Audio-Driven Stylized Gesture Generation with Flow-Based Model (Supplementary Material)

Sheng Ye[1], Yu-Hui Wen[1], Yanan Sun[1], Ying He[2], Ziyang Zhang[3], Yaoyuan Wang[3], Weihua He[4], and Yong-Jin Liu[1]

[1] CS Dept, BNRist, Tsinghua University
[2] Nanyang Technological University
[3] Advanced Computing and Storage Lab, Huawei Technologies Co Ltd.
[4] Department of Precision Instrument, Tsinghua University.
{ye-c18,wenyh1616,sunyn20,hwh20,liuyongjin}@tsinghua.edu.cn,
yhe@ntu.edu.sg, {zhangziyang11,wangyaoyuan1}@huawei.com

In this supplementary document, we provide further details of the network structure, the objective evaluation metrics used in our experiments, as well as more experimental results.

1 Detailed Structure of the Invertible Flow

Our neural network builds upon MoGlow. We show the detailed structure of the invertible flow in Fig.1. The Invertible Flow is stacked by 16 flow steps. Each flow step transformation consists of three parts: the Actnorm layer for normalization, the affine coupling layer that performs multiplication and addition transformation, and the invertible 1×1 convolution layer for permutation. Full details can be checked in our code that will be publicly available.

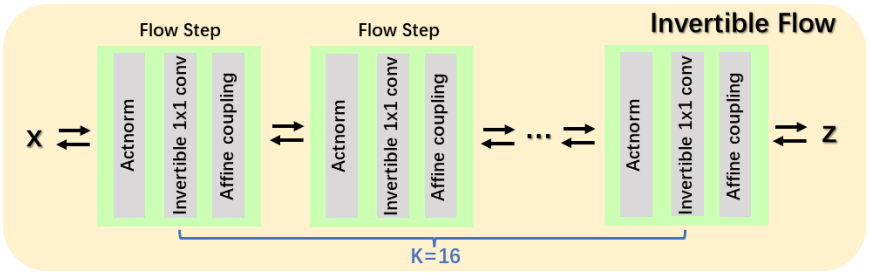

Fig. 1. Detailed structure of the invertible flow.

Y.H Wen and Y.J Liu are the corresponding authors.

2 User Study

We conducted a preliminary user study to evaluate the quality of our results based on human perception. We recruited 12 participants (6 males, 6 females), whose ages are between 21 and 31. The user study is described as follows.

We compared the gestures produced by our method with real gestures on the Trinity Dataset. Each participant watched six randomly ordered gesture videos, of which three are generated gestures, and the other three are real gestures. The participants then rated the gesture in each video clip in terms of Realism ("Does this gesture look real and natural?"), Matching Degree ("Does this gesture match the audio input?"), and Diversity ("Does this gesture have rich details?") using a 5-point scale, with 1 for the worst, and 5 for the best.

Table 1. User study results of comparing gestures produced by our method with real gestures. Higher numbers indicate better results.

Configuration	Realism	Matching Degree	Diversity
Our method	3.41	3.27	3.42
Ground truth	3.50	3.28	3.50

Table 1 reports the results of our preliminary user study. We observed that the participants gave similar ratings to our generated gestures and real gestures. These results confirmed that the generated gestures are diverse and human-like, and match the input audio well.

3 Objective Evaluation Metrics

Percent of Correct Keypoints (PCK) computes the percentage of correctly predicted pose joints. If the L_2 distance between a predicted joint and its target is less than a threshold δ, then this joint is considered to be correct. Previous studies [3, 6] suggest that this metric can be used to evaluate the realism of the generated poses. Specifically, PCK can be calculated as:

$$PCK = \frac{1}{T \times J} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1}[\left\| \hat{x}_t^j - x_t^j \right\|_2 < \delta] \tag{1}$$

where $\hat{x}_t^j$ and x_t^j are the j-th joint of the synthesized pose and the ground-truth pose at the t-th frame, and $\mathbf{1}$ is the indicator function. We set $\delta = 0.1$ in our experiments.

