Supplementary Technical Report for Music-to-Dance Synthesis

Huang Hu and Ruozi Huang

1 Introduction

Automatic dance choreography from music is a challenging problem in machine learning research, which has attracted the fastly growing interests in recent years. Most relevant area to this task is human motion prediction [12], which is a challenging task in computer vision. It suffers from the high spatial-temporal complexity, *i.e.*, human motions are highly diverse in the space and have the temporal dependency property in the time dimension. Early works in this area try to tackle this task by the hidden markov models [9], Gaussian processes [16, 17] and restricted boltzmann machines [14], while recent works leverage recurrent neural networks (RNNs) [3, 7, 4] to model the spatial-temporal dependency among human motions. However, the longest motion sequences produced by these works can only last several seconds under 30 frame per seconds (FPS), which is far from the requirement that a formal dance should last one minute at least.

Although the music-to-dance synthesis is an interesting research topic that can exhibit the higher-level machine intelligence, it is still extremely hard for researchers to design the efficient models that can generate long-term smooth and realistic dances from music due to the following challenges:

(1) Lack of the high-quality music-dance paired training data. Most existing works leverage the human pose estimation models like OpenPose [1] to extract motion sequences from real dance videos. While these extracted motions usually contain lots of noisy keyjoints due to imperfect detection and vary a lot in the shape of human skeletons;

(2) The generated dance motion sequences need to be consistent with the input music in terms of style, rhythm, beat and etc;

(3) The synthesized dance motions should be realistic as much as possible, *i.e.*, the produced motions should conform to the 2D or 3D spatial structure of human skeletons;

(4) Dance poses are complex and diverse in the space. The pose at the certain time-step has the highly-diverse subsequent poses, *i.e.*, the successive motions have the randomness property while simultaneously the transitions among them need to be naturally smooth too;

(5) Dance is a long-term sequence composed of human motions with the spatio-temporal structure. Hence, generating dances also suffers from the **exposure bias issue**[1] [11, 5, 20, 6, 18, 13, 15, 19] known in natural language generation (NLG), referring to the train-test discrepancy of autoregressive models. This issue would quickly accumulate the prediction errors at inference and thus make the generated motion sequences rapidly converge to the mean poses, *i.e.*, "freezing motions". Moreover, it becomes much more severe in the long sequence generation of real-valued vectors in continuous space (each motion is represented by a dozen of 2D or 3D keyjoints), compared to NLG (sequence generation of symbols in discrete space). This is one of most crucial challenges in the music-to-dance synthesis and human motion prediction tasks.

[1] After communicating with some researchers in the computer vision community, we found few of them knew the exposure bias problem in the sequence generation due to the knowledge bias.

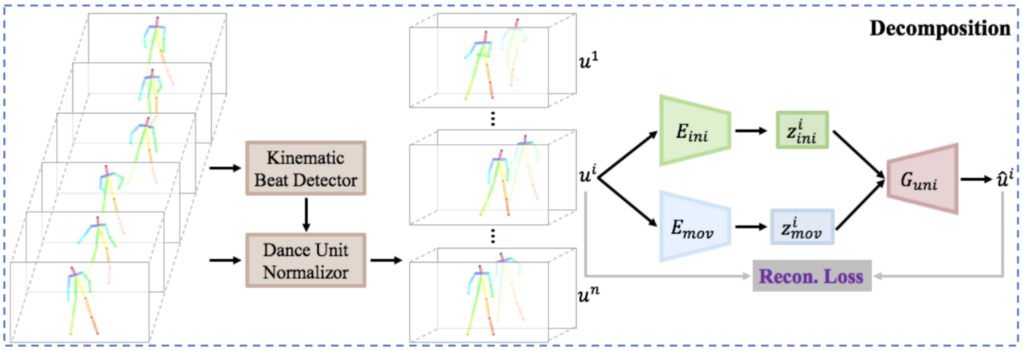

Figure 1: The flowchart of decomposition stage.

2 Paper Analysis for Dancing to Music

This paper [8] published in NeurIPS 2019 makes the primary attempt on this direction. They propose a decomposition-to-composition method to address the dance generation with music under GAN framework. Specifically, their method first decomposes the whole dance motion sequence into small dance units and generates them individually. Then these units are composed by using the last dance motion of current unit to initialize the first dance motion of the next unit. Hence, this work simplifies the problem by the proposed decomposition-to-composition framework, and avoid to directly handle the severe exposure bias issue in the long-term generation since each unit has only 32 dance movements.

