Music Gesture for Visual Sound Separation

Citation: Gan, Chuang et al. "Music Gesture for Visual Sound Separation." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020, Seattle, Washingston, Institute of Electrical and Electronics Engineers, August 2020. © 2020 IEEE

As Published: http://dx.doi.org/10.1109/cvpr42600.2020.01049

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: https://hdl.handle.net/1721.1/130393

Version: Original manuscript: author's manuscript prior to formal peer review

Music Gesture for Visual Sound Separation

Chuang Gan[1,2], Deng Huang[2], Hang Zhao[1], Joshua B. Tenenbaum[1], Antonio Torralba[1]
[1] MIT, [2] MIT-IBM Watson AI Lab

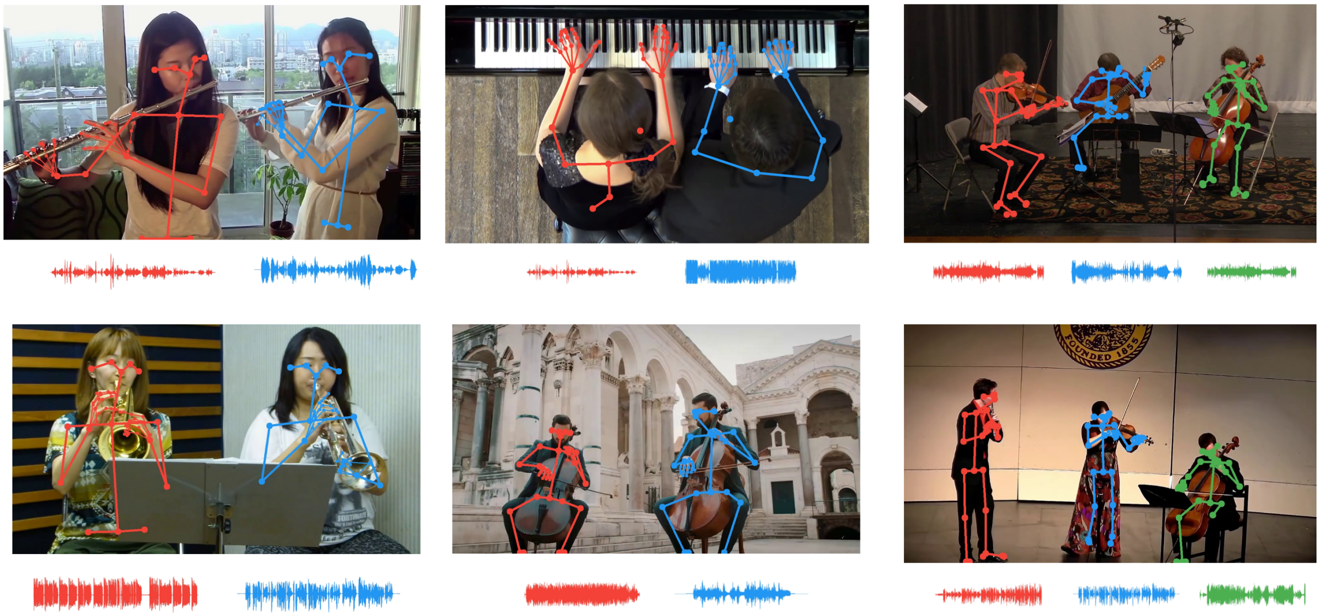

Figure 1: We propose to leverage explicit body dynamics motion cues for visual sound separation in music performances. We show that our new model can perform well on both heterogeneous and homogeneous music separation tasks.

Abstract

Recent deep learning approaches have achieved impressive performance on visual sound separation tasks. However, these approaches are mostly built on appearance and optical flow like motion feature representations, which exhibit limited abilities to find the correlations between audio signals and visual points, especially when separating multiple instruments of the same types, such as multiple violins in a scene. To address this, we propose "Music Gesture," a keypoint-based structured representation to explicitly model the body and finger movements of musicians when they perform music. We first adopt a context-aware graph network to integrate visual semantic context with body dynamics, and then apply an audio-visual fusion model to associate body movements with the corresponding audio signals. Experimental results on three music performance datasets show: 1) strong improvements upon benchmark metrics for hetero-musical separation tasks (i.e. different instruments); 2) new ability for effective homo-musical separation for piano, flute, and trumpet duets, which to our best knowledge has never been achieved with alternative methods. Project page: `http://music-gesture.csail.mit.edu`.

