The Supplemental Material for "RSMT: Real-time Stylized Motion Transition for Characters"

ACM Reference Format:
. 2023. The Supplemental Material for "RSMT: Real-time Stylized Motion Transition for Characters". In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 6–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3588432.3591514

1 METHODS

1.1 Network details of manifold

The encoder consists of a three-layer network with 256 units, which is followed by ELU activation in hidden layers. The output is divided into two 32-dimensional variables, μ and σ, in order to achieve the latent variable z through reparameterization. In the hidden layers, the gating network has three linear layers with 128 units and dropouts with a probability of 0.3. The blending weights for all eight experts are calculated by the gating network using the next phase p^{i+1} and latent variable z. The eight parameters will then be used to weight the eight experts in the decoder. Each expert has three linear layers, each with 512 units, followed by ELU activation.

1.2 Network details of sampler

The style encoder is made up of four convolution layers, each with 512 dimensions, followed by ReLU activation. The kernel sizes of the four convolution layers are 1, 3, 3, 5, respectively. In Style Embedding, $k \in \mathbb{R}^{512\times T}$ is embedded by a FiLM and an Attention block, each followed by an ELU activation. The FiLM block scales and shifts the latent variable via the style feature:

$$\mathbf{FiLM}(x) = \mathbf{LN}(x) \cdot \Phi_\sigma(Pooling(k)) + \Phi_\mu(Pooling(k)), \quad (1)$$

where **LN** represents the layer normalisation and x is the latent variable. *Pooling* is a pooling layer that removes the temporal axis from k. Φ_σ and Φ_μ are linear layers to produce the FiLM vectors. For the Attention block, we compute the weighted-sum of k along the temporal dimension based on the similarity between the style feature at each frame and the predicted latent variable (a.k.a the attention mechanism):

$$\mathbf{ATN}(x) = x + \Phi_s(k) \cdot \mathbf{softmax}(\frac{\Phi(x)^T \cdot \Phi_o(k)}{\sqrt{512}})^T, \quad (2)$$

where $x \in \mathbb{R}^{512}$ is the latent variable, $\Phi_s, \Phi, and \Phi_o$ are three linear layers. The State Encoder, Target Encoder, and Offset Encoder all use the same architecture, which includes a 512-unit hidden layer and a 256-unit output layer. PLU is used as the activation in all layers. The style embedding network uses a linear layer to transform the input to 512 dimensions before embedding the style using FiLM and Attention blocks.

The hidden layer of the LSTM contains 1024 units. The decoder has four linear layers, three of which are hidden. The hidden layer has 1024, 512, and 512 units. All layers are followed by ELU activation. The style embedding occurs following the second hidden layer.

2 ABLATION STUDY

2.1 Attention block

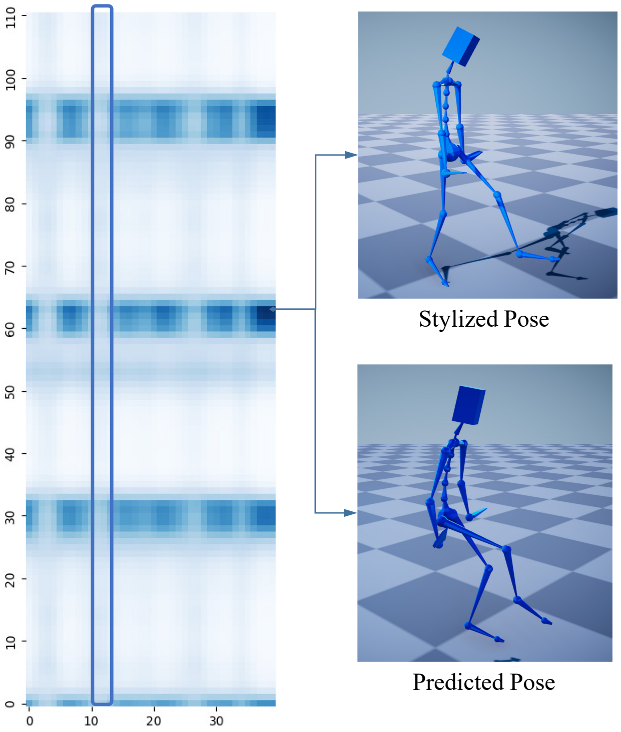

Figure 1: Visualization of an attention map of kick style.