2.1 Decomposition Stage

Figure 1 shows the decomposition process which consists of two steps. The first step is to divide the full dance motion sequences into the small dance units by a developed kinematic beat detector, while the second is learning a generation model to generate the dance motions for each unit.

In the music information retrieval (MIR) community, the beat tracking [2] is usually utilized to detect the music beats and related detection methods have been already well-developed. Popular libraries for audio and music analysis in MIR include Librosa [10], Essentia[2] and etc. While different from music beat detection, there are few works studied on detecting dance beats in existing literature. This paper proposes to find the cutting position by detecting whether its motion has the sudden deceleration at current time-step. Concretely, they compute the motion magnitude and angle of each keyjoint between neighboring motions, and track the magnitude and angle trajectories to spot when a dramatic decrease in the motion magnitude or a substantial change in the motion angle happens. Besides, another commonsense is that human dancers usually step on music beats during dancing but would not step on every music beat. In other words, the kinematic beats are aligned to music beats but do not need to be aligned with every music beat. Figure 2 demonstrates the alignment between music beats and kinematic beats from their paper, which is consistent with the commonsense. Hence, this work divides the whole dance motion sequence into small dance units with the same number of motions to simplify the long-term sequential generation.

The next step is to learn a generation model to generate the motion sequence for each unit. This work proposes a dance unit variational auto-encoder (DU-VAE) to disentangle the dance unit into two latent spaces, namely an initial movement space $\mathcal{Z}_{ini}$ and a normal movement space $\mathcal{Z}_{mov}$. The reconstruction loss and KL loss can be calculated as follows:

$$\begin{aligned} L^u_{\text{recon}} &= E[\|G_{uni}(z_{ini}, z_{mov}) - u\|_1], \\ L^u_{\text{KL}} &= E[\text{KL}(\mathcal{Z}_{ini}\|N(0,\text{I}))] + E[\text{KL}(\mathcal{Z}_{mov}\|N(0,\text{I}))], \end{aligned} \tag{1}$$

They claim that this design can facilitate the long-term sequential generation since the last dance motion of current unit could be used to initialize the generation of next unit.

[2] http://essentia.upf.edu/

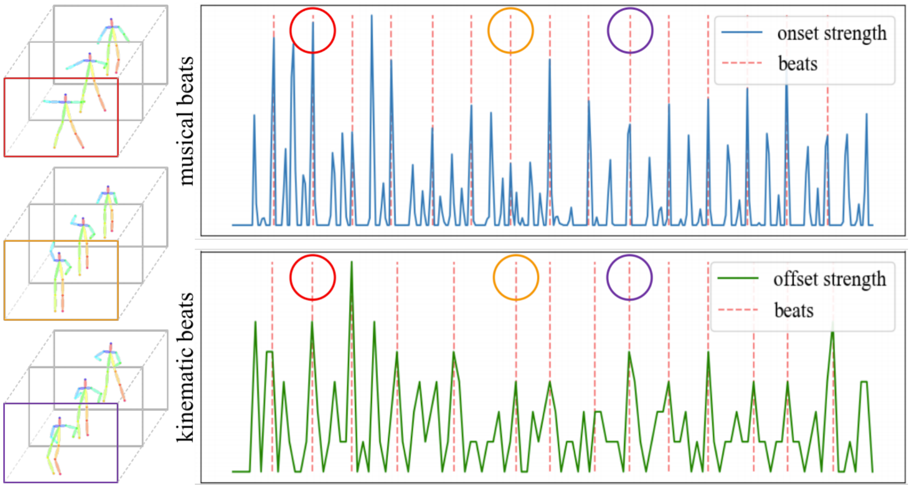

Figure 2: The extracted music beats and kinematic beats from a piece of dance. The dash lines denote the positions where the beats occur.