Diversity metric measures whether a model can generate rich and diverse motions, following Li et al. [6]. This metric calculates the distance between different

synthesized gestures. We denote the batch size at test time as B, then the *Diversity* metric is defined as:

$$Div = \frac{2}{B \times (B-1)} \sum_{i=1}^{B-1} \sum_{j=i+1}^{B} \left\| \hat{x}_i - \hat{x}_j \right\|_1 \tag{2}$$

where $\hat{x}_i$ and $\hat{x}_j$ are the i-th and j-th synthesized pose sequence in a batch, and B is set to 50 in our experiments.

Fréchet Gesture Distance (FGD) [8] is a plausible metric consistent with human judgment. FGD computes the Fréchet distance between the Gaussian mean and covariance of the latent feature distributions of synthesized gestures and real gestures. The latent features are extracted by a feature extraction network trained on the Human3.6M dataset. The FGD metric can be calculated as follows:

$$FGD = \left\| \mu_r - \mu_s \right\|^2 + Tr(\Sigma_r + \Sigma_s - 2(\Sigma_r \Sigma_s)^{\frac{1}{2}}) \tag{3}$$

where μ_s and Σ_s are the mean and covariance of latent feature distribution of synthesized gestures, and μ_r and Σ_r are the mean and covariance of latent feature distribution of real gestures.

Beat Alignment Score (BA) is a metric for the correlation between the audio and the motion [7]. This metric measures the average distance between each motion beat (extracted as the local minima of the kinetic velocity) and its nearest corresponding audio beat. The BeatAlign can be computed as :

$$BA = \frac{1}{n} \sum_{i=1}^{n} exp(-\frac{min_{\forall b_j^a \in B^a} \left\| b_i^m - b_j^a \right\|^2}{2\sigma^2}) \tag{4}$$

where $B^m = \{b_i^m\}$ denotes the motion beats, and $B^a = \{b_j^a\}$ denotes the audio beats. We set $\sigma = 5$ in our experiments.

Multi-modality metric (MM) aims to measure how many different motion clips can be sampled for a single audio input. Mathematically, it resembles the Div metric, which can be defined as:

$$MM = \frac{2}{K \times (K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left\| m_i - m_j \right\|_1 \tag{5}$$

where m_i and m_j are generated motion clips given the same audio input, and K is the number of samples.

4 Failure Cases

In rare cases where a long audio sequence contains short segments of silent pause, our method generates dynamic gestures during silent pauses by referring to the target gesture style (Fig.2). The generated gestures may not be desired since silent pauses often correspond to little or even no movement. This problem can be overcome by detecting silent pauses in audio and designing a special module in the network to generate more reasonable gestures. We leave it as future work.

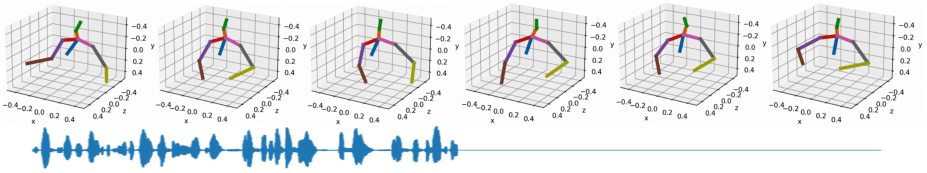

Fig. 2. Our method generates dynamic gestures during silent pauses.

5 Qualitative Comparison of Different Methods

On the TED Dataset, we qualitatively compared our method with two recent approaches: GTC [8] and S2AG [2] (see Fig. 3). We observed that our method can generate plausible and appropriate gestures that align well with the input speech. However, the poses generated by GTC sometimes mismatch the speech and also lack diversity. S2AG generates results visually similar to those of GTC, thereby exhibiting the same artifacts.

On the Trinity Dataset, we qualitatively compared our method with Gesticulator [4] and MoGlow [1] (see Fig. 4). We observed that Gesticulator can only generate tedious and monotonous poses, which are not visually appealing. MoGlow produced some odd-looking, unnatural gestures (see the third row, red rectangle areas in Fig. 4). On the contrary, we observed plausible and realistic gestures generated by our proposed model. We assumed that this is because the global information can be better utilized by adding our proposed global encoder and therefore avoiding the local artifacts.