1. Introduction

Music performance is a profoundly physical activity. The interactions between body and the instrument in nuanced gestures produce unique sounds [21]. When performing, pianists may strike the keys at a lower register or "tickle the ivory" up high; Violin players may move vigorously through a progression while another player sways gently with a melodic base; Flautists press a combination of keys to produce a specific note. As humans, we have the remarkable ability to distinguish different sounds from one another, and associate the sound we hear with the corresponding visual perception from the musician's bodily gestures.

Inspired by this human ability, we propose "Music Gesture" (shown in Figure 1), a structured keypoint-based visual representations to makes use of the body motion cues for sound source separation. Our model is built on the mix-and-separate self-supervised training procedure initially proposed by Zhao *et al*. [57]. Instead of purely relying on visual semantic cues [57, 17, 19, 53] or low-level optical-flow like motion representations [56], we consider to exploit the explicit human body and hand movements in the videos. To achieve this goal, we design a new framework, which consists of a video analysis network and an audio-visual separation network. The video analysis network extracts body dynamics and semantic context of musical instruments from video frames. The audio-visual separation network is then responsible for separating each sound source based on the visual context. In order to better leverage the body dynamic motions for sound separations, we further design a new audio-visual fusion module in the middle of the audio-visual separation network to adjust sound features conditioned on visual features.

We demonstrate the effectiveness of our model on three musical instrument datasets, URMP [31], MUSIC [57] and AtinPiano [36]. Experimental results show that by explicitly modeling the body dynamics through the keypoint-based structured visual representations, our approach performs favorably against state-of-the-art methods on both hetero-musical and homo-musical separation task. In summary, our work makes the following contributions:

- We pave a new research direction on exploiting body dynamic motions with structured keypoint-based video representations to guide the sound source separation.
- We propose a novel audio-visual fusion module to associate human body motion cues with the sound signals.
- Our system outperforms previous state-of-the-arts approaches on hetero-musical separation tasks by a large margin.
- We show that the keypoint-based structured representations open up new opportunities to solve harder homo-musical separation problem for piano, flute, and trumpet duets.

2. Related Work

Sound separation. Sound separation is a central problem in the audio signal processing area [34, 22], while the classic solutions for it are based on Non-negative Matrix Factorization (NMF) [52, 11, 48]. These are not very effective as they rely on low-level correlations in the signals. Deep learning based methods are taking over in the recent years. Simpson *et al*. [47] and Chandna *et al*. [9] proposed CNN models to predict time-frequency masks for music source separation and enhancement. Another challenging problem in speech separation is identity permutation: a spectrogram classification model could not deal with the case with arbitrary number of speakers talking simultaneously. To solve this problem, Hershey *et al*. [24] proposed Deep Clustering and Yu *et al*. [55] proposed a speaker-independent training framework.

Visual sound separation. Our work falls into the category of visual sound separation. Early works [5] leveraged the tight associations between audio and visual onset signal to perform audio-visual sound attribution. Recently, Zhao *et al*. [57] proposed a framework that learns from unlabeled videos to separate and localize sounds with the help of visual semantics cues. Gao *et al*. [17] combined deep networks with NMF for sound separation. Ephrat *et al*. [12] and Owens *et al*. [38] proposed to used vision to improve the quality of speech separation. Xu *et al*. [53] and Gao *et al*. [19] further improved the models with recursive models and co-separation loss. Those works all demonstrated how semantic appearances could help with sound separation. However, these methods have limited capabilities to capture the motion cues, thus restricts their applicability to solve harder sound source separation problems.

Most recently, Zhao *et al*. [56] proposed to leverage temporal motion information to improve the vision sound separation. However, this algorithm has not yet seen wide applicability to sound separation on real mixtures. This is primarily due to the trajectory and optical flow like motion features they used are still limited to model the human-object interactions, thus can not provide strong visual conditions for the sound separation. Our work overcomes these limitations in that we study the explicit body movement cues using structured keypoint-based structured representations for audio-visual learning, which has never been explored in the audio-visual sound separation tasks.

Audio-visual learning. With the emergence of deep neural networks, bridging signals of different modalities becomes easier. A series of works have been published in the past few years on audio-visual learning. By learning audio model and image model jointly or separately by distillation, good audio/visual representations can be achieved [39, 4, 2, 33, 32, 14, 30]. Another interesting problem is sounding object localization, where the goal is to associate sounds in the visual input spatially [26, 25, 3, 44, 57]. Some other interesting directions include biometric matching [37], sound generation for videos [58], auditory vehicle tracking [16], emotion recognition [1], audio-visual co-segmentation [43], audio-visual navigation [15], and 360/stereo sound from videos [18, 35].

