Style-ERD: Responsive and Coherent Online Motion Style Transfer Supplementary Material

A. Implementation Details

Our method is implemented in PyTorch and trained on a PC with an 8-core AMD Ryzen 7 CPU with a single NVIDIA GeForce GTX 1060 GPU. The overall training process takes about 10 hours to complete. Both the generator and discriminator are optimized by Adam optimizer [2] with learning rates of 10^{-4} and 5×10^{-5} respectively. Our model is trained for 2000 epochs with a batch size of 16. We adopt the default initialization offered by PyTorch for all the neural networks in our model.

The encoder in the style transfer module consists of two MLP layers with $(64, 32)$ neurons and ReLU activation. The latent code z has a dimension of 32. The residual module comprises 6 LSTM layers for the neutral branch r_0 and 4 LSTM layers for other branches $[r_1, \ldots, r_{n_S}]$. The decoder is conditioned on target style labels and comprises 4 MLP layers with $(58, 86, 128, 184)$ neurons and ReLU activation. For the discriminator, we find that replacing ReLU with LeakyReLU accelerates the training.

We experiment with two different loss functions on the quaternion values: L2-norm and $\mathcal{L}_{quat}$ in Eq. (2). Though the four coefficients of a quaternion are continuous and smooth, the rotation value is not evenly-spaced. Optimizing in this not evenly-spaced space with L2-norm can lead to jerky motion, which can be solved by adopting $\mathcal{L}_{quat}$.

We pre-train a denoising autoencoder with 1D convolution layers as the feature extractor for Frechet Motion Distance (FMD). At training time, the joint rotations are sampled from the dataset, and we add random noise sampled from a normal distribution, $\mathcal{N}(0, 0.03)$ to the joint rotations. The denoising autoencoder is trained to reconstruct the noisy joint rotation input with $\mathcal{L}_{quat}$ in Eq. (2) as the loss function. After the training, we use the activation of the last convolution layer in the encoder as features for the FMD score.

Code, models and demo videos are available at https://tianxintao.github.io/Online-Motion-Style-Transfer.

B. User Study

We performed a user study to compare the quality of the style transfer results. We presented the participants with the rendered motions in a side view angle.

Method	Style	Content	Quality
Aberman *et al.* [1]	9.5%	4.5%	6.1%
Park *et al.* [4]	12.3%	8.6%	5.8%
Ours	**78.2%**	**86.8%**	**88.1%**

Table 1. User study results. Each entry in the table corresponds to the ratio of being selected as the optimal among the alternatives in the three aspects.

The questionnaire contained eight sets of transfer results in the target style and also the input motion for reference. Each set contained the motions generated by *Style-ERD* and alternative methods [1, 4]. The users were supposed to select the best results in three aspects: style expressiveness, content preservation and overall quality.

We collected a total of 243 responses from 31 participants. As shown in Tab. 1, 78.2%, 86.8% and 88.1% of the responses select our results as the best among all the choices in the three aspects respectively. The results of the user study suggest that *Style-ERD* outperforms the alternative methods in terms of the style transfer quality.

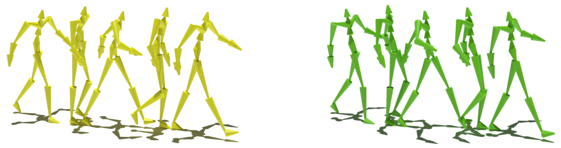
C. Discussions

C.1. Limitations

Our method still requires paired motion data in different styles. This could possibly be addressed by adopting the cycle consistency idea [5]. Also, our model is conditioned on content labels, which could be unavailable. Additionally, *Style-ERD* has an increasing runtime with a growing input length. As with other approaches, we still rely on inverse kinematics based cleanup to remove minor foot sliding. As future work, we are also interested in optimizing our framework to quickly adapt to unseen styles in a few-shot by matrix decomposition as shown in [3].

