Autoregressive Stylized Motion Synthesis with Generative Flow Supplementary Materials

Yu-Hui Wen[1]†, Zhipeng Yang[2]†, Hongbo Fu[3], Lin Gao[2,4]*, Yanan Sun[1], Yong-Jin Liu[1]*
[1]CS Dept, BNRist, Tsinghua University
[2]University of Chinese Academy of Sciences
[3]School of Creative Media, City University of Hong Kong
[4]Beijing Key Laboratory of Mobile Computing and Pervasive Device, ICT, CAS
{wenyh1616,liuyongjin}@tsinghua.edu.cn, {yangzhipeng19s,gaolin}@ict.ac.cn,
hongbofu@cityu.edu.hk, sunyn20@mails.tsinghua.edu.cn

In this supplementary document, we provide further details of the implementation used in our experiments, as well as more experimental results.

1. Experiment and Evaluation

1.1. Implementation Details

As illustrated in our paper (Section 3.2), our generative model is trained to learn the motion distribution at frame t with τ previous poses and control signals in the current and τ previous frames. In this way, the generative model based on flows is able to synthesize motions autoregressively. Specifically, we set $\tau = 10$.

Each step of flow contains a coupling layer. We first split the input a of the coupling layer into two parts $a = [a', a'']$ in the channel dimension. In more detail, the input a contains the data of τ previous poses and control signals in the current and τ previous frames. Then, we transform one part of the input a'' based on the scale and translation parameters, which are extracted from the remaining part a' by using the invertible Transformer. The invertible Transformer in each coupling layer consists of two layers, followed by a linear transformation (Section 4.1 in our paper). During training, the linear transformation is initialized by zeros. Thus, the scale parameters are initialized close to ones and translation parameters are initialized with zeros [2, 3]. Consequently, the effect of the coupling layer is initially close to an identity transformation to facilitate training our deep networks [5].

During motion style transfer, we generally infer the latent style from the input style motion. In addition, we edit the style code in the latent space to generate multiple plausible results. In more detail, we infer the latent codes z_{s_i} from motions in the style s_i and calculate the difference of $\max(z_{s_i}) - \min(z_{s_i})$ as a direction of the latent space to manipulate the motion style s_i. Note that our method is unsupervised during training and uses the style labels only during testing.

1.2. Latent Code Visualization

As discussed in our paper (Section 3.2), we use the latent codes inferred from the generative flow model to control motion styles. To get a better understanding of how the generative flows learn to synthesize stylized human motions, we infer the latent codes by our proposed generative flow model and project the latent codes onto a 2D space by using t-SNE (Section 4.2). Below, we show more experiment results of latent style codes.

Latent Style Code. In our paper (Section 4.2), we evaluate the clustering results of latent codes from the generative flows with different invertible transformation settings (i.e., baseline: without imposing an invertible transformation, with an invertible LSTM, and with an invertible Transformer). The clustering results show that the invertible Transformer outperforms the other settings (Section 4.2). We also evaluate the motion style transfer results qualitatively to confirm the superiority of the invertible Transformer. As shown in Figure 1, our model with the invertible Transformer can transfer the "sexy" style of a walking motion to a running motion, while the other settings fail to transfer the style and preserve the content.

In our paper (Section 4.2), the latent codes inferred from walking samples of dataset A are plotted to show how the generative flows learn to synthesize stylized motions in the same content (walking). Here, we show the projected 2D latent codes from random motion samples of different motion contents (i.e., jump, kick, punch, run, trans, walk) in dataset A. As shown in Figure 2 (b), the latent style codes inferred from our model can be clustered in accordance with motion contents. We notice that the latent codes of "trans"

