Foley Music: Learning to Generate Music from Videos

Chuang Gan[1,2], Deng Huang[2], Peihao Chen[2],
Joshua B. Tenenbaum[1], and Antonio Torralba[1]

[1] Massachusetts Institute of Technology
[2] MIT-IBM Watson AI Lab
http://foley-music.csail.mit.edu

Abstract. In this paper, we introduce *Foley Music*, a system that can synthesize plausible music for a silent video clip about people playing musical instruments. We first identify two key intermediate representations for a successful video to music generator: body keypoints from videos and MIDI events from audio recordings. We then formulate music generation from videos as a motion-to-MIDI translation problem. We present a *Graph−Transformer* framework that can accurately predict MIDI event sequences in accordance with the body movements. The MIDI event can then be converted to realistic music using an off-the-shelf music synthesizer tool. We demonstrate the effectiveness of our models on videos containing a variety of music performances. Experimental results show that our model outperforms several existing systems in generating music that is pleasant to listen to. More importantly, the MIDI representations are fully interpretable and transparent, thus enabling us to perform music editing flexibly. We encourage the readers to watch the supplementary video with audio turned on to experience the results.

Keywords: Audio-Visual, Sound Generation, Pose, Foley

1 Introduction

Date Back to 1951, British computer scientist, Alan Turing was the first to record computer-generated music that took up almost an entire floor of the laboratory. Since then, computer music has become an active research field. Recently, the emergence of deep neural networks facilitates the success of generating expressive music by training from large-scale music transcriptions datasets [40,28,11,62,46]. Nevertheless, music is often accompanied by the players interacting with the instruments. Body and instrument interact with nuanced gestures to produce unique music [23]. Studies from cognitive psychology suggest that humans, including young children, are remarkably capable of integrating the correspondences between acoustic and visual signals to perceive the world around them. For example, the McGurk effect [37] indicates that the visual signals people receive from seeing a person speak can influence the sound they hear.

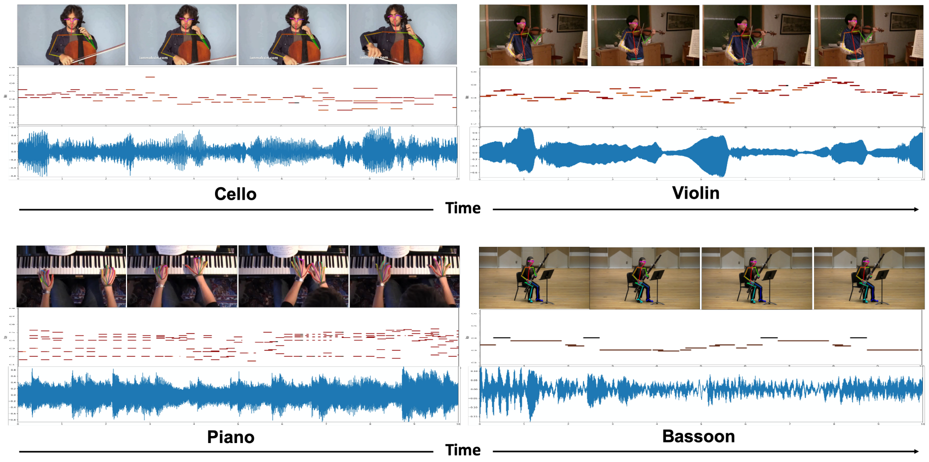

Fig. 1. Given a video of people playing instrument, our system can predict the corresponding MIDI events, and generate plausible musics.

An interesting question then arises: given a silent video clip of a musician playing an instrument, could we develop a computational model to automatically generate a piece of plausible music in accordance with the body movements of that musician? Such capability serves as the foundations for a variety of applications, such as adding sound effects to videos automatically to avoid tedious manual efforts; or creating auditory immersive experiences in virtual reality;

In this paper, we seek to build a system that can learn to generate music by seeing and listening to a large-scale music performance videos (See Figure 1). However, it is an extremely challenging computation problem to learn a mapping between audio and visual signals from unlabeled video in practice. First, we need a visual perception module to recognize the physical interactions between the musical instrument and the player's body from videos; Second, we need an audio representation that not only respects the major musical rules about structure and dynamics but also easy to predict from visual signals. Finally, we need to build a model that is able to associate these two modalities and accurately predict music from videos.

To address these challenges, we identify two key elements for a successful video to music generator. For the visual perception part, we extract key points of the human body and hand fingers from video frames as intermediate visual representations, and thus can explicitly model the body parts and hand movements. For the music, we propose to use Musical Instrument Digital Interface (MIDI), a symbolic musical representation, that encodes timing and loudness information for each note event, such as note-on and note-off. Using MIDI musical representations offers several unique advantages: 1) MIDI events capture the expressive timing and dynamics information contained in music; 2) MIDI is a sequence of symbolic representation, thus relatively easy to fit into machine learning models; 3) MIDI representation is fully interpretable and flexible; 4) MIDI could be easily converted to realistic music with a standard audio synthesizer.

