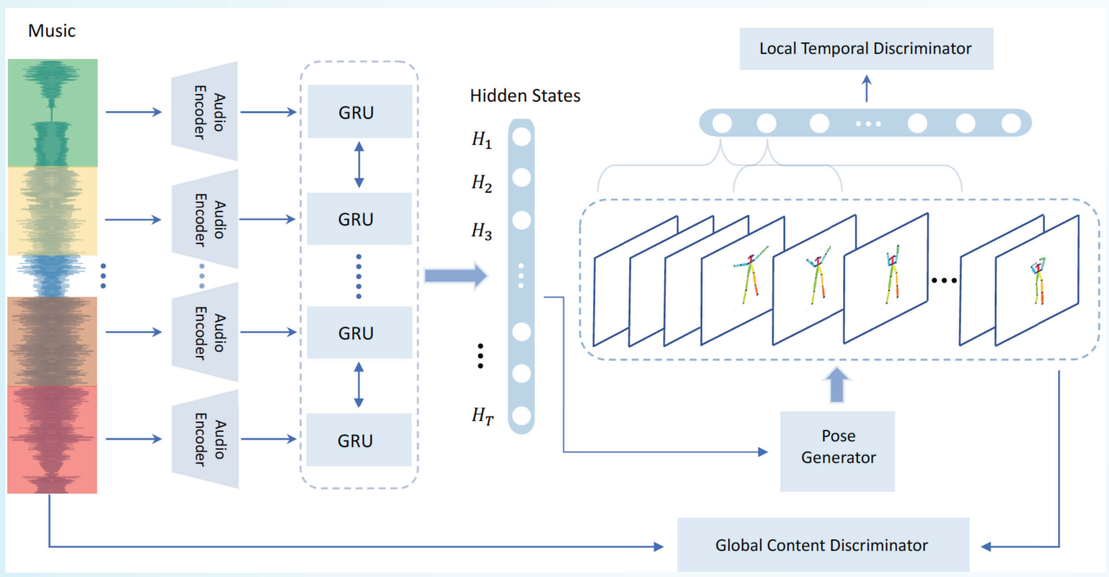
Undergraduate Research Opportunities Program

Self-supervised Dance Video Synthesis Conditioned on Music

Student Name:	REN Xuanchi	**Project Supervisor:**	Prof. Qifeng Chen
Major / School:	COSC / SENG	**Dept. / School :**	COMP / SENG

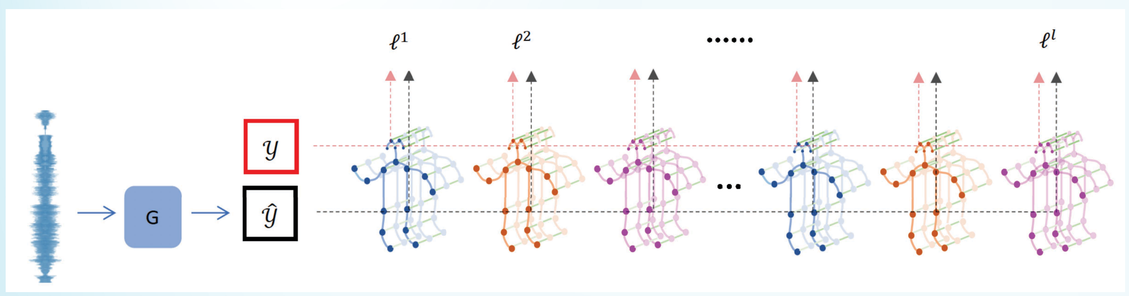
We present a self-supervised approach with pose perceptual loss for automatic dance video generation. Our method can produce a realistic dance video that conforms to the beats and rhymes of given music. To achieve this, we firstly generate a human skeleton sequence from music and then apply the learned pose-to-appearance mapping to generate the final video. In the stage of generating skeleton sequences, we utilize two discriminators to capture different aspects of the sequence and propose a novel pose perceptual loss to produce natural dances. Besides, we also provide a new cross-modal evaluation to evaluate the dance quality, which is able to estimate the similarity between two modalities (music and dance). Finally, experimental qualitative and quantitative results demonstrate that our dance video synthesis approach produces realistic and diverse results.

Overview



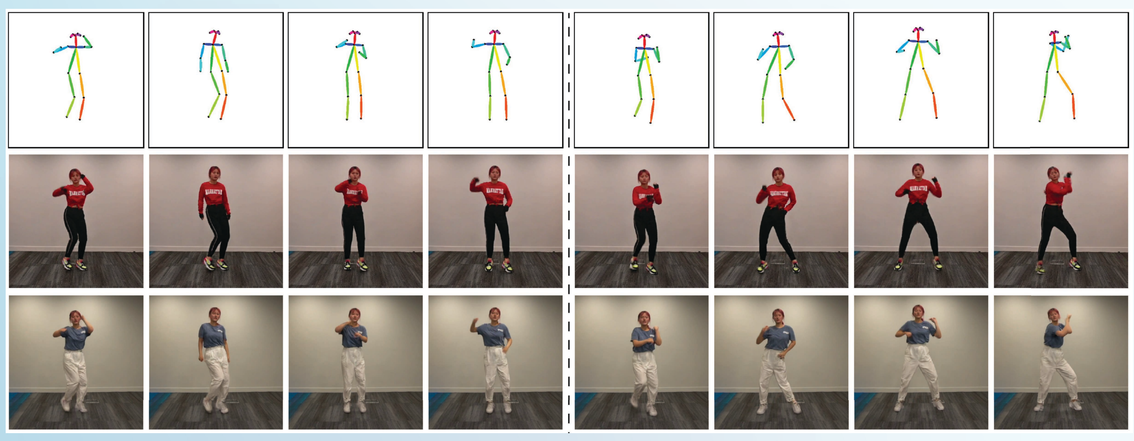
Our framework for music-oriented dance skeleton sequence synthesis. The input music signals are first divided into pieces of 0.1-second music. The generator in our model contains an audio encoder, a bidirectional GRU, and a pose generator. The output skeleton sequence of the generator is fed into the Global Content Discriminator to evaluate the consistency with the input music. The generated skeleton sequence is also divided into overlapping sub-sequences, which are fed into the Local Temporal Discriminator for local temporal consistency.

