
Q14(S)	3	4	3	4	3	3	4	3	2	1
Q15(S)	0	4	3	1	2	0	1	3	4	2
Q16(A)	3	4	4	4	3	4	3	3	4	2
Q17(S)	1	0	3	0	2	3	0	3	4	0
Q18(A)	4	4	3	4	1	0	2	3	4	3
Q19(A)	1	1	3	0	2	3	1	3	4	0
Q20(S)	3	3	4	4	2	3	1	3	4	3
Q21(A)	4	1	4	4	2	1	4	3	4	0
Q22(S)	2	2	3	4	2	1	3	3	1	2
Q23(S)	0	2	3	4	2	1	2	3	2	2
Q24(A)	3	2	2	4	0	0	0	3	4	0
Q25(S)	2	4	3	4	1	1	2	3	3	3
Q26(S)	1	4	4	4	2	3	1	3	0	3
Q27(A)	3	4	3	4	3	1	3	2	4	3
Q28(A)	3	3	3	4	1	0	3	3	4	3

Table 1: Raw experimental data from amateur participants for the evaluation on synthetic data. (S) represents synthetic data and (A) represents real amatuer data.

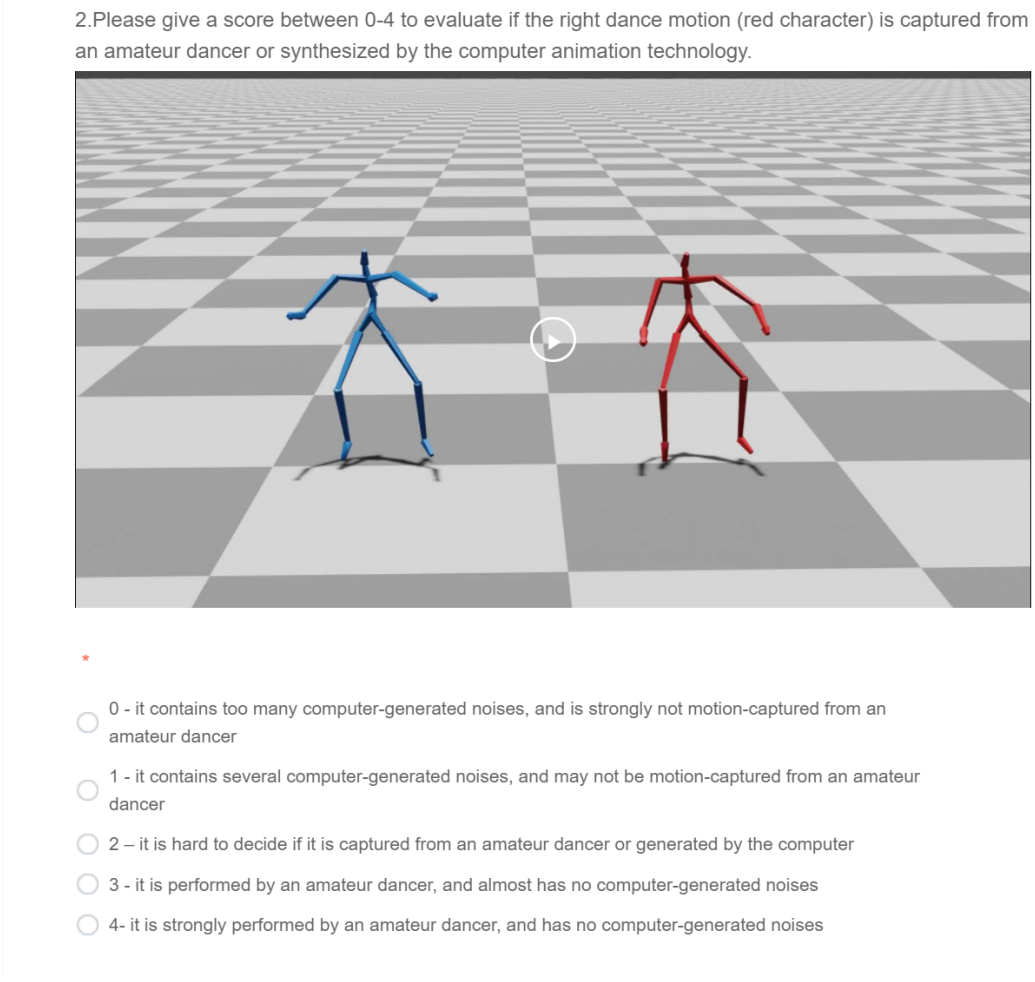

Figure 5: The examplar task page in the perceptual study of synthetic amateur dataset evaluation.

motions and answer three questions regarding the three professionalism aspects (including the motion fluency, the naturalism of physical amplitude, and the music-dance synchronization). Note that one of the dance motions in each task is always generated by our method, and the other is generated from the alternatives. Figure 6 shows the introduction and task page.