Given paired data of body keypoints and MIDI events, music generation from videos can be posed as a motion to MIDI translation problem. We develop a *Graph−Transformer* module, which consists of a GCN encoder and a Transfomer decoder, to learn a mapping function to associate them. The GCN encoder takes input the coordinates of detected keypoints and applies a spatial-temporal graph convolution strategy to produce latent feature vectors over time. The transformer decoder can then effectively capture the long-term relationships between human body motion and MIDI events using the self-attention mechanism. We train the model to generate music clips of accordion, bass, bassoon, cello, guitar, piano, tuba, ukulele, and violin, using large-scale music performance videos. To evaluate the quality of our predicted sounds, we conduct listener study experiments measured by correctness, least noise, synchronization, and overall preferences. We show the music generated by our approach significantly outperforms several strong baselines. In summary, our work makes the following contributions:

- We present a new model to generate synchronized and expressive music from videos.
- This paper proposes body keypoint and MIDI as an intermediate representation for transferring knowledge across two modalities, and we empirically demonstrate that such representations are key to success.
- Our system outperforms previous state-of-the-art systems on music generation from videos by a large margin.
- We additionally demonstrate that MIDI musical representations facilitate new applications on generating different styles of music, which seems impossible before.

2 Related Work

2.1 Audio-Visual Learning

Cross-modal learning from vision and audio has attracted increasing interest in recent years [44,2,4,57,32]. The natural synchronization between vision and sound has been leveraged for learning diverse tasks. Given unlabeled training videos, Owens *et al.* [44] used sound clusters as supervision to learn visual feature representation, and Aytar *et al.* [4] utilized the scene to learn the audio representations. Follow up works [2,33] further investigated to jointly learn the visual and audio representation using a visual-audio correspondence task. Instead of learning feature representations, recent works have also explored to localize sound source in images or videos [29,26,3,48,64], biometric matching [39], visual-guided sound source separation [64,15,19,60], auditory vehicle tracking [18], multi-modal action recognition [36,35,21], audio inpainting [66], emotion recognition [1], audio-visual event localization [56], multi-modal physical scene understanding [16], audio-visual co-segmentation [47], aerial scene recognition [27] and audio-visual embodied navigation [17].

2.2 Motion and Sound

Several works have demonstrated the strong correlations between sound and motion. For example, the associations between speech and facial movements can be used for facial animations from speech [31,55], generating high-quality talking face from audio [54,30], separate mixed speech signals of multiple speakers [14,42], and even lip-reading from raw videos [12]. Zhao *et al.* [63] and Zhou *et al.* [68] have demonstrated to use optical flow like motion representations to improve the quality of visual sound separations and sound generations. There are also some recent works to explore the correlations between body motion and sound by predicting gestures from speech [22], body dynamics from music [50], or identifying a melody through body language [15]. Different from them, we mainly focus on generating music from videos according to body motions.

2.3 Music Generation

Generating music has been an active research area for decades. As opposed to handcrafted models, a large number of deep neural network models have been proposed for music generation [40,11,28,62,24,59,46,65,8]. For example, MelodyRNN [59] and DeepBach [24] can generate realistic melodies and bach chorales. WaveNet [40] showed very promising results in generating realistic speech and music. Song from PI [11] used a hierarchical RNN model to simultaneously generate melody, drums, and chords, thus leading to a pop song. Huang *et al.* [28] proposed a music transformer model to generate expressive piano music from MIDI event. Hawlhorne *et al.* [25] created a new MAESTRO Dataset to factorize piano music modeling and generation. A detailed survey on deep learning for music generation can be found at [5]. However, there is little work on exploring the problem of generating expressive music from videos.

2.4 Sound Generation from Videos

Back in the 1920s, Jack Foley invented *Foley*, a technique that can create convincing sound effects to movies. Recently, a number of works have explored the ideas of training neural networks to automate Foley. Owens *et al.* [43] investigated the task of predicting the sound emitted by interacting objects with a drumstick. They first used a neural network to predict sound features and then performed an exemplar-based retrieval algorithm instead of directly generating the sound. Chen *et al.* [10] proposed to use the conditional generative adversarial networks for cross-modal generation on lab-collected music performance videos. Zhou *et al.* [68] introduced a SampleRNN-based method to directly predict a generate waveform from an unconstraint video dataset that contains 10 types of sound recorded in the wild. Chen *et al.* [9] proposed a perceptual loss to improve the audio-visual semantic alignment. Chen *et al.* [45] introduced an information bottleneck to generate visually aligned sound. Recent works [20,38,67] also attempt to generate 360/stereo sound from videos. However, these works all use appearances or optical flow for visual representations, and spectrograms or waveform for audio representations. Concurrent to our work, [32,52] also study using MIDI for music transcription and generation.

3 Approach

In this section, we describe our framework of generating music from videos. We first introduce the visual and audio representations used in our system (Section 3.1). Then we present a new Graph−Tansformer model for MIDI events prediction from body pose features (Section 3.2). Finally, we introduce the training objective and inference procedures (Section 3.3). The pipeline of our system is illustrated in Figure 2.