To clearly evaluate the ATN mechanism, we visualize the attention vector of each frame and concatenate them along the temporal axis to generate an attention map. Each column is an attention vector, and each row is a predicted frame along the temporal axis. As shown in Figure 1, most attention maps exhibit regular patterns, with the highlight areas being the phases performing the most stylized pose. The figure illustrates the kick style's in-between sequences. The highlight area for each column is the phase in which the leg reaches its highest position. Besides, the periodic brightness change of each row corresponds to the predicted motion's phase

SIGGRAPH '23 Conference Proceedings, August 6–10, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0159-7/23/08…$15.00
https://doi.org/10.1145/3588432.3591514

Table 1: Manifold and Sampler are trained on two different datasets. We use "BA on C" to represent that the manifold is trained on B dataset, the sampler is trained on A dataset and the experiment is tested on C dataset.

	L2 norm of global position		
Frames	10	20	40
BA on A	0.64 (+0.05)	0.80 (+0.04)	1.38 (+0.07)
AA on A	**0.59**	**0.76**	**1.31**
AB on A	0.80 (+0.21)	1.19 (+0.43)	1.92 (+0.61)
BB on A	0.78 (+0.19)	1.15 (+0.39)	1.85 (+0.54)
BA on B	0.62 (+0.09)	0.89 (+0.21)	1.53 (+0.42)
AA on B	0.63 (+0.10)	0.94 (+0.26)	1.63 (+0.52)
AB on B	0.58 (+0.05)	0.74 (+0.06)	1.23 (+0.12)
BB on B	**0.53**	**0.68**	**1.11**
BA on C	0.82 (+0.02)	**1.180**	1.97 (+0.02)
AA on C	**0.80**	1.21 (+0.03)	**1.95**
AB on C	0.98 (+0.18)	1.51 (+0.33)	2.44 (+0.49)
BB on C	0.97 (+0.17)	1.56 (+0.38)	2.53 (+0.58)
		NPSS	
Interpolation on A	0.0062	0.0327	0.2716
BA on A	0.00502	0.01828	**0.09576**
AA on A	**0.00435**	**0.01640**	0.09850
AB on A	0.00534	0.02141	0.11461
BB on A	0.00518	0.02195	0.13408
BA on B	0.00612	0.02395	0.18043
AA on B	0.00549	0.02838	0.20981
AB on B	0.00504	0.01922	0.12387
BB on B	**0.00380**	**0.01442**	**0.10067**
BA on C	0.00743	0.04057	0.18372
AA on C	**0.00634**	**0.03798**	**0.16784**
AB on C	0.01032	0.05541	0.19864
BB on C	0.00872	0.05046	0.26806
		Foot skate	
Ground Truth on A	0.161	0.167	0.167
BA on A	0.202	0.222	0.321
AA on A	**0.171**	**0.197**	**0.297**
AB on A	0.199	0.230	0.340
BB on A	0.217	0.250	0.353
Ground Truth on B	0.184	0.181	0.186
BA on B	0.256	0.232	0.346
AA on B	**0.211**	**0.218**	0.324
AB on B	0.221	0.249	0.343
BB on B	0.213	0.231	**0.309**
Ground Truth on C	0.271	0.269	0.255
BA on C	0.210	**0.241**	**0.302**
AA on C	**0.196**	0.264	0.336
AB on C	0.218	0.279	0.358
BB on C	0.235	0.276	0.345
		Diversity	
BA on A	0.908	2.210	7.423
AA on A	0.869	2.172	7.194
AB on A	**0.938**	**2.303**	**7.587**
BB on A	0.899	2.283	7.446
BA on B	**0.892**	**2.191**	**7.612**
AA on B	0.856	2.189	7.471
AB on B	0.875	2.096	6.989
BB on B	0.834	2.049	6.738
BA on C	1.013	2.629	**8.170**
AA on C	0.965	2.539	7.830
AB on C	**1.024**	**2.663**	8.062
BB on C	0.995	2.623	7.844

Table 2: Comparison of different methods on reconstruction, foot skating and diversity metrics.