Pose Perceptual Loss



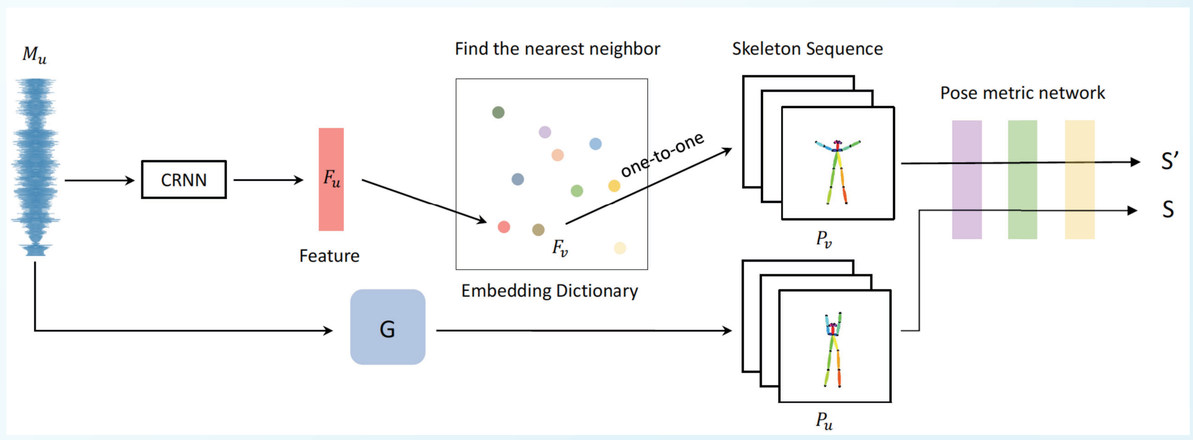
We propose to directly match features in a pose recognition network that takes human skeleton sequences as input. We use ST-GCN that is a Graph Convolutional Network (GCN) for a pose recognition to extract deep features. ST-GCN utilizes a spatial-temporal graph to form the hierarchical representation of skeleton sequences to learn both spatial and temporal patterns from data. As shown in the above figure, our generator can stably generate poses with the pose perceptual loss.

Pose to Video



Synthesized video conditioned on the music “LIKEY" by TWICE. For each 5-second dance video, we show four frames. The top row shows the skeleton sequence, and the bottom rows show the synthesized video frames conditioned on different target videos.

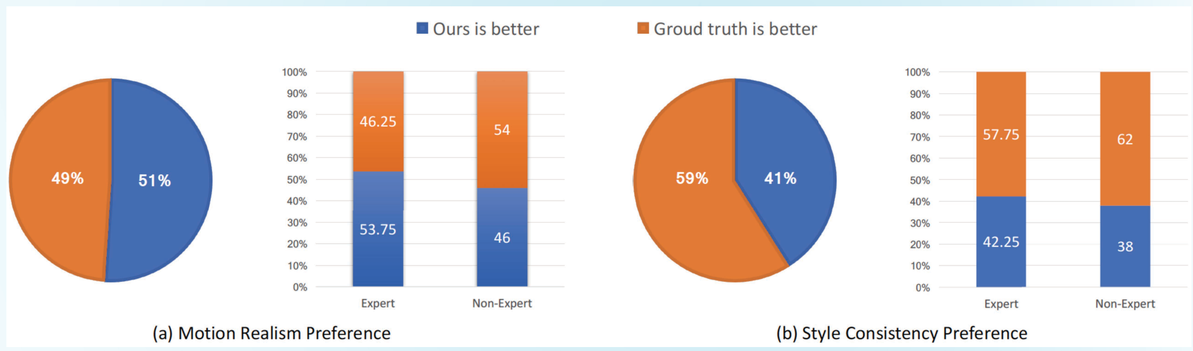
Evaluation



We propose a cross-modal metric, as shown in Figure 8, to estimate the similarity between music and dance.

	FID	Diversity	Cross-modal
Real	–	26.12	0.043
L_1	37.92	17.71	0.312
Global D	18.04	20.33	0.094
Local D	15.86	19.57	0.068
Our model	**3.80**	**25.63**	**0.046**

Comparison between our model and baselines. For FID and the cross-modal evaluation, lower is better. For Diversity, higher is better.



Results of user study on comparisons between our synthesized skeleton sequence and the ground truth. For each comparison, the participant should select the dances that are more realistic regardless of music and better match the style of music.

Qualitative Evaluation