3.1 Visual and Audio Representations

Visual Representations. Existing work on video to sound generation either use the appearances [43,68] or optical flow [68] as the visual representations. Though remarkable results have achieved, they exhibit limited abilities to applications that require the capture of the fine-grained level correlations between motion and sound. Inspired by previous success on associating vision with audio signals through the explicit movement of the human body parts and hand fingers [50,22], we use the human pose features to capture the body motion cues. This is achieved by first detecting the human body and hand keypoints from each video frame and then stacking their 2D coordinates over time as structured visual representations. In practice, we use the open-source OpenPose toolbox [6] to extract the 2D coordinates of human body joints and adopt a pre-trained hand detection model and the OpenPose [6] hand API [51] to predict the coordinates of hand keypoints. In total, we obtain 25 keypoints for the human body parts and 21 keypoints for each hand.
Audio Representations. Choosing the correct audio representations is very important for the success of generating expressive music. We have explored several audio representations and network architectures. For example, we have explored to directly generate raw waveform using RNN [43,68] or predict sound spectrograms using GAN [10]. However, none of these models work well on generating realistic music from videos. These results are not surprising since music is highly compositional and contains many structured events. It is extremely hard for a machine learning model to discover these rules contained in the music.

We choose the Musical Instrument Digital Interface (MIDI) as the audio representations. MIDI is composed of timing information note-on and note-off events. Each event also defines note pitch. There is also additional velocity information contained in note-on events that indicates how hard the note was played. We first use a music transaction software [1] to automatically detect MIDI events from the audio track of the videos. For a 6-second video clip, it typically contains

[1] https://www.lunaverus.com/

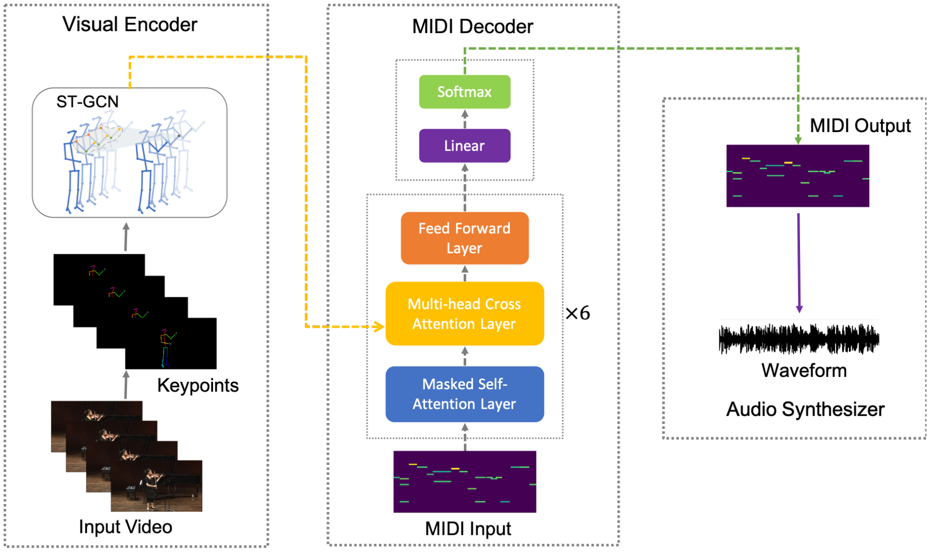

Fig. 2. An overview of our model architecture. It consists of three components: a visual encoder, a MIDI decoder, and an audio synthesizer. The visual encoder takes video frames to extract keypoint coordinates, use GCN to capture the body dynamic and produce a latent representation over time. The MIDI decoder take the video sequence representation to generate a sequence of MIDI event. Finally the MIDI event is converted to the waveform with a standard audio synthesizer.

around 500 MIDI events, although the length might vary for different music. To generate expressive timing information for music modeling, we adopt similar music performance encoding proposed by Oore *et al.* [41], which consists of a vocabulary of 88 note-on events, 88 note-off events, 32 velocity bins and 32 time-shift events. These MIDI events could be easily imported into a standard synthesizer to generate the waveforms of music.

3.2 Body Motions to MIDI Predictions

We build a $Graph-Tansformer$ module to model the correlations between the human body parts and hand movements with the MIDI events. In particular, we first adopt a spatial-temporal graph convolutional network on body keypoint coordinates over time to capture body motions and then feed the encoded pose features to a music transformer decoder to generate a sequence of the MIDI events.

Visual Encoder. Given the 2D keypoints coordinates are extracted from the raw videos, we adopt a Graph CNN to explicitly model the spatial-temporal relationships among different keypoints on the body and hands. Similar to [61], we first represent human skeleton sequence as an undirected spatial-temporal graph $G = (V, E)$, where the node $v_i \in \{V\}$ corresponds to a key point of the human body and edges reflect the natural connectivity of body keypoints.

The input for each node are 2D coordinates of a detected human body keypoint over time T. To model the spatial-temporal body dynamics, we first perform a spatial GCN to encode the pose features at each frame independently and then a standard temporal convolution is applied to the resulting tensor to aggregate the temporal cues. The encoded pose feature P is defined as:

$$P = AXW_SW_T, \tag{1}$$

where $X \in R^{V \times T \times C_n}$ is the input features; V and C_n represent the number of keypoints and the feature dimension for each input node, respectively; $A \in R^{V \times V}$ is the row-normalized adjacency matrix of the graph; W_S and W_T are the weight matrices of spatial graph convolution and temporal convolution. The adjacency matrix is defined based on the joint connections of the body and fingers. Through GCN, we update the keypoint node features over time. Finally, we aggregate the node features to arrive an encoded pose feature $P \in R^{T_v \times C_v}$, where T_v and C_v indicate the number of temporal dimension and feature channels.