Audio and body dynamics. There are numerous works to explore the associations between speech and facial movements [7, 6]. Multi-model signals extracted from face

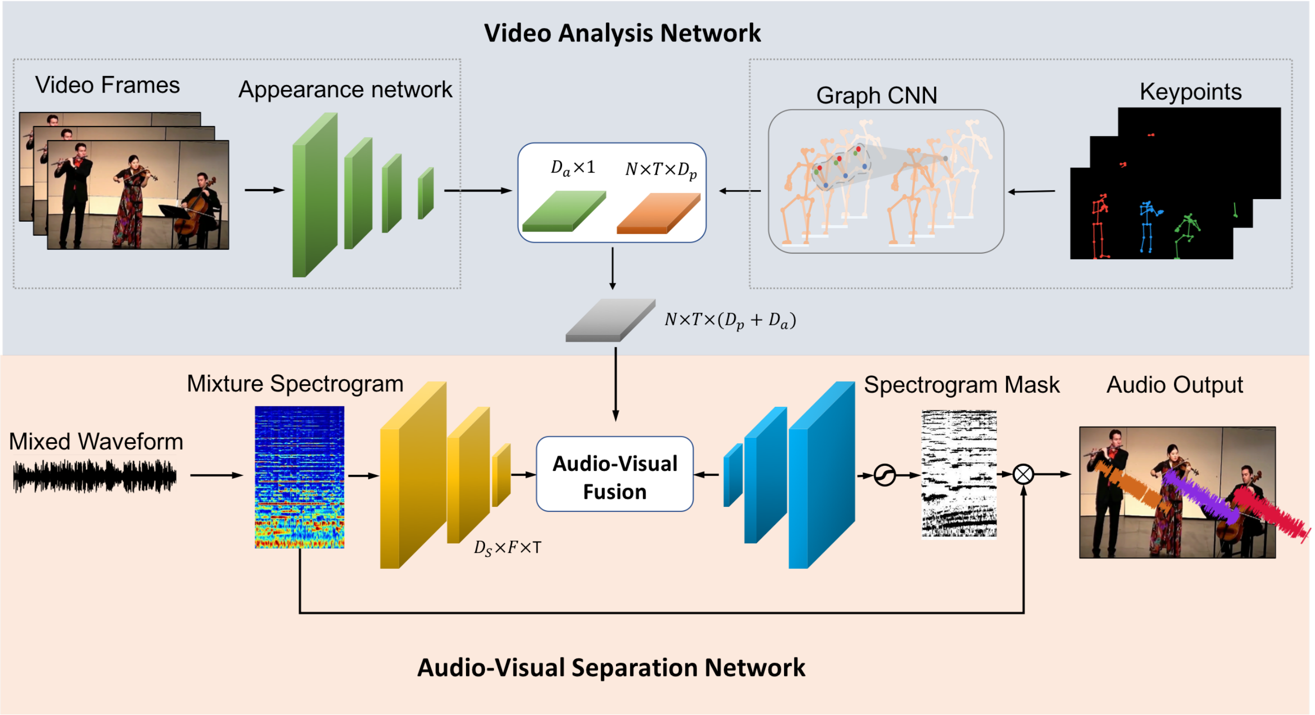

Figure 2: **An overview of our model architecture.** It consists of two components: a video analysis network and a visual-audio separation network. The video analysis network first takes video frames to extract global context and keypoint coordinates; Then a GCN is applied to integrate the body dynamic with semantic context, and outputs a latent representation. Finally, an audio-visual separation network separates sources form the mixture audio conditioned on the visual features.

and speech has been used to do facial animations using speech [28, 50], generate high-quality talking face from audio [49, 27] separate mixed speech signals of multiple speakers [12], on/off screen audio source separation[38], and lip reading from raw videos [10]. In contrast, the correlations between body pose with sound were less explored. The most relevant to us are recent works on predicting body dynamics from music [45] and body rhythms from speech [20]. This is the inverse of our goal to separate sound sources using body dynamic cues.

3. Approach

We first formalize the visual sound separation task and summarize our system pipeline in Section 3.1. Then we present the video analysis network for learning structured representation (Section 3.2) and audio-visual separation model (Section 3.3). Finally, we introduce our training objective and inference procedures in Section 3.4.

