Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning

Ruozi Huang[1], Huang Hu[2], Wei Wu[2], Kei Sawada[2],
Mi Zhang[1] and Daxin Jiang[2]

1. Fudan University, Shanghai, China
2. Microsoft STCA, Beijing, China

Microsoft

ICLR

Motivations

- Dancing to music is one of human's innate abilities. The neurological mechanism behind dancing behavior motivates us to **explore a computational approach to dance creation from music**

- Technologies on automatic dance generation from music can potentially **help choreographers design new dances** for given music

- With the help of 3D animation software, such technologies can drive various 3D virtual characters and **show the great potential for virtual advertisement video generation**

Challenges

- Prediction of human motion dynamics suffers from **high spatial-temporal complexity**

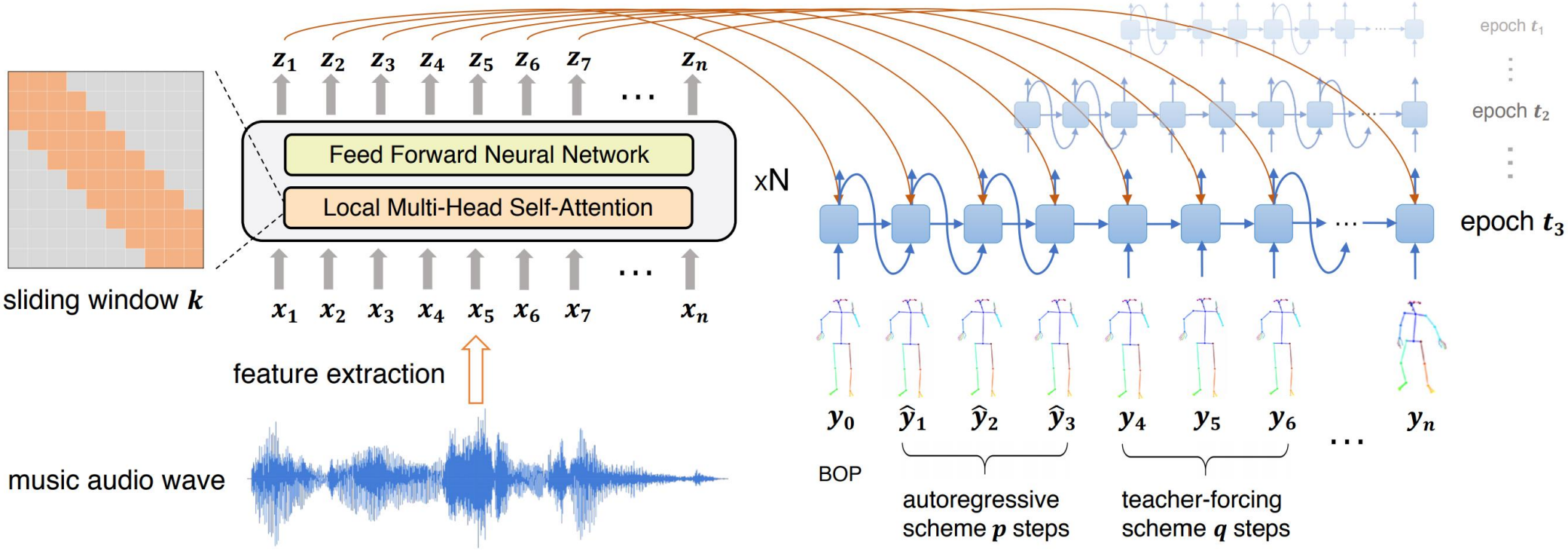
- Existing methods synthesize new human motion sequences through autoregressive models, which often generate short sequences due to an **accumulation of prediction errors** that are fed back into neural network

- Existing methods do **NOT consider the consistency** between dance and music in terms of **style, rhythm and beat** during modeling

Contributions

- Propose a novel Seq2Seq architecture for music-conditioned dance generation task
- Propose a curriculum learning strategy to alleviate the error accumulation of autoregressive models in long motion sequence generation
- Release a high-quality 2D human dance motion dataset paired with music, which contains three genres, i.e., ballet, hip-hop and Japanese pop