Among the 20 participants, 15 of them are amateur dancers, who have studied dance for less than one year. Five participants of them are expert dancers, who have more than eight years of dance study experience. We show the original votes of amateur participants for different methods in Table 2, and that of the expert participants in Table 3. Those raw data were used to calculate the average percentage of participants that voted for our results.

Professionalism Evaluation on Real Motions. We then used the 12 true, motion-captured dance sequences performed by amateur dancers, to further evaluate the performance of our method on real amateur data. In this survey, we recruited 20 participants, out of which 15 were amateur dancers, and five were expert dancers. Similar to the previous study, each participant was shown 12 pairs of dance motions and asked to answer three questions regarding the professionalism aspects. Note that in each pair of dance motions, one is always the true amateur motion and the other is generated by our method. Table 4 and 5 show the original votes of amateur and expert observers, respectively.

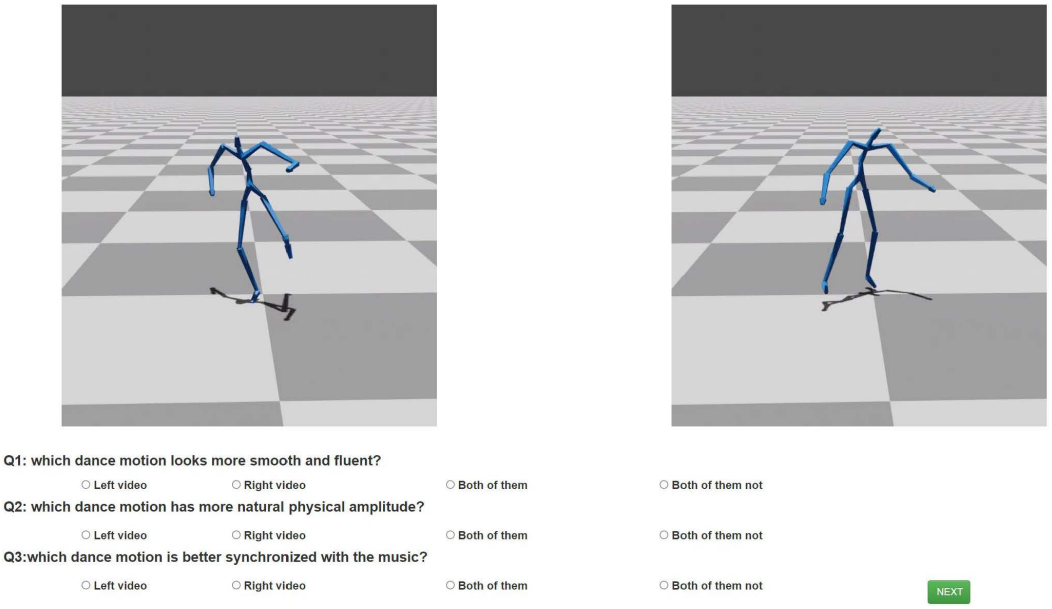

Figure 6: The examplar task page in our perceptual study of professionalism evaluation.

