Let's Play Music: Audio-driven Performance Video Generation

Hao Zhu[1,2], Yi Li[2,2], Feixia Zhu[4], Aihua Zheng[4], Ran He[2,3]

[1]Anhui University [2]NLPR & CEBSIT & CRIPAC, CASIA [3]University of Chinese Academy of Sciences

Background

New Task:

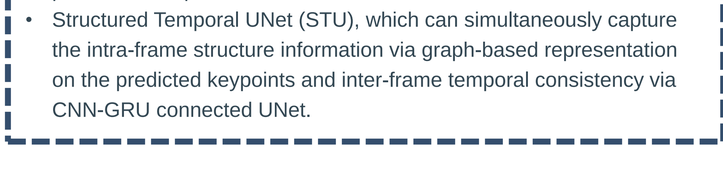
- Audio-driven Performance Video Generation (APVG), which aims to synthesize the performance video from the given music

Premise:

- A professional can know relationship between a given piece of music and corresponding performance movements.

Challenges:

- Generate precise motion details such as body and fingers from the low-dimensional audio information.



Motivation

Multi-stage generation:

- It is difficult to directly generate high quality video from music.
- The coarse video frames provide the texture information
- Keypoints provide motion information

Intra-frame structured information

- Performing video requires maintaining finger movements.
- Graph Convolutional Network to discover the intraframe structured information from feature block

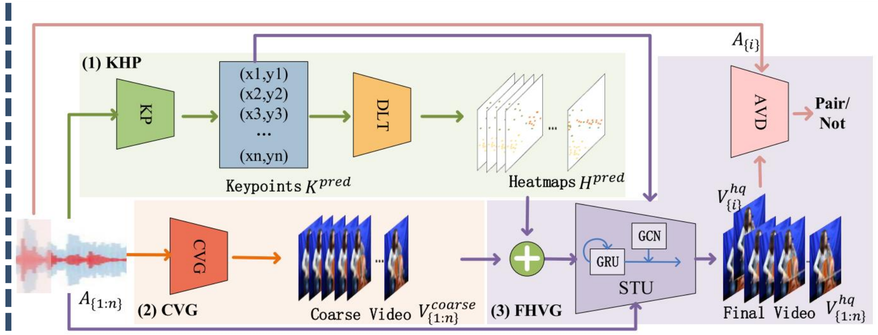
Inter-frame temporal information

- Temporal consistency is a important metric in video generation.
- GRU to preserve the inter-frame temporal consistency

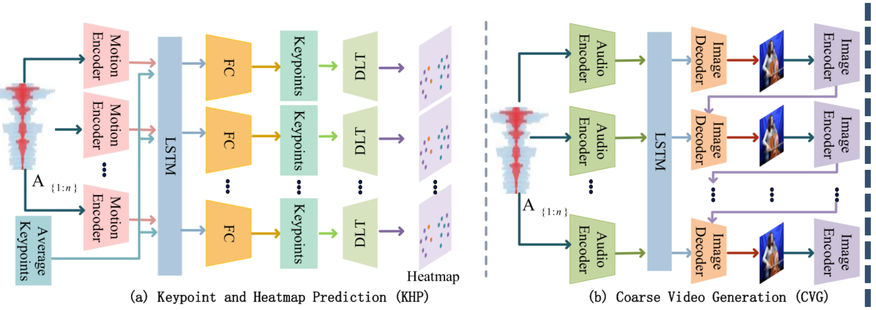
Contribution

- A multi-stage adversarial generation model to achieve the APVG task, which casts a new challenging problem for audio-visual computation and provides a baseline framework for related researches and potential applications.
- Transform the predicted keypoints to corresponding heatmap by utilizing a differentiable landmark transformer (DLT) to provide more precise local spatial information
- Structured Temporal UNet (STU), which can simultaneously capture the intra-frame structure information via graph-based representation on the predicted keypoints and inter-frame temporal consistency via CNN-GRU connected UNet.