Model Overview



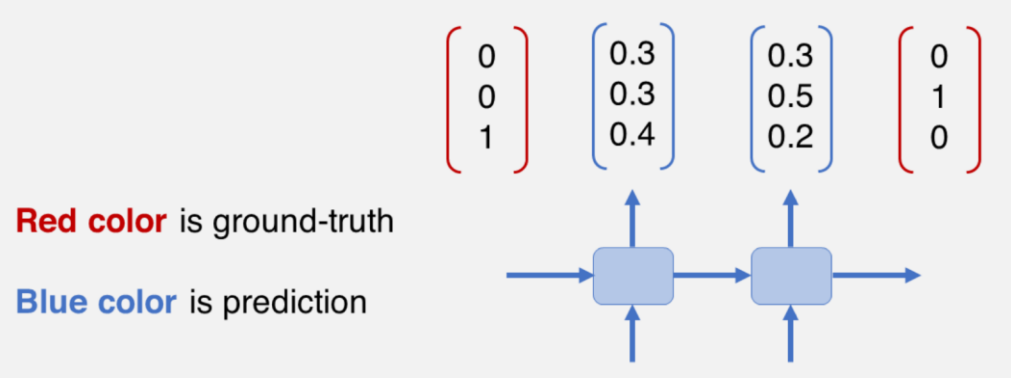
$$p = f(t) \in \{const, \lfloor \lambda t \rfloor, \lfloor \lambda t^2 \rfloor, \lfloor \lambda e^t \rfloor\}$$

Transformer-based encoder with local self-attention mechanism

RNN-based decoder with proposed curriculum learning strategy

Dynamic Auto-Conditioned Learning Approach

- Scheduled sampling is proposed for NLG (**discrete tokens**) while our learning approach is tailored for motion sequence generation (**each motion is a real-valued vector in continuous space**)
- In NLG, the bias of predicted probability distribution over vocabulary can be corrected by sampling strategy. While any small bias of predicted motion at each step will be accumulated and propagated to the future



If the blue vectors represent generated motions,

$$bias = (|0 - 0.3| + |0 - 0.3| + |1 - 0.4|) + (|0 - 0.3| + |1 - 0.5| + |0 - 0.2|)$$

If the blue vectors represent probability distributions, it can still generate the target words

Dataset

- Collect the solo dance videos from YouTube by crowd-workers
- Trim the beginning and ending few seconds for each video to remove silent parts, and split them into 1-min video clips
- Extract 2D pose data from these clips using OpenPose [Cao et al., CVPR 2017] with 15 FPS

Table 1: Statistics of the dataset.

Category	Num of Clips	Clip Length (min)	FPS	Resolution
Ballet	136	1	15	720P
Hiphop	298	1	15	720P
Japanese Pop	356	1	15	720P

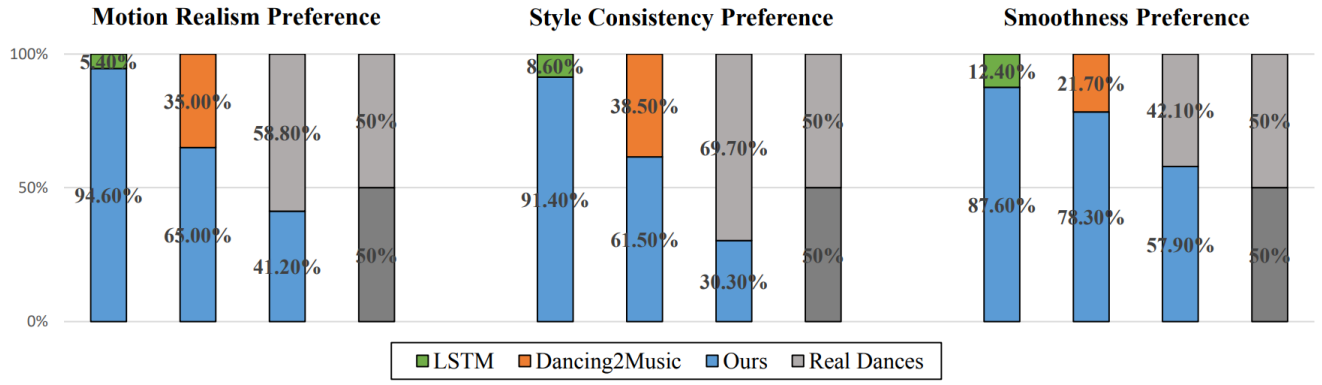
Baselines

- **Dancing2Music** [Lee et al., NeurIPS 2019]
  - Primary baseline, the state-of-the-art method on music-conditioned dance generation task.
  - Propose a decomposition-to-composition framework and leverage GANs to generate motion sequence for each dance unit
- **AudMoCoGAN**
  - Auxiliary baseline, modify MoCoGAN [Tulyakov et al., CVPR 2018] to take audio as the input, which maps a sequence of random vectors (audio feature vectors) to a sequence of video frames
- Audio to body dynamics (**LSTM**) [Shlizerman et al., CVPR 2018]
  - Map audio to arm and body dynamics