3.1. Pipeline Overview

Our goal is to associate the body dynamics with the audio signals for sound source separation. We adopt the commonly used "mix-and-separate" self-supervised training procedure introduced in [57]. The main idea of this training procedure is to create synthetic training data by mixing arbitrary sound sources from different video clips. Then the learning objective is to separate each sound from mixtures conditioned on its associated visual context.

Concretely, our framework consists of two major components: a video analysis network and a audio-visual separation network (see Figure 2). During training, we randomly select N video clips with paired video frames and audio signal $\{V_k, S_k\}$, and then mix their audios by linear combinations of the audio inputs to form a synthetic mixture $S_{mix} = \sum_{k=1}^{N} S_k$. Given a video clip V_k, the video analysis network extracts global context and body dynamic features from videos. The audio-visual separation network is then responsible for separating its audio signal S_k from the mixture audio S_{mix} conditioned on the corresponding visual context V_k. To be noted, we trained the neural network in a supervised fashion, but it learned from unlabeled video data. Therefore, we consider the training pipeline as self-supervised learning.

3.2. Video Analysis Network

Our proposed video analysis network integrates keypoint-based structured visual representations, together with global semantic context features.

Visual semantic and keypoint representations. To extract global semantic features from video frames, we use ResNet-50 [23] to extract the features after the last spatial average pooling layer from the first frame of each video clip. Therefore, we obtain a 2048-dimensional context feature vector for each video clip. We also aim to capture the explicit movement of the human body parts and hand fingers through the keypoint representations. To achieve that, we adopt the AlphaPose toolbox [13] to estimate the 2D locations of human body joints. For estimation of hand pose, we first apply a pre-trained hand detection model and then use the OpenPose [8] hand API [46] to estimate the coordinates of hand keypoints. As a result, we extract 18 keypoints for human body and 21 keypoints for each hand. Since the keypoints estimation in videos in the wild is challenging and noisy, we maintain both 2D coordinates (X, Y) and the confidence score of each estimated keypoint.

Context-Aware Graph CNN. Once the visual semantic feature and keypoints are extracted from the raw video, we adopt a context-aware Graph CNN (CT-GCN) to fuse the semantic context of instruments and human body dynamics. This architecture is designed for the non-grid data, suitable for explicitly modeling the spatial-temporal relationships among different keypoints on the body and hands.

The network architecture design is inspired by previous work on action recognition [54] and human shape reconstruction [29]. Similar to [54], we start by constructing a undirected spatial-temporal graph $G = \{V, E\}$ on a human skeleton sequence. In this graph, each node $v_i \in \{V\}$ corresponds to a keypoint of the human body; edges reflect the natural connectivity of body keypoints.

The input features for each node is represented as 2D coordinates and the confidence score of a detected keypoint over time T. To model the spatial-temporal body dynamics, we first apply a Graph Convolution Network to encode the pose at each time step independently. Then, we perform a standard temporal convolution on the resulting tensor to fuse the temporal information. The encoded pose feature f_v is defined as follows:

$$f_v = \hat{A} X W_s W_t, \tag{1}$$

where $X \in R^{N \times T \times D_n}$ is the input features, W_s and W_t are the weight matrices of spatial graph convolution and 2D convolution, and $\hat{A} \in R^{N \times N}$ is the row-normalized adjacency matrix of the graph; N represents the number of keypoints; D_n represents the feature dimension for each input node. Inspired by previous work [54], we define the adjacency matrix based on the joint connections of the body and fingers. The output of the GCNs is updated features of each keypoint node.

To further incorporate the visual semantic cues, we concatenated the visual appearance context features to each node feature as the final output of the video analysis network. The context-aware graph CNN is capable of modeling both semantic context and body dynamics, thus providing strong visual cues to guide sound separations. There could be other model designs options. We leave this to future work.

3.3. Audio-Visual Separation Network

Finally, we have an audio-visual separation network, which takes the spectrogram of mixture audio with visual representation produced by the video analysis network as input, to predict a spectrogram mask and generate the audio signal for the selected video.

Audio Network. We adopt a U-Net style architecture [42], namely an encoder-decoder network with skip connections for the audio network. It consists of 4 dilated convolution layers and 4 dilated up-convolution layers. All dilated convolutions and up-convolutions use 3 × 3 spatial filters with stride 2, dilation 1 and followed by a BatchNorm layer and a Leaky ReLU. The input of the audio network is a 2D time-frequency spectrogram of mixture sound and the output is a same-size binary spectrogram mask. We infuse the visual features into the middle part of the U-Net for guiding the sound separation.