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Q1	Aberman	1	2	2	1	4	0	1	5	3	0	2	4	1	3	2
	Ours	6	5	3	6	7	7	7	6	6	7	6	7	6	6	4
	Holden	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	Ours	7	7	3	7	7	7	7	7	7	7	7	7	7	6	4
	Input	0	3	3	2	2	1	3	4	1	1	0	4	5	5	2
	Ours	7	7	3	6	7	6	6	6	6	6	7	7	2	3	4
	GT	6	4	3	6	7	3	5	7	6	4	6	7	6	7	6
	Ours	4	5	2	5	7	5	5	5	5	6	2	6	4	1	3
Q2	Aberman	2	4	2	1	5	1	2	4	3	0	2	6	1	3	2
	Ours	5	7	3	5	6	7	7	7	6	7	6	7	6	6	5
	Holden	1	0	0	0	0	0	0	0	0	0	0	0	0	4	0
	Ours	7	7	3	7	7	7	7	7	7	7	7	7	7	3	3
	Input	1	3	3	6	1	0	0	3	1	1	0	5	5	7	2
	Ours	7	7	3	1	7	5	7	7	6	6	7	7	2	3	4
	GT	6	4	3	7	6	4	5	7	5	4	6	7	6	7	6
	Ours	4	5	2	3	6	5	5	6	5	6	2	7	4	2	3
Q3	Aberman	1	4	2	1	3	0	0	3	2	0	1	2	2	1	3
	Ours	6	4	3	5	7	5	6	7	5	7	6	5	7	6	6
	Holden	0	0	0	0	0	0	0	1	0	0	0	0	0	2	2
	Ours	7	7	3	7	7	6	6	6	7	7	7	5	6	5	4
	Input	0	2	3	5	2	1	2	5	3	1	0	1	5	3	2
	Ours	7	7	3	3	7	5	4	5	5	6	6	4	5	4	4
	GT	5	6	3	7	7	4	4	7	6	4	6	7	5	5	5
	Ours	5	6	2	4	6	3	4	6	4	6	2	3	4	3	4

Table 2: Raw experimental data from amateur participants for the professionalism evaluation on synthetic amateur data.

		P1	P2	P3	P4	P5
Q1	Aberman	0	3	3	3	3
	Ours	7	6	3	3	5
	Holden	0	0	0	0	0
	Ours	7	7	7	7	7
	Input	1	4	4	4	2
	Ours	6	6	4	3	6
	GT	6	7	6	6	5
	Ours	5	1	2	1	4
Q2	Aberman	0	5	3	5	3
	Ours	5	7	3	2	4
	Holden	0	0	0	0	0
	Ours	7	7	7	7	7
	Input	1	4	4	3	1
	Ours	6	7	4	4	6
	GT	6	6	6	4	6
	Ours	1	3	2	3	4
Q3	Aberman	0	1	3	4	0
	Ours	7	3	4	2	6
	Holden	0	1	0	0	1
	Ours	7	6	7	7	6
	Input	1	3	6	5	1
	Ours	5	4	4	2	6
	GT	6	7	6	5	5
	Ours	5	2	5	2	5

Table 3: Raw experimental data from expert participants for the professionalism evaluation on synthetic amateur data.

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Q1	Input	7	8	9	8	8	5	6	5	6	7	7	6	10	5	7
	Ours	10	9	9	7	9	7	6	9	11	9	9	12	11	9	6
Q2	Input	8	9	8	7	7	6	5	6	7	8	8	10	12	4	7
	Ours	7	8	7	5	7	6	7	7	10	5	10	11	12	9	5
Q3	Input	8	5	7	7	6	5	7	8	6	6	7	7	6	6	8
	Ours	10	8	6	8	8	7	7	8	9	8	9	11	9	9	9

Table 4: Raw experimental data from amateur participants for the professionalism evaluation on true amateur data.

		P1	P2	P3	P4	P5
Q1	Input	9	5	6	4	8
	Ours	7	7	8	8	9
Q2	Input	7	6	7	4	8
	Ours	6	6	8	6	9
Q3	Input	5	6	7	5	9
	Ours	8	5	9	7	7

Table 5: Raw experimental data from expert participants for the professionalism evaluation on true amateur data.

4 Qualitative Comparison on Real Captured Data

In our paper, we show qualitative comparison between our method and the alternatives on the synthetic dataset. Additionally, we captured twelve dance motion sequences performed by amateur dancers, and compared with the alternatives on those data. Figure 7 shows the qualitative comparison on one of the real captured data. It can be observed that our method produces exaggerated dance poses compared to the two baseline methods, in order to enhance the professionlism. Please refer to our supplementary video for animated results.

Figure 7: Qualitative comparison on real captured data.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Trans. Graph.*, 39(4), jul 2020.

[2] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), jul 2016.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.