Evaluation Results

- Our proposed approach outperforms the baselines on all automatic metrics except for Multimodality
- Human judgement further demonstrates the superior performance of our approach
- Generated demo is here

Method	FID	ACC (%)	Beat Coverage (%)	Beat Hit Rate (%)	Diversity	Multimodality
Real Dances	2.6	99.5	56.4	63.9	40.2	-
LSTM	51.9	12.1	4.3	9.7	16.8	-
Aud-MoCoGAN	48.5	33.8	10.9	28.5	30.7	16.4
Dancing2Music	22.7	60.4	15.7	65.7	30.8	**18.9**
Ours	**6.5**	**77.6**	**21.8**	**68.4**	**36.9**	15.3



Long-Term Generation

- Split generated dance of 1-min into 15 4-second clips and measure FID of each clip
- FID scores of LSTM and Aud-MoCoGAN grow rapidly since the generated dance motions quickly become frozen
- Dancing2Music maintains the relatively stable FID scores all the time, which benefits from its decomposition-to-composition method
- Our approach performs well all the time

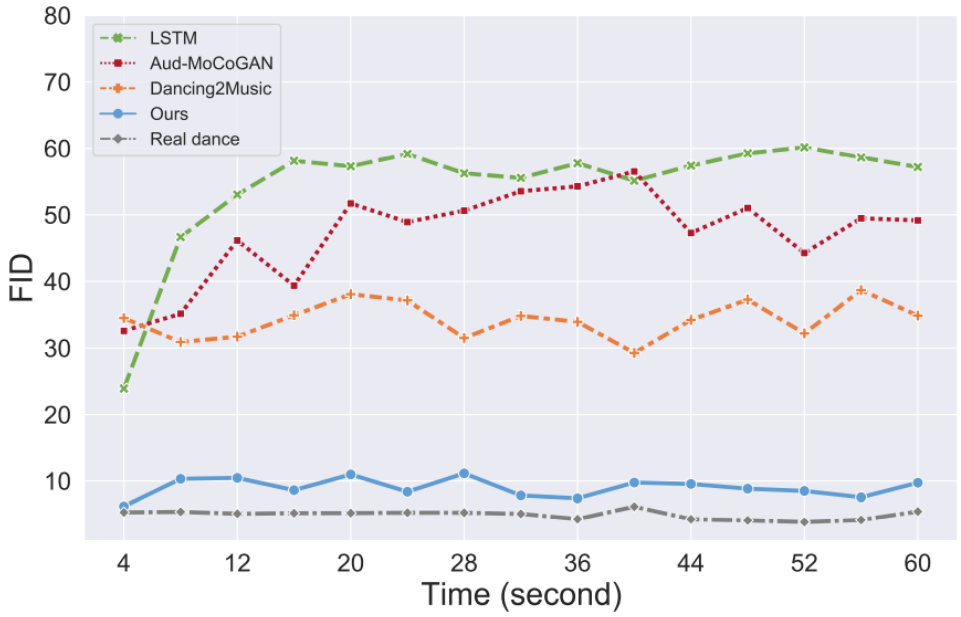


Figure 4: FID curves of different methods over time.

Ablation Study

- The FID and accuracy of local self-attention is close to the global one with less GPU memory usage and it is obviously better than other alternatives

- Our learning approach significantly outperforms the teacher-forcing and auto-condition. Among them, linear growth function achieves the best performance

Encoder Structure	FID	ACC (%)
LSTM	7.9	45
ConvS2S encoder	13.3	54
Global self-attention	3.6	80
Local self-attention	4.3	78.5

Learning Approach	FID	ACC (%)
Teacher-forcing	61.2	5.3
Auto-condition ($const$)	15.7	35
Ours ($\lfloor \lambda e^t \rfloor$)	9.8	69
Ours ($\lfloor \lambda t^2 \rfloor$)	6.4	73
Ours ($\lfloor \lambda t \rfloor$)	5.1	77.6

Conclusion

- The proposed Seq2Seq architecture can efficiently process long sequences of music features with local self-attention mechanism, and generate dance sequences that are **beat-match, style-consistent** with music

- The dynamic auto-conditioned learning approach can effectively alleviate error accumulation, and enable the decoder to generate **long sequences** (**1-min length**) **of non-freezing motions**

- Release 2D dance dataset to facilitate the further research on this topic

Thanks

Code&data link: https://github.com/stonyhu/DanceRevolution