MIDI Decoder. Since the music signals are represented as a sequence of MIDI events, we consider music generation from body motions as a sequence prediction problem. To this end, we use the decoder portion of the transformer model [28], which has demonstrated strong capabilities to capture the long-term structure in sequence predictions.

The transformer model [58] is an encoder-decoder based autoregressive generative model, which is originally designed for machine translation applications. We adapt this model to our motion to MIDI translation problem. Specifically, given a visual representation $P \in R^{T_v \times C_v}$, the decoder of transformers is responsible for predicting a sequence of MIDI events $M \in R^{T_m \times L}$, where T_m and L denote a total number of MIDI events contained in a video clip and the vocabulary size of MIDI events. At each time step, the decoder takes the previously generated feature encoding over the MIDI event vocabulary and visual pose features as input and predicts the next MIDI event.

The core mechanism used in the Transformer is the *scale dot-product self-attention* module. This self-attention layer first transforms a sequence of vectors into query Q, key K, and values V, and then output a weighted sum of value$V$, where the weight is calculated by dot products of the key K and query Q. Mathematical:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^t}{\sqrt{D_k}})V \tag{2}$$

Instead of performing single attention function, *multi-head attention* is a common used strategy, which allows the model to integrate information from different independent representations.

Different from the vanilla Transformer model, which only uses positional sinusoids to represent timing information, we adopt relative position representations [49] to allow attention to explicitly know the distance between two tokens in a sequence. This is s critically important for modeling music application [28], since music has rich polyphonic sound, and the relative difference matter significantly to timing and pitch. To address this issue, we follow the strategy used in [28] to jointly learn an ordered relative position embedding R for each possible pairwise distance among pairs of query and key on each head as:

$$\text{Relative Attention}(Q, K, V) = \text{softmax}(\frac{QK^t + R}{\sqrt{D_k}})V \tag{3}$$

For our MIDI decoder, we first use a masked self-attention module with relative position embedding to encode input MIDI events, where queries, keys, and values are all from the same feature encoding and only depend only on the current and previous positions to maintain the auto-aggressive property. The output of masked self-attention module $M \in R^{T_m \times C_m}$ and pose features $P \in R^{T_v \times C_v}$ are then passed into a multi-head attention module, computed as:

$$\text{Cross Attention}(M, P) = \text{softmax}(\frac{MW^M(PW^P)^t}{\sqrt{D_k}})(PW^V) \tag{4}$$

The pointwise feed-forward layer takes the input from cross multi-head attention layer, and further transforms it through two fully connected layers with ReLU activation as:

$$\text{Feed Foward} = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

The output of feed-forward layers is passed into a softmax layer to produce probability distributions of the next token over the vocabulary.

Music Synthesizer. MIDI can get rendered into a music waveform using a standard synthesizer. It is also possible to train a neural synthesizer [25] for the audio rendering. We leave it to future work.

3.3 Training and Inference

Our graph−transformer model is fully differentiable, thus can be trained in an end-to-end fashion. During training, we take input 2D coordinates of the human skeleton and predict a sequence of MIDI events. At each generation process, the MIDI decoder takes visual encoder features over time, previous and current MIDI event tokens as input and predict the next MIDI event. The training objective is to minimize the cross-entropy loss given a source target sequence of MIDI events. Given the testing video, our model generates MIDI events by performing a beam-search with a beam size of 5.

4 Experiments

In this section, we introduce the experimental setup, comparisons with state-of-the-arts, and ablation studies on each model component.

4.1 Experimental Setup

Datasets: We conduct experiments on three video datasets of music performances, namely URMP [34], AtinPiano and MUSIC [64]. URMP is a high-quality multi-instrument video dataset recorded in a studio and provides MIDI file for each recorded video. AtinPiano is a YouTube channel, including piano video recordings with camera looking down on the keyboard and hands. We use [53] to extract the hands from the videos. MUSIC is an untrimmed video dataset downloaded by querying keywords from Youtube. It contains around 1000 music performance videos belonging to 11 categories. In the paper, we MUSIC and AtinPiano datasets for comparisons with state-of-the-arts, and URMP dataset for ablated study.

Implementation Details: We implement our framework using Pytorch. We first extract the coordinates of body and hand keypoints for each frame using OpenPose [6]. Our GCN encoder consists of 10-layers with residual connections. When training the graph CNN network, we use a batch normalization layer for input 2D coordinates to keep the scale of the input the same. During training, we also perform random affine transformations on the skeleton sequences of all frames as data augmentationto avoid overfitting. The MIDI decoder consists of 6 identical decoder blocks. For each block, the dimension of the attention layer and feed-forward layer are set to 512 and 1024, respectively. The number of attention head is set to 8. For the audio data pre-processing, we first use the toolbox to extract MIDI events from audio recordings. During training, we randomly take a 6-second video clip from the dataset. A software synthesizer[2] is applied to obtain the final generated music waveforms.