6 Additional Results

Fig. 5 shows more qualitative results of our method on the TED Dataset [9] and the Trinity Dataset [5]. We also provided a demo video.

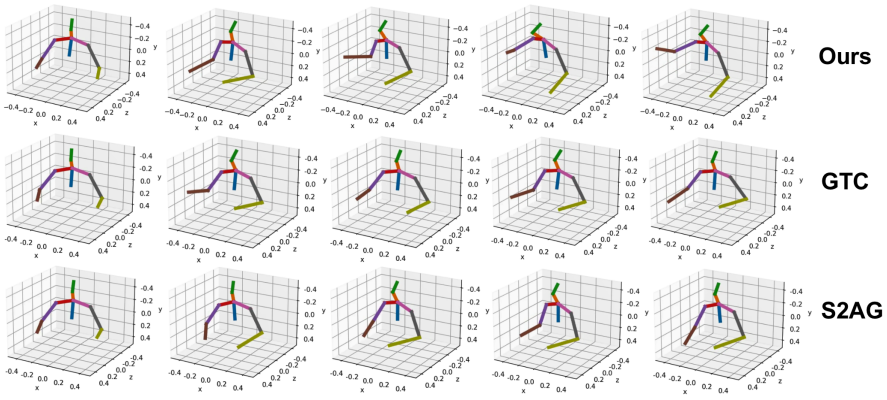

Fig. 3. Qualitative comparisons of our method to two methods: GTC [8] and S2AG [2] on the TED Dataset.

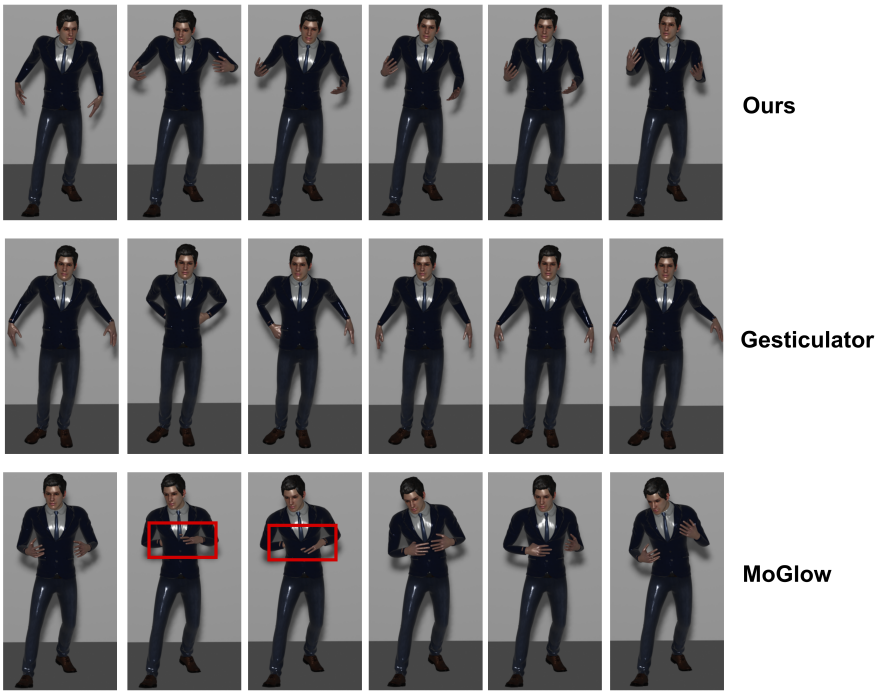

Fig. 4. Qualitative comparisons of our method to the approaches of Kucherenko et al. [4] and Alexanderson et al. [1] on the Trinity Dataset.