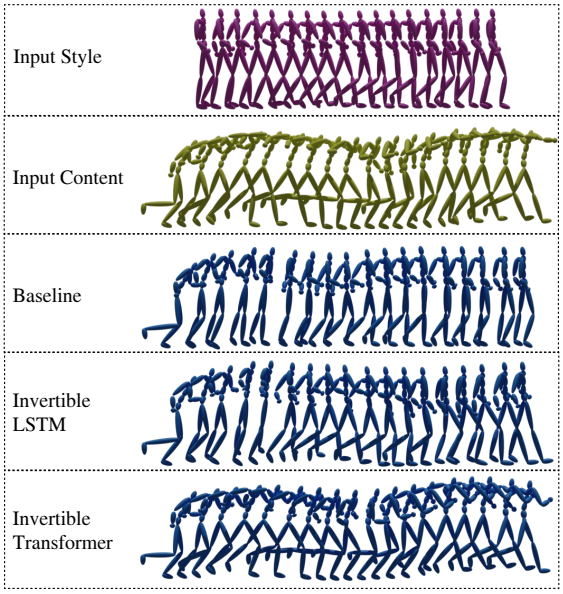


Figure 1. Qualitative comparison of our model with different settings, namely, without imposing an invertible transformation (Baseline), with an invertible LSTM, and with an invertible Transformer. The style of a walking motion (in a "sexy" style) is transferred to a running motion (in an "angry" style). The full video sequences can be found in the supplemental video.

motions are not clustered into a group, because the motions in the "trans" content represent transition movements (e.g., from "walk" to "run"), which are similar to other motions in contents. The result implies that our model can synthesize stylized motions, considering not only the style properties but also the content properties of motions. In Figure 2 (a), it can be seen that the style codes extracted by the network of Aberman et al. [1] are not related to the motion contents. As illustrated in [1], their network encodes motion styles and contents into latent codes, separately. However, it is hard to perform motion style transfer in a similar way for different motion contents. For example, the "childlike" style may have an influence on the "walk" motion with lively movements of upper limbs, and may have an influence on the "kick" motion with slow movements of lower limbs. As shown in Figure 3, the model of Aberman et al. [1] fails to transfer the "childlike" style of a running motion to a kicking motion and preserve the kicking motion content. However, our model is able to synthesize a new kicking motion in the "childlike" style successfully.

The motion samples of a specific style in dataset A are captured in different contents, separately. For example, childlike walking movements and childlike kicking movements are captured in different motion samples. Thus, the style latent codes from motion samples in the same content (walking) in dataset A can be clustered according to

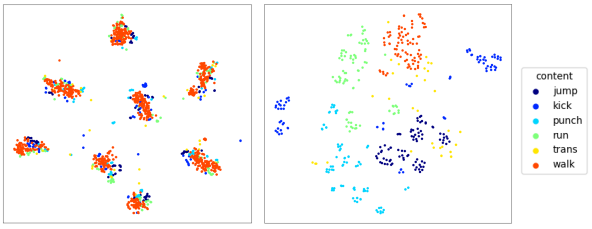


Figure 2. Style latent codes extracted by the network of Aberman et al. [1] (a) and our generative flow model (b). The latent codes are extracted from motion samples in dataset A and are projected onto 2D space by using t-SNE, and colored according to their motion contents.

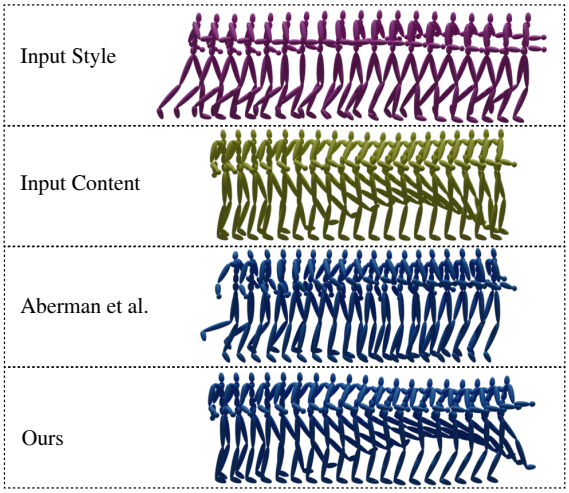


Figure 3. Qualitative comparison between our model and that of Aberman et al. [1]. Our model is better at transferring a "childlike" style of a running motion) to a kicking motion while preserving the motion content. The full video sequences can be found in the supplemental video.