We train our model using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. We schedule the learning rate during training with a warm-up period. Specifically, the learning rate is linearly increased to 0.0007 for the first 4000 training steps, and then decreased proportionally to the inverse square root of the step number.

4.2 Comparisons with State-of-the-arts

We use 9 instruments from MUSIC and AtinPiano dataset to compare against previous systems, including accordion, bass, bassoon, cello, guitar, piano, tuba, ukulele, and violin.

Baseline: we consider 3 state-of-the-art systems to compare against. For fair comparisons, we use the same pose feature representations extracted from GCN for all these baselines.

- **SampleRNN:** We follow the sequence-to-sequence pipeline used in [68]. Specifically, we used the pose features to initial the coarsest tier RNN of the SampleRNN, which serves as a sound generator.
- **WaveNet:** We take a conditional WaveNet as our sound generator. To consider the video content during sound generation, we use pose features as the local condition. All other settings are the same as [40].

[2] https://github.com/FluidSynth/fluidsynth

Table 1. Human evaluation on model comparisons.

Method	GAN-based	SampleRNN	WaveNet	Ours
Accordion	12%	16%	8%	**64%**
Bass	8%	8%	12%	**72%**
Bassoon	10%	14%	6%	**70%**
Cello	8%	14%	12%	**66%**
Guitar	12%	26%	6%	**56%**
Piano	14%	10%	10%	**66%**
Tuba	8%	20%	10%	**62%**
Ukulele	10%	14%	14%	**62%**
Violin	10%	18%	14%	**58%**

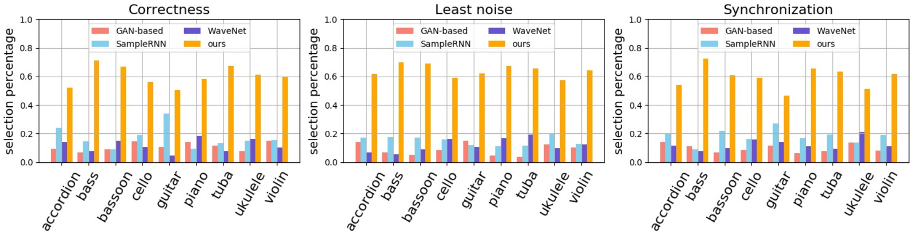

Fig. 3. Human evaluation results of forced-choice experiments in term of correctness, least noise, and synchronization.

- **GAN-based Model:** We adopt the framework proposed in [10]. Specifically, taking the pose feature as input, an encoder-decoder is adopted to generate a spectrogram. A discriminator is designed to determine whether the spectrogram is real or fake, conditional on the input pose feature. We transform the spectrogram to waveform by iSTFT.

Qualitative Evaluation with Human Study: Similar to the task of image or video generation, the quality of the generated sound can be very subjective. For instance, it could be possible to generate music not similar to the ground truth by applying distance metrics, but still sound like a reasonable match to the video content. We carried out a listening study to qualitatively compare the perceived quality of generated music on the Amazon Mechanical Turk (AMT).

We first conduct a forced-choice evaluation [68] to directly compare the proposed method against three baselines. Specifically, we show the four videos with the same video content but different sounds synthesized from our proposed method and three baselines to AMT turkers. They are instructed to choose the best video-sound pair. We use four criteria proposed in [68]:

- **Correctness:** which music recording is most relevant to video content;
- **Least noise:** which music recording has least noise;
- **Synchronization:** which music recording temporally aligns with the video content best;

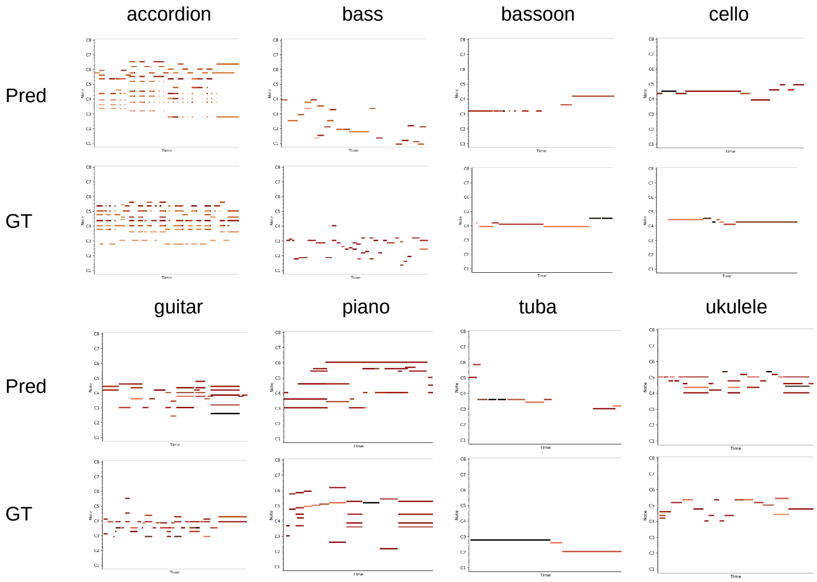

Fig. 4. Visualization of MIDI prediction results.