	L2 norm of global position		
Frames	10	20	40
w/o phase	0.532	**0.658**	1.155
w/o ATN	0.562	0.730	1.230
Our method	**0.525**	0.680	**1.148**
		NPSS	
w/o phase	**0.00361**	**0.01434**	**0.08756**
w/o ATN	0.00398	0.01563	0.099645
Our method	0.00384	0.01507	0.09659
		Foot skate	
w/o phase	0.217	0.215	0.290
w/o ATN	0.186	0.221	0.306
Our method	**0.174**	**0.194**	**0.272**
		Diversity	
w/o phase	0.630	1.564	5.364
w/o ATN	**1.116**	**2.148**	**7.257**
Our method	0.910	2.017	6.683

Table 3: Comparison on the L2 norm of global position of last predicted frame and the target, foot skating metrics, under the conditions that change the time duration and locations of target frame of 40 missing frames.

	L2 norm of global position of the last frame			
Conditions	dt=2, d=1	dt=0.5, d=1	dt=1, d=2	dt=1, d=-1
w/o phase	0.573	0.503	0.480	0.494
w/o ATN	0.330	0.430	0.338	0.351
Our method	**0.292**	**0.358**	**0.302**	**0.303**
		Foot skate		
w/o phase	**0.238**	**0.794**	0.689	0.563
w/o ATN	0.237	1.081	0.696	0.589
Our method	0.307	1.018	**0.592**	**0.557**

change, with the highlighted areas following the same phase as the most stylized pose. Moreover, the attention vector shows the averaged gray color (the blue box area) for the predicted phase, which is not the most stylized.

As a result, when predicting the motion of the specific phases, the ATN block emphasizes the stylized pose. As a result, the model receives explicit guidance that it performs better reconstruction accuracy with a little less diversity, as shown in Table 4.

2.2 Manifold Ablation

To make a fair comparison, we replace the CVAE sampler with our sampler, which accepts the style code as input. The manifold design distinguishes the two methods. CVAE does not employ a phase manifold (see w/o phase in Tabs 2 and 3). Adding the style code to the sampler has little effect on the reconstruction, NPSS, foot skating, and diversity metrics. When it comes to foot skating and motion diversity, our method still outperforms the CVAE.

Furthermore, even after explicitly imposing style, CVAE cannot generate vividly styled motions if the motion differs from the distribution of the training set. When the character slows down, for example, the phase of our method drives the character to maintain stylized motion while the CVAE performs slow and less-stylized movement. Please watch the accompanying video for the animations. More information about phase can be found in the following section.

2.2.1 Phase discussion. A phase is a spatial-temporal structure that captures the characteristics of a short clip of stylized motion. We conclude from the comparison of our method and CVAE that the benefit of style preservation is due in part to the phase.

We found that the less-stylized clips of the generated motion have different hip velocities than the stylized clips when we examined CVAE failure cases. In these cases, the character would adopt a less stylized pose in order to meet the velocity requirement and avoid foot skating. In contrast, our method learns a stylized strategy to meet the phase and hip velocity requirements at the same time.

To validate this, we randomly sample a Gaussian noise vector to replace the latent variable sampled by the sampler. CVAE would generate a random pose based on the previous pose and the predicted hip velocity, whereas our method requires an additional input: the predicted phase vector. CVAE motion cannot maintain the same dynamic as stylized motion because CVAE is more likely to sample a common pose (neutral walking or standing still) than a stylized one from a Gaussian distribution with no prior knowledge. The predicted phase of our method, on the other hand, always drives the character to continue performing the same dynamic as the stylized motion.