style labels (as illustrated in our paper, Section 4.2). However, the motion samples of a specific style in dataset B are performed by a character in different contents without partitions between the contents. Thus, there are many transition motions (e.g., from "walk" to "turn" and from "walk" to "kick") in dataset B. Here, we also show the projected 2D latent codes, which are inferred from random motion samples in dataset B (Figure 4). It can be seen from Figure 4 that the latent codes learned by the generative flow model are not clustered according to style labels. This may be due to the fact that most of the motion samples in dataset B are transition motions. The latent codes learned from a mixture of transition motions tend to manipulate the style properties, relying on motion contents. However, our model can still successfully learn to synthesize stylized motions based on dataset B, as shown in Figure 5 and our supplementary

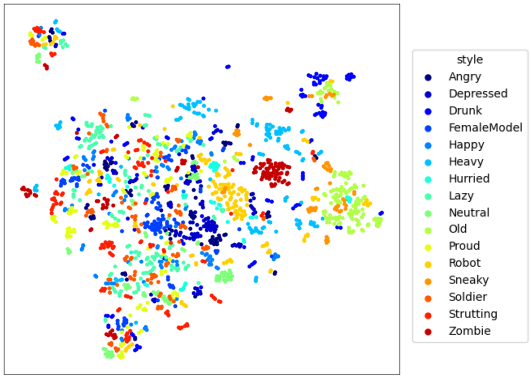


Figure 4. Style latent codes extracted from motion samples in dataset B are projected onto 2D space by using t-SNE, and colored according to their style labels.

video.

Unseen Styles. As shown in our paper (Section 4.2), our generative flows can not only learn to cluster the action samples in the same style, but also generalize to the samples in an unseen style. Here, we perform experiments to verify the generalization ability of the generative flows in a more challenging scenario. We retrain our model on dataset A without the motions that are labeled by the "strutting" label or "old" label, and then test the retrained model using the motions including the "strutting" and "old" styles. The latent codes of unseen "strutting" and "old" styles extracted by the network of Aberman et al. [1], are close to the "proud" and "angry" styles (Figure 6), respectively. In contrast, the latent codes of unseen styles from our model are clustered successfully. The experiment result demonstrates that the generative flows can infer latent style properties from unseen style motion samples with very limited training data, implying that the proposed generative flow model has a good generalization ability for applications.

1.3. Plausible Results.

As illustrated in our paper (Section 3), our model is probabilistic and is able to generate plausible motions in a specific style. In more detail, we edit the style latent code inferred from a style motion in the latent space to generate multiple plausible results, as shown in Figure 7.

1.4. User Study.

As illustrated in our paper (Section 4.3), we have compared our model to other methods with qualitative experiments. We also conduct a user study to qualitatively evaluate our synthesized motion results in terms of the realism, style expressiveness, and content preservation as suggested by [1]. 50 subjects have been invited to participate in the user study, and most of them have no experience (about 52%) or are beginners (about 32%) in the research on human motion. For evaluating our model, the participants are shown to the human motions represented by using 3D stick skeletons in a fixed camera angle.

	Realism	Style	Content
Mocap	53.6%	–	–
Holden et al. [4]	36.6%	20.4%	24.2%
Aberman et al. [1]	39.8%	32.4%	40.2%
Ours	47.8%	47.2%	35.6%

Table 1. User study results.

First, we evaluate the realism of different motions. We prepare in total 10 sets of motions. Each set of motions contains 4 motions, obtained from four different sources: (1) Results from Mocap datasets, (2) Results of Holden et al. [4], (3) Results of Aberman et al. [1], (4) Results of our method. Specifically, the results of (2), (3) and (4) are generated with the same inputs. The users are asked to answer a question: "Which of the motions is realistic?", and to choose one of the five answers: (1), (2), (3), (4) or None.

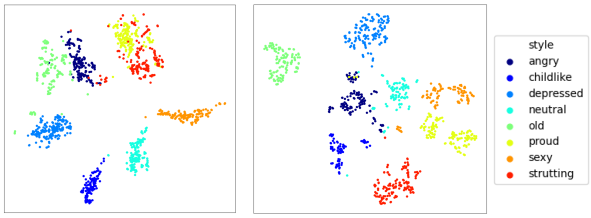
We receive 500 (= 50 participants $\times$ 10 sets of motions per participant) responses for the question of realism evaluation, and report the realism ratios for each motion source in Table 1. It is shown that only 53.6% of the motions from Mocap datasets are judged as realistic. This is possibly because most of the users are not familiar with human motion studies. Specifically, the users tend to judge the motions in special styles (e.g., running in the "old" style) as not realistic. However, the results still confirm that the motions generated by our model are the most realistic compared to those by the other two methods.