Table 2. Human evaluation on real-fake. Success mean the percentage of generate sound that were considered real by worker.

Method	Sample RNN	WaveNet	GAN	Ours	Oracle
Success	12%	8%	12%	**38%**	50%

– **Overall:** which sound they prefer to listen to overall.

For each instrument category, we choose 50 video clips for evaluation. There are 450 video clips in total. Every question for each test video has been labeled by three independent turkers and the results are reported by majority voting. Table 1 shows overall preference rate for all categories. We find that our method beat the baseline systems for all the instrument categories. To in-depth understand the benefit of our approach, we further analyze the correctness, least noise and synchronization in Figure 3. We can observe that our approach also consistently outperform baseline systems across all the evaluation criteria by a large-margin. These results further support our claims that the MIDI event representations help improve sound quality, semantic alignment, and temporal synchronization for music generation from videos.

Visualizations: In figure 4, we first show the MIDI prediction and ground truth. We can observe that our predicted MIDI event are reasonable similar to the ground truth. We also visualize the sound spectrogram generated by different approaches in Figure 5. We can find that our model does generate more structured harmonic components than other baselines.

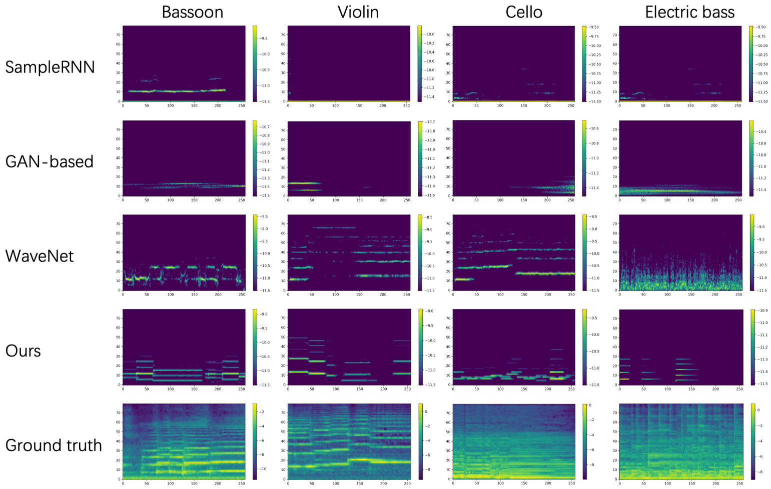

Fig. 5. Qualitative comparison results on sound spectrogram generated by different methods. We report the fraction of generated images

Qualitative Evaluation with real or fake study In this task, we would like to assess whether the generated audios can fool people into thinking that they are real. We provide two videos with real (originally belonging to this video) and fake (generated by computers) audio to the AMT turkers. The turkers are required to choose the video that they think is real. The criteria for being fake can be bad synchronization, artifacts, or containing noise. We evaluated the ranking of 3 AMT turkers, each was given 100 video pairs. To be noted, an oracle score of 50% would indicate perfect confusions between real and fake. The results in Table 2 demonstrate that, our generated music was hard to distinguish from the real audio recordings than other systems.

Quantitative Evaluation with Automatic Metrics We adopt the Number of Statistically-Different Bins (NDB) [13] as automatic metrics to evaluate the diversity of generated sound. Specifically, we first transform sound to log-spectrogram. Then, we cluster the spectrogram in the training set into $k = 50$ Voronoi cells by k-means algorithm. Each generated sound in the testing set is assigned to the nearest cell. NDB indicated the number of cells in which the training samples are significantly different from the number of testing examples. Except for the baselines mentioned above, we also compare with VIG baseline [8] which uses perception loss. The results are listed in Table 3. Our method achieve significantly lower NDB, demonstrating that we can generate more diverse sound.

4.3 Ablated Study

In this section, we perform in-depth ablation studies to assess the impact of each component of our model. We use 5 instruments from URMP dataset for

Table 3. Automatic metrics for different models. For NDB, lower is better.

Metric	VIG	WaveNet	GAN	SampleRNN	Ours
NDB	33	32	25	30	**20**

Table 4. Ablated study on visual representation in term of NELL loss on MIDI prediction. **Lower** number means **Better** results.

Method	violin	viola	cello	trumpet	flute
RGB image	1.586	3.772	3.077	2.748	2.219
Optical Flow	1.581	3.859	3.178	3.013	2.046
Skeleton (Ours)	**1.558**	**3.603**	**2.981**	**2.512**	**1.995**

quantitative evaluations, including violin, viola, cello, trumpet, and flute. Since this dataset provides the ground-truth MIDI file, we use negative log-likelihood (NLL) of MIDI event prediction on the validation set as an evaluation metric.

The effectiveness of Body Motions. In our system, we exploit explicit body motions through keypoint-based structure representations to guide music generation. To further understand the ability of these representations, we conduct an ablated study by replacing keypoint-based structure representation with RGB image and optical flow representation. For these two baselines, we extract the features using I3D network [7] pre-trained on Kinetics. As results shown in Table 4, keypoint-based representation achieve better MIDI prediction accuracy than other options. We hope our findings could inspire more works using the keypoints-based visual representations to solve more challenging audio-visual scene analysis tasks.