Pipeline

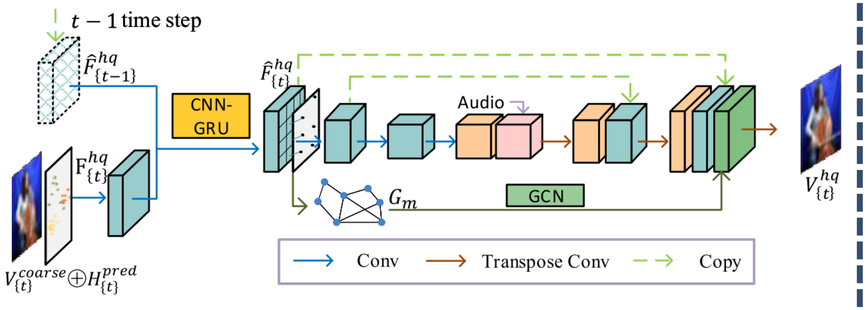


The pipeline of our proposed model.



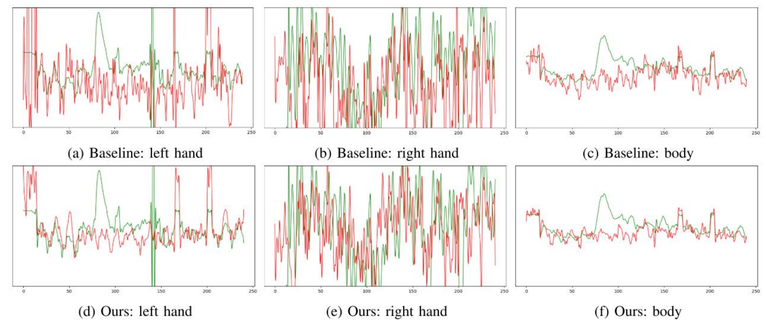
(1) Keypoint and Heatmap Prediction (KHP) which predicts the keypoints from the given music, and then transforms the them into corresponding heatmap.

(2) Coarse Video Generation (CVG) which generates the coarse video from given music



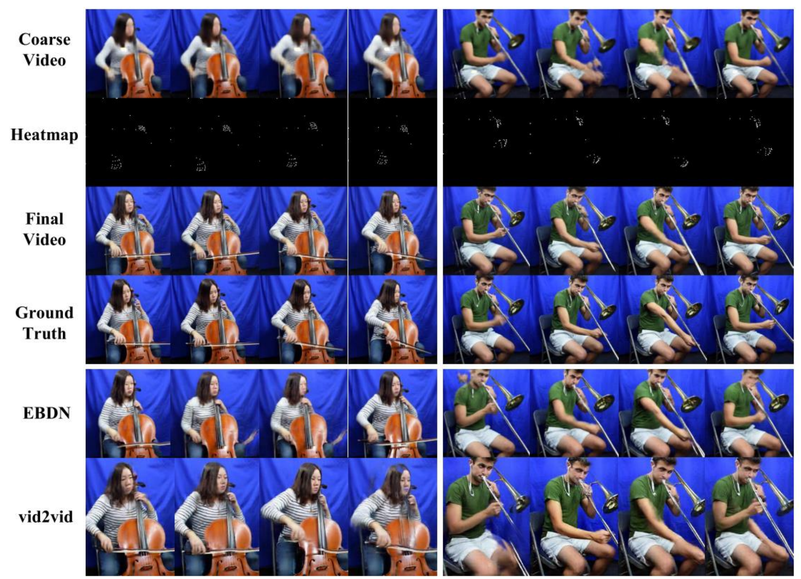
(3) Final Performance Video Generation (FPVG), which integrates the graph represented intra-frame structure information from predicted keypoints via GCN module and temporal information via CNN-GRU module.

Experiments on Keypoints Prediction



Visualization of cello keypoints, where X-axis and Y-axis denote each sample and the 1-D PCA feature respectively. The red line and green line indicate the PCA features of predicted and ground truth keypoints respectively

Experiments on Video Generation



The generation examples of our model