Second, we compare our style transfer results to those of Holden et al. [4] and Aberman et al. [1] in terms of style transfer and content preservation. Similarly, we prepare 10 sets of motions, each including a style input, a content input, and three transferred results by the three compared methods. Among each set of 3 evaluated transferred results, we ask the users to first select the motion whose style is closer to the input style motion (" Which of the motions is more similar to the input style motion in style?"), and then select the motion whose content is closer to the input content motion (" Which of the motions is more similar to the input content motion in content?"). 7 sets of the above motions involve input motions in different styles but in the same content, because it is difficult for common users who are not familiar with human motion studies to judge style transfer results from input motions in different contents (e.g., transfer the style of a kicking motion in the "childlike" style to a running motion in the "old" style).

We receive 500 responses respectively for evaluating the style transfer and content preservation. The results are shown in Table 1. It is shown that our method is judged

(a) (b) (c)

Figure 5. Samples of two representative style transfer results on database B. In each example, the input style motion (a) and input content motion (b) are used to synthesize an output motion (c). The style of a walking motion (in a "depressed" style) is transferred to a running motion in the first row, and the style of a walking motion (in a "zombie" style) is transferred to a transition motion (from walking to kicking). The full video sequences can be found in the supplemental video.



(a) (b)

Figure 6. The t-SNE visualization of latent codes from unseen styles, extracted by the network of Aberman et al. [1] (a) and our generative flow model (b). Both of the models are trained on dataset A excluding the action samples in the "strutting" label and "old" styles to evaluate their generalization abilities.

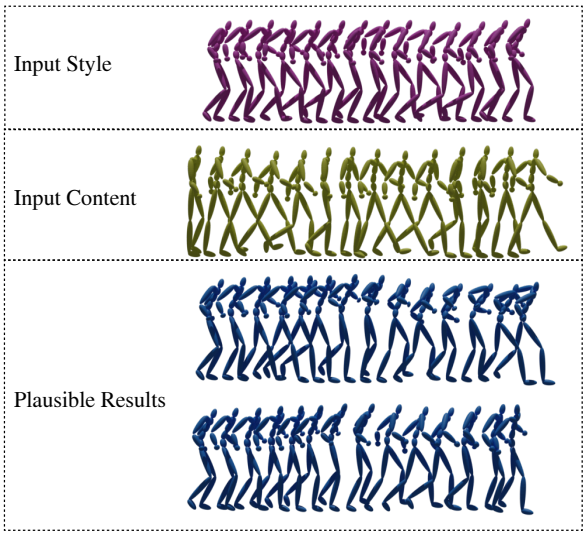


Figure 7. Plausible results of stylized motion synthesis by our generative flow model.

better than the methods of Holden et al. [4] and Aberman et al. [1] in the aspect of style transfer. However, the method of Aberman et al. [1] is judged better than our method in the aspect of content preservation. The reason is that 70% of the evaluated motions provide input motions in the same content to make it easy for the common users to evaluate the style transfer results. Our model tends to generate stylized motions more similar to the style motion with input motions in the same content (as illustrated in our paper Section 4.2). Then, some of the users judge such results as not similar to the content motion. Furthermore, we evaluate the content preservation with input motions in different contents only (3 sets of motions are used), and the results are 19.3%, 38.7% and 42% for the method of Holden et al. [4], the method of Aberman et al. [1] and our method, respectively. It can be seen that our method outperforms the methods of Holden et al. [4] and Aberman et al. [1] in the aspect of content preservation based on input motions in different contents.

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):64, 2020. 2, 3, 4

[2] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In *ICLR (Workshop)*, 2015. 1

[3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *ICLR*, 2017. 1

[4] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4):138:1–138:11, 2016. 3, 4

[5] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, pages 10236–10245, 2018. 1