The effectiveness of Music Transformers. We adopt a music transformers framework for the sequence predictions. To verify its efficacy, we replace this module with GRU, and keep the other parts of the pipeline the same. The comparison results are shown in Table 5. We can find that the music transformer module improves NEL loss over the GRU baseline. These results demonstrated the benefits of our designed choices using the transformer to capture the long-term dependencies in music.

4.4 Music Editing with MIDI

Since MIDI representation is fully interpretable and transparent, we can easily perform the music editing by manipulating the MIDI file. To demonstrate the flexibility of MIDI representations, we show an example in Figure 6. Here, we simply manipulate the key of the predicted MIDI, showing its capability to generate music with different styles. These result validate that the MIDI events are flexible and interpretable, thus enabling new applications on controllable music generation, which seem impossible for previous systems which use the waveform or spectrogram as the audio representations.

Table 5. Ablated study on sequence prediction model in term of NELL loss on MIDI prediction. **Lower** number means **Better** results.

Method	violin	viola	cello	trumpet	flute
GRU	1.631	3.747	3.06	2.631	2.101
Transformers w/o hands (Ours)	1.565	3.632	3.014	2.805	2.259
Transformers w hands (Ours)	**1.558**	**3.603**	**2.981**	**2.512**	**1.995**

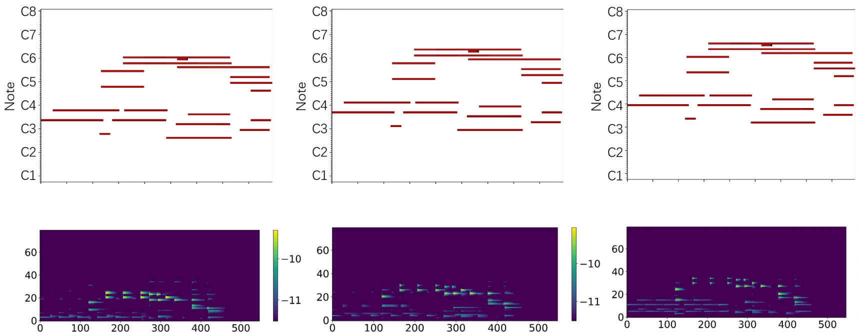

Fig. 6. Music key editing results by manipulating MIDI.

5 Conclusions and Future Work

In this paper, we introduce a *foley music* system to generate expressive music from videos. Our model takes video as input, detects human skeletons, recognizes interactions with musical instruments over time and then predicts the corresponding MIDI files. We evaluated the quality of our approach using human evaluation, showing that the performance of our algorithm was significantly better than baselines. The results demonstrated that the correlations between visual and music signals can be well established through body keypoints and MIDI representations. We additionally show our framework can be easily extended to generate music with different styles through the MIDI representations.

In the future, we plan to train a WaveNet [40] like neural music synthesizer that can generate waveform from MIDI events. Therefore, the whole system can be end-to-end trainable. We envision that our work will open up future research on studying the connections between video and music using intermediate body keypoints and MIDI event representations.

Acknowledgement. This work is supported by ONR MURI N00014-16-1-2007, the Center for Brain, Minds, and Machines (CBMM, NSF STC award CCF-1231216), and IBM Research.

References

1. Albanie, S., Nagrani, A., Vedaldi, A., Zisserman, A.: Emotion recognition in speech using cross-modal transfer in the wild. ACM Multimedia (2018) 3
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 609–617. IEEE (2017) 3
3. Arandjelović, R., Zisserman, A.: Objects that sound. arXiv preprint arXiv:1712.06651 (2017) 3
4. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. pp. 892–900 (2016) 3
5. Briot, J.P., Hadjeres, G., Pachet, F.D.: Deep learning techniques for music generation–a survey. arXiv preprint arXiv:1709.01620 (2017) 4
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018) 5, 9
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 4724–4733. IEEE (2017) 13
8. Chen, K., Zhang, C., Fang, C., Wang, Z., Bui, T., Nevatia, R.: Visually indicated sound generation by perceptually optimized classification. In: ECCV. vol. 11134, pp. 560–574 (2018) 4, 12
9. Chen, K., Zhang, C., Fang, C., Wang, Z., Bui, T., Nevatia, R.: Visually indicated sound generation by perceptually optimized classification. In: The European Conference on Computer Vision. pp. 560–574 (2018) 4
10. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: ACM Multimedia 2017. pp. 349–357 (2017) 4, 5, 10
11. Chu, H., Urtasun, R., Fidler, S.: Song from pi: A musically plausible network for pop music generation. ICLR (2017) 1, 4
12. Chung, J.S., Senior, A.W., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: CVPR. pp. 3444–3453 (2017) 4
13. Engel, J.H., Agrawal, K.K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A.: GANSynth: Adversarial neural audio synthesis. In: ICLR (2019) 12
14. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) **37**(4), 112 (2018) 4
15. Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., Torralba, A.: Music gesture for visual sound separation. In: CVPR. pp. 10478–10487 (2020) 3, 4
16. Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., et al.: Threedworld: A platform for interactive multi-modal physical simulation. arXiv preprint arXiv:2007.04954 (2020) 3
17. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. ICRA (2020) 3
18. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: ICCV. pp. 7053–7062 (2019) 3
19. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV. pp. 35–53 (2018) 3