Audio-visual fusion. To better leverage body dynamic cues to guide the sound separation, we adopt a self-attention [51] based cross-modal early fusion module to capture the correlations between body movement with the sound signals. As shown in Figure 3 , the fused feature z_t at each time step t is defined as follows:

$$h_t = Softmax(f_s^t \cdot {f_v^t}^T) f_v^t + f_s^t, \tag{2}$$

$$z_t = MLP(h_t) + h_t, \tag{3}$$

where $f_s^t \in R^{F \times D_s}$ and $f_v^t \in R^{N \times D_v}$ represents visual and sound features at time step t. F, D_v, and D_s denote the frequency bases of the sound spectrogram, the dimensions of

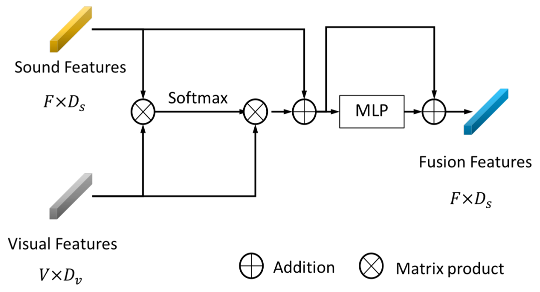

Figure 3: Audio-visual fusion module of the model in Figure 2.

visual features, and the dimension of sound features, respectively. The softmax computation is along the dimension of visual feature channels. The visual feature is then weighted by attention matrix and concatenated with the sound feature. We further add a multi-layer perceptron (MPL) with residual connection to produce the output features. The MLP is implemented with two fully-connected layers with a ReLU activation function. This attention mechanism enforces the model to focus more on the discriminative body keypoints, and associate them with the corresponding sound components on the spectrogram.

3.4. Training and Inferences

The learning objective of our model is to estimate a binary mask M_k. The ground truth mask of k-th video is calculated whether the target sound is the dominant component in the input mixed sound on magnitude spectrogram S, *i.e.*,

$$M_k(t,f) = [\![S_k(t,f) \geq S_{mix}(t,f)]\!], \quad \forall k = (1, ..., N), \tag{4}$$

where (t, f) represents the time-frequency coordinates in the sound spectrogram. The network is trained by minimizing the per-pixel sigmoid cross entropy loss between the estimated masks and the ground-truth binary masks. Then the predicted mask is thresholded and multiplied with the input complex STFT coefficients to get a predicted sound spectrogram. Finally, we apply an inverse short-time Fourier Tranform (iSTFT) with the same transformation parameters on the predicted spectrogram to reconstruct the waveform of separated sound.

During testing, our model takes a single realistic multi-source video to perform sound source separation. We first localize human in the video frames. For each detected person, we use the video analysis network to extract visual feature to isolate the portion of the sound belonging to this musician from the mixed audio.

4. Experiments

In this section, we discuss our experiments, implementation details, comparisons and evaluations.

4.1. Dataset

We perform experiments on three video music performance datasets, namely MUSIC-21[56], URMP [31] and AtinPiano [36]. MUSIC-21 is an untrimmed video dataset crawled by keyword query from Youtube. It contains music performances belonging to 21 categories. This dataset is relatively clean and collected for the purpose of training and evaluating visual sound source separation models. URMP [31] is a high quality multi-instrument video dataset recorded in studio and provides ground truth labels for each sound source. AtinPiano [36] is a dataset where the piano video recordings are filmed in a way that camera is looking down on the keyboard and hands.

4.2. Hetero-musical Separation

We first evaluate the model performance in the task of separating sounds from different kinds of instruments on the MUSIC dataset.

Baseline and evaluation metrics We consider 5 state-of-the-art systems to compare against.

- **NMF** [52] is a well established pipeline for audio-only source separation based on matrix factorization;
- **Deep Separation** [9] is a CNN-based audio-only source separation approach;
- **MIML** [17] is a model that combines NMF decomposition and multi-instance multi-label learning;
- **Sound of Pixels** [57] is a pioneering work that uses vision for sound source separations;
- **Co-separation** [19] devices a new model that incorporates an object-level co-separation loss into the mix-and-separate framework [57];
- **Sound of Motions** [56] is a recently proposed self-supervised model which leverages trajectory motion cues.

We adopt the blind separation metrics, including signal-to-distortion ratio (SDR), and signal-to-interference ratio (SIR) to quantitatively compare the quality of the sound separation. The results reported in this paper were obtained by using the open-source `mir_eval` [41] library.