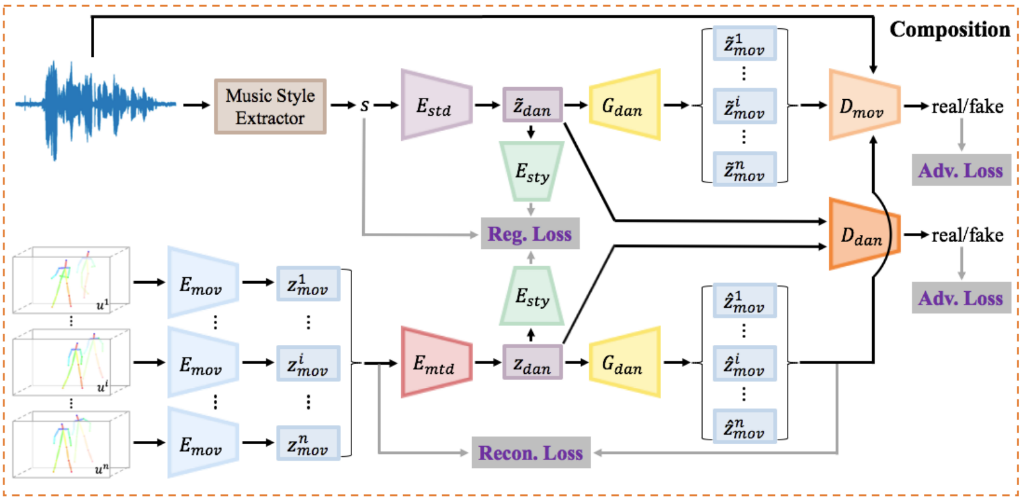

Figure 3: The flowchart of composition stage.

2.2 Composition Stage

For composition stage shown in Figure 3, this paper proposes a music-to-movement GAN (MM-GAN) to model the music-conditioned dance generation. Based on the movement space $\mathcal{Z}_{mov}$ learned in decomposition stage, it introduces a movement-to-dance encoder E_{mtd}: $\{z^i_{mov}\} \rightarrow z_{dan}$ to compress $\{z^i_{mov}\}$ into a latent variable z_{dan} of dance, and a generator G_{dan} to reconstruct z_{dan} back to $\{\hat{z}^i_{mov}\}$. Thus, the reconstruction loss can be formulated as follow:

$$L^m_{\text{recon}} = E[\left\|\{\hat{z}^i_{mov}\} - \{z^i_{mov}\}\right\|_1]. \tag{2}$$

For the music part, this paper only utilizes the style information s which is extracted by a pre-trained music style classifier, *i.e.*, the predicted probability vector of the classifier. Then it introduces a style-to-dance encoder E_{std}: $(s, \epsilon) \rightarrow \tilde{z}_{dan}$ to encode s into another dance latent space $\tilde{z}_{dan}$, where ϵ is a Gaussian noise. After that, $\tilde{z}_{dan}$ is reconstructed back to $\{\tilde{z}^i_{mov}\}$ by the same generator G_{dan}. Finally, a discriminator D_{mov} is introduced to judge the real or fake for inputs and the adversarial loss is defined as:

$$L^m_{\text{adv}} = E[\log D_{mov}(\{\hat{z}^i_{mov}\}, a) + \log\left(1 - D_{mov}(\{\tilde{z}^i_{mov}\}, a)\right)], \tag{3}$$

where a is the style feature of the music. Compared to directly aligning dance movements generated from music and real dance movements by AE, utilizing a discriminator to judge the real or fake for the given inputs is a much looser objective that allows the generated dances to have the certain diversity while simultaneously ensuring the match of generated dance and given music.

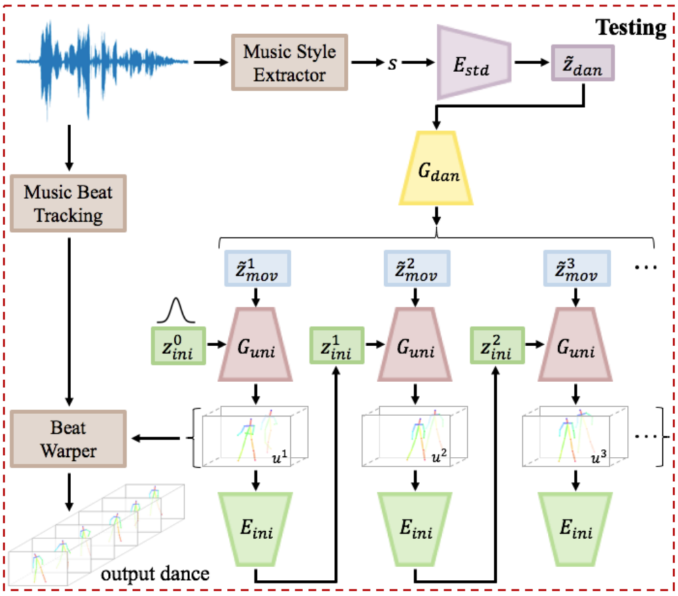

Figure 4: The flowchart of inference time.