20. Gao, R., Grauman, K.: 2.5 d visual sound. arXiv preprint arXiv:1812.04204 (2018) 4
21. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: CVPR. pp. 10457–10467 (2020) 3
22. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: CVPR. pp. 3497–3506 (2019) 4, 5
23. Godøy, R.I., Leman, M.: Musical gestures: Sound, movement, and meaning. Routledge (2010) 1
24. Hadjeres, G., Pachet, F., Nielsen, F.: Deepbach: a steerable model for bach chorales generation. In: ICML. pp. 1362–1371 (2017) 4
25. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.Z.A., Dieleman, S., Elsen, E., Engel, J., Eck, D.: Enabling factorized piano music modeling and generation with the maestro dataset. ICLR (2019) 4, 8
26. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: Solla, S.A., Leen, T.K., Müller, K. (eds.) Advances in Neural Information Processing Systems 12, pp. 813–819 (2000) 3
27. Hu, D., Li, X., Mou, L., Jin, P., Chen, D., Jing, L., Zhu, X., Dou, D.: Cross-task transfer for multimodal aerial scene recognition. ECCV (2020) 3
28. Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A.M., Hoffman, M.D., Dinculescu, M., Eck, D.: Music transformer: Generating music with long-term structure (2018) 1, 4, 7, 8
29. Izadinia, H., Saleemi, I., Shah, M.: Multimodal analysis for identification and segmentation of moving-sounding objects. IEEE Transactions on Multimedia **15**(2), 378–390 (2013) 3
30. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. International Journal of Computer Vision pp. 1–13 (2019) 4
31. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) **36**(4), 94 (2017) 4
32. Koepke, A.S., Wiles, O., Moses, Y., Zisserman, A.: Sight to sound: An end-to-end approach for visual piano transcription. In: ICASSP. pp. 1838–1842 (2020) 3, 5
33. Korbar, B., Tran, D., Torresani, L.: Co-training of audio and video representations from self-supervised temporal synchronization. arXiv preprint arXiv:1807.00230 (2018) 3
34. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Transactions on Multimedia **21**(2), 522–535 (2018) 9
35. Long, X., Gan, C., De Melo, G., Liu, X., Li, Y., Li, F., Wen, S.: Multimodal keyless attention fusion for video classification. In: AAAI (2018) 3
36. Long, X., Gan, C., de Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: CVPR (2018) 3
37. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746–748 (1976) 1
38. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: NIPS (2018) 4
39. Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces: Cross-modal biometric matching. arXiv preprint arXiv:1804.00326 (2018) 3
40. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. ICLR (2017) 1, 4, 9, 14

41. Oore, S., Simon, I., Dieleman, S., Eck, D., Simonyan, K.: This time with feeling: Learning expressive musical performance. Neural Computing and Applications pp. 1–13 (2018) 6
42. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. ECCV (2018) 4
43. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2405–2413 (2016) 4, 5
44. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: European Conference on Computer Vision. pp. 801–816. Springer (2016) 3
45. Peihao, C., Yang, Z., Mingkui, T., Hongdong, X., Deng, H., Chuang, G.: Generating visually aligned sound from videos. IEEE Transactions on Image Processing (October 2020) 4
46. Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D.: A hierarchical latent vector model for learning long-term structure in music. arXiv preprint arXiv:1803.05428 (2018) 1, 4
47. Rouditchenko, A., Zhao, H., Gan, C., McDermott, J., Torralba, A.: Self-supervised audio-visual co-segmentation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2357–2361. IEEE (2019) 3
48. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. arXiv preprint arXiv:1803.03849 (2018) 3
49. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018) 7
50. Shlizerman, E., Dery, L., Schoen, H., Kemelmacher-Shlizerman, I.: Audio to body dynamics. In: CVPR. pp. 7574–7583 (2018) 4, 5
51. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017) 5
52. Su, K., Liu, X., Shlizerman, E.: Audeo: Audio generation for a silent performance video. arXiv preprint arXiv:2006.14348 (2020) 5
53. Submission, A.: At your fingertips: Automatic piano fingering detection. In: ICLR (2020) 9
54. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) **36**(4), 95 (2017) 4
55. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Transactions on Graphics (TOG) **36**(4), 93 (2017) 4
56. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV. pp. 247–263 (2018) 3
57. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. ECCV (2020) 3
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017) 7
59. Waite, E., et al.: Generating long-term structure in songs and stories. Web blog post. Magenta **15** (2016) 4
60. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 882–891 (2019) 3
61. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018) 6

62. Yang, L.C., Chou, S.Y., Yang, Y.H.: Midinet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847 (2017) 1, 4
63. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. ICCV (2019) 4
64. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: The European Conference on Computer Vision (ECCV) (September 2018) 3, 9
65. Zhao, K., Li, S., Cai, J., Wang, H., Wang, J.: An emotional symbolic music generation system based on lstm networks. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). pp. 2039–2043 (2019) 4
66. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: ICCV. pp. 283–292 (2019) 3
67. Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z.: Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In: ECCV (2020) 4
68. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. CVPR (2018) 4, 5, 9, 10