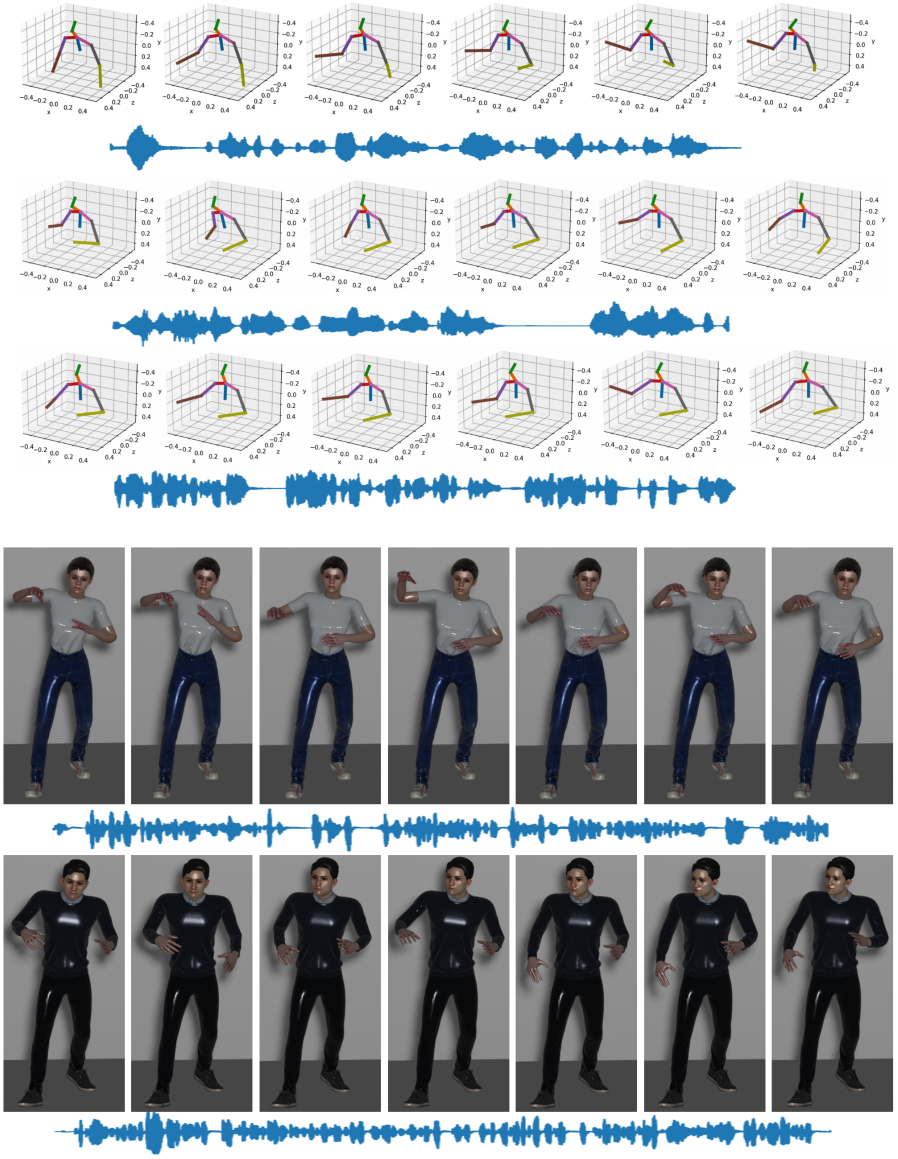

Fig. 5. Additional qualitative results of our method on the TED Dataset and the Trinity Dataset.

References

1. Alexanderson, S., Henter, G.E., Kucherenko, T., Beskow, J.: Style-controllable speech-driven gesture synthesis using normalising flows. In: Computer Graphics Forum. vol. 39, pp. 487–496. Wiley Online Library (2020)
2. Bhattacharya, U., Childs, E., Rewkowski, N., Manocha, D.: Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2027–2036 (2021)
3. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019)
4. Kucherenko, T., Jonell, P., van Waveren, S., Henter, G.E., Alexandersson, S., Leite, I., Kjellström, H.: Gesticulator: A framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 242–250 (2020)
5. Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., Henter, G.E.: A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In: 26th International Conference on Intelligent User Interfaces. pp. 11–21 (2021)
6. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021)
7. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)
8. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
9. Yoon, Y., Ko, W.R., Jang, M., Lee, J., Kim, J., Lee, G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4303–4309. IEEE (2019)