Note that even though removing the auto-endcoder (AE) in dance part, we can still train the model. However, introducing such an AE brings two advantages: (1) The reconstruction loss in Eq. 2 can be used to enhance the ability of G_{dan} since z_{dan} is a relatively simple distribution, and thus the majority of information is compressed into the generator; (2) It can align the two latent variables in z_{dan} space from music and dance respectively following:

$$\begin{gathered} L^d_{\text{adv}} = E[\log D_{dan}(z_{dan}) + \log(1 - D_{dan}(\tilde{z}_{dan}))], \\ L^d_{\text{KL}} = E[\text{KL}(\mathcal{Z}_{dan} \| N(0, \text{I}))]. \end{gathered} \tag{4}$$

Considering the style consistency regularization defined as:

$$L^s_{\text{recon}} = E[\|E_{sty}(z_{dan}) - s\|_1 + \|E_{sty}(\tilde{z}_{dan}) - s\|_1]. \tag{5}$$

The final training objective of MM-GAN is given by:

$$L_{\text{comp}} = L^m_{\text{recon}} + \lambda^s_{\text{recon}} L^s_{\text{recon}} + \lambda^m_{\text{adv}} L^m_{\text{adv}} + \lambda^d_{\text{adv}} L^d_{\text{adv}} + \lambda^d_{\text{KL}} L^d_{\text{KL}}, \tag{6}$$

2.3 Inference

When the training is finished, the framework can now follow Figure 4 to do the generation for a given music. Specifically, it first encodes extracted music style feature s into the learned latent dance space $\tilde{z}_{dan}$ and utilize the generator G_{dan} learned in composition stage to generate the latent unit sequence $\{\tilde{z}^i_{mov}\}$. Then, another generator G_{uni} learned in decomposition stage is applied to generating short motion sequences for each unit in an autoregressive manner as:

$$u^i = G_{uni}(z^{i-1}_{ini}, z^i_{mov}), \qquad z^i_{ini} = E_{ini}(u^i(-1)), \tag{7}$$

where z^0_{ini} is initialized by sampling from a standard normal distribution.

2.4 Summary

This paper proposes a decomposition-to-composition framework to simplify the problem of long-term dance generation with music, and thus avoid the severe exposure bias problem in long-term motion sequence generation. However, in practice, we found the dance motion sequences generated by their method can only last 20 to 30 seconds at most. Besides, the way of composing small motion units into a full dance sequence prevents their approach from producing the naturally smooth dance sequences. Since the concatenation positions of neighboring dance units have the obvious motion jumps in the experiment. Although this work tries to tackle the music-to-dance synthesis task under GAN framework, it still fails to directly handle the exposure bias issue (*i.e.*, the key challenge) in the long-term dance generation with music.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1302–1310. IEEE Computer Society, 2017.

[2] Daniel PW Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[3] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4346–4354. IEEE Computer Society, 2015.

[4] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[5] Anirudh Goyal, Alex Lamb, Ying Zhang, Saizheng Zhang, Aaron C. Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4601–4609, 2016.

[6] Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Quantifying exposure bias for neural language generation. *arXiv preprint arXiv:1905.10617*, 2019.

[7] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5308–5317. IEEE Computer Society, 2016.

[8] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3581–3591, 2019.

[9] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1314–1321. IEEE Computer Society, 2014.

[10] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015.

[11] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[12] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

[13] Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 157–167, Hong Kong, 2019. Association for Computational Linguistics.

[14] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1345–1352. MIT Press, 2006.

[15] Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online, 2020. Association for Computational Linguistics.

[16] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 1441–1448, 2005.

[17] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007.

[18] Yifan Xu, Kening Zhang, Haoyu Dong, Yuezhou Sun, Wenlong Zhao, and Zhuowen Tu. Rethinking exposure bias in language modeling. *arXiv preprint arXiv:1910.11235*, 2019.

[19] Liping Yuan, Jiangtao Feng, Xiaoqing Zheng, and Xuanjing Huang. Alleviate exposure bias in sequence prediction with recurrent neural networks. *arXiv preprint arXiv:2103.11603*, 2021.

[20] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy, 2019. Association for Computational Linguistics.