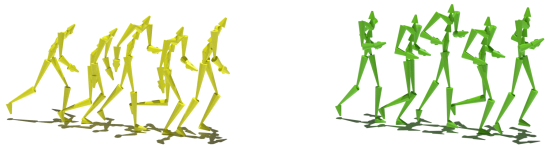
C.2. Societal Impact

The motion data is biased towards one default skeleton, ignoring the needs of minority groups, *e.g.*, children, elders and people with physical disabilities. This can be mitigated



(a) Input: Angry walk (b) Output: Strutting walk

Figure 1. Style transfer with non-neutral input motion. Our model is not restricted by the input motion style.



(a) Input: Depressed run (b) Output: Neutral run

Figure 2. Transfer non-neutral input to neutral style.

by motion retargetting in the pre-processing step. Motion style transfer could be used for creating human animations with the intention to mislead.

C.3. Style Transfer with Unpaired Data

We design our framework to perform motion style transfer with paired motion data. For example, to transfer a neutral walking motion to angry, the dataset should contain an angry walking motion as reference, which does not need to be temporally registered. However, some researchers have demonstrated style transfer with unpaired motion data is also feasible [1]. Such style transfer with unpaired data is defined as heterogeneous style transfer. Exploring style transfer with unpaired data remain as a future work since it lowers the requirement on the dataset and enables to learn more diverse style transfer effects.

C.4. Choice of Input Style

Most qualitative results in this work use input motion in neutral style. Style transfer starting from neutral motion is a more practical problem because more neutral motion exists in the available database. Nevertheless, our framework is not restricted by the input style to be neutral. Fig. 1 shows our method can successfully transfer input motion in angry style to motion in strutting style. Given the training process, transferring non-neutral motion to another style other than neutral is an unseen task at inference time. Moreover, we also experiment with transferring stylized motion to its neutral counterpart, and show the results in Fig. 2. Such results reveal that the encoder E in the style transfer module *Style-ERD* can normalize input motion to neutral style as expected.

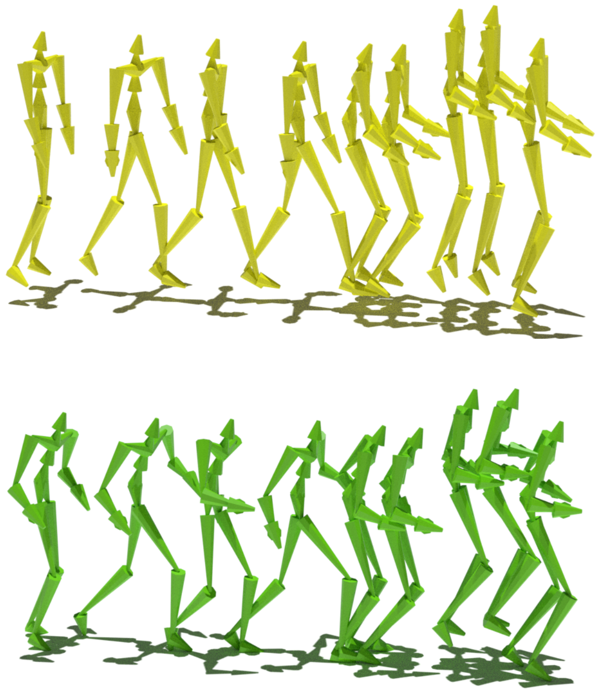
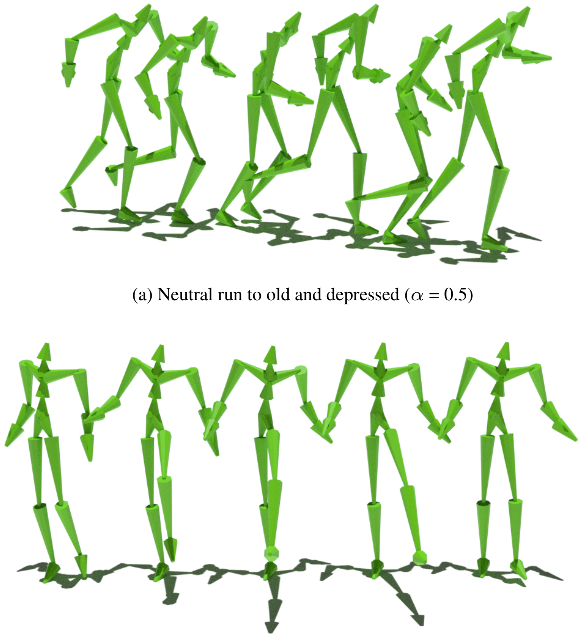


Figure 3. Style transfer on heterogeneous motion sequence. Task: Neutral walk and jump motion to old.