Experimental Setup Following the experiment protocol in Zhao *et al.* [56], we split all videos on MUSCI dataset into a training set and a test set. We train and evaluate our model using mix-2 and mix-3 samples, which contain 2 and 3 sound sources of different instruments in mixtures. Since the real mix video data with multiple sounds on the MUSIC dataset do not have ground-truth labels for quantitative evaluation, we construct a synthetic testing set by mixing solo videos. The result of model performances are reported on a validation set with 256 pairs of sound mixtures, the same as [56]. We also perform a human study on the real mixtures on MUSIC and URMP dataset to measure human's perceptual quality.

Implementation Details We implement our framework using Pytorch. We first extract a global context feature from a video clip using ResNet-50 [23] and the coordinates of body and hand key points for each frame using OpenPose [8] and AlphaPose [13]. Our GCN model consists of 11-layers with residual connections. When training the graph CNN network, we first pass the keypoint coordinates to a batch normalization layer to keep the scale of the input same. During training, we also randomly move the coordinates as data augmentation to avoid overfitting.

Methods	2-Mix		3-Mix	
	SDR	SIR	SDR	SIR
NMF [52]	2.78	6.70	2.01	2.08
Deep Separation [9]	4.75	7.00	-	-
MIML [17]	4.25	6.23	-	-
Sound of Pixels [57]	7.52	13.01	3.65	8.77
Co-Separation [19]	7.64	13.8	3.94	8.93
Sound of Motion [56]	8.31	14.82	4.87	9.48
Our	**10.12**	**15.81**	**5.41**	**11.47**

Table 1: Sound source separation performance ($N = 2, 3$ mixture) on different instruments. Compared to previous approaches, our models with body dynamic motion information perform better in sound separation.

For the audio data pre-processing, we first re-sample the audio to 11KHz. During training, we randomly take a 6-second video clip from the dataset. The audio-visual separation network takes a 6-second mixed audio clip as input, and transforms it into spectrogram by Short Time Fourier Transform (STFT). We set the frame size and hop size as 1022 and 256, respectively. The spectrogram is then fed into a U-Net with 4 dilated convolution and 4 deconvolution layers. The ouput of U-Net is an estimated binary mask. We set a threshold of 0.7 to obtain a binary mask, and then multiply it with the input mixture sound spectrogram. An iSTFT with the same parameters as the STFT is applied to obtain the final separated audio waveforms.

We train our model using SGD optimizer with 0.9 momentum. The audio separation Network and the fusion module use a learning rate of 1e-2; the ST-GCN Network and Appearance Network use a learning rate of 1e-3.

Quantitative Evaluation. Table 1 summarizes the comparison results against state-of-the-art methods on MUSIC. We observe that our method consistently outperforms all baselines in separation accuracy, as captured across metrics. Remarkably, our system outperforms a previous state-of-the-art algorithm [56] by 1.8dB on 2-mix and 0.6dB on 3-mix source separation in term of SDR score. These quantitative results suggest that our model can successfully exploit the explicit body dynamic motions to improve the sound separation quality.

Qualitative evaluation on real mixtures. Our quantitative results demonstrate that our model achieves better results than baselines. However, these metrics are limited in their ability to reflect the actual perceptual quality of the sound separation result on real-world videos. Therefore, we further conduct a subjective human study using real mixture videos from MUSIC and URMP datasets on Amazon Mechanical Turk (AMT).

Specifically, we compare sound separation results of our own model with best baseline system [56][1] The AMT workers are required to compare these two systems and answer the following question: "Which sound separation result is better?." We randomly shuffle the orders of two models to avoid shortcut solutions. Each job is performed independently by 3 AMT workers. Results are shown in Table 2 using majority voting. From this table, we find workers favor our system for both 2-mix and 3-mix sound separation.

Method	2-Mix	3-Mix
Sound of Motions [56]	24%	16%
Ours	76%	84%

Table 2: Human evaluation results for the sound source separation on mixtures of the different instruments.

4.3. Ablated study

In this section, we perform in-depth ablation studies to evaluate the impact of each component of our model.

Keypoint-based representation. The main contribution of our paper is to use explicit body motions through keypoint-based structure representations for source separation. To further understand the ability of these representations, we conduct an ablated study using the keypoint-based structure representation only, without the RGB context features. Interestingly,we can observe that keypoint-based representations alone could also achieve very strong results (see Table 3). We hope our findings could inspire more works using structured