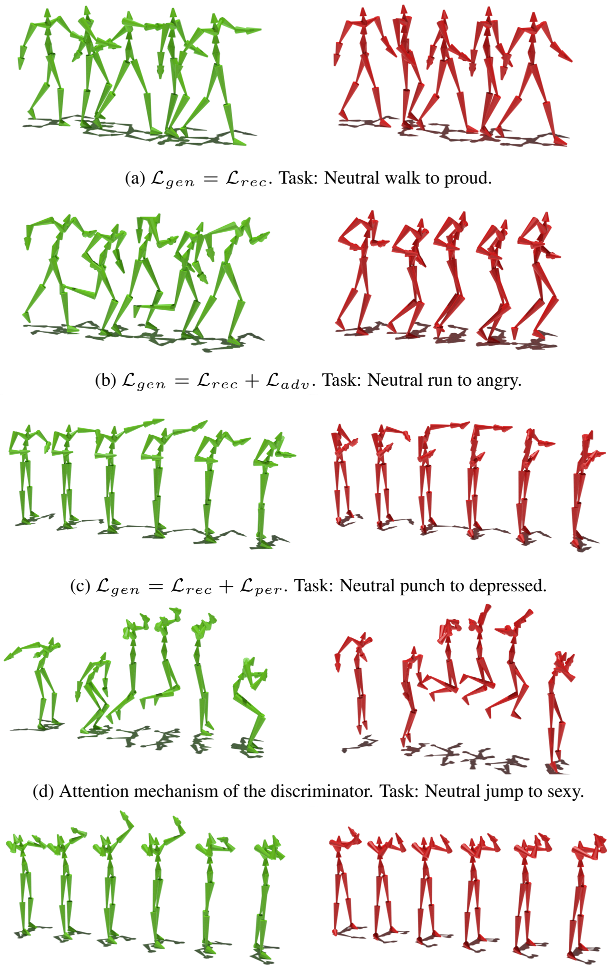


(a) Neutral run to old and depressed ($\alpha = 0.5$)

(b) Neutral kick to childlike and strutting ($\alpha = 0.5$)

Figure 4. Mixture of two existing styles.

When the input motion is not clearly depicted with a style label, the motion can usually be considered neutral for the transfer algorithm and produces satisfying style transfer results. For example, the *Mixamo* data we used in the gen-



(a) $\mathcal{L}_{gen} = \mathcal{L}_{rec}$. Task: Neutral walk to proud.

(b) $\mathcal{L}_{gen} = \mathcal{L}_{rec} + \mathcal{L}_{adv}$. Task: Neutral run to angry.

(c) $\mathcal{L}_{gen} = \mathcal{L}_{rec} + \mathcal{L}_{per}$. Task: Neutral punch to depressed.

(d) Attention mechanism of the discriminator. Task: Neutral jump to sexy.

(e) Learnable initial hidden states. Task: Neutral punch to childlike.

Figure 5. Ablation studies. Left (green): results produced by the full model, Right (red): ablation experiment results.

eralization test does not own a specific style label. We treat it as neutral motion and feed it to the *Style-ERD* module to generate the motions shown in Fig. 7 of the paper. Alternatively, a style-classification model can be trained to predict the style of the input motion. Therefore, *Style-ERD* can accept the style prediction of the classification model as the input style label for the transfer task.

D. Extra Visualization Results

In this section, we show the results that are omitted in the paper, including style transfer results on heterogeneous motion sequence in Fig. 3, mixture of two styles in Fig. 4, and ablation study results in Fig. 5.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64–1, 2020. 1, 2

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura. Few-shot learning of homogeneous human locomotion styles. In *Computer Graphics Forum*, volume 37, pages 143–153. Wiley Online Library, 2018. 1

[4] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–17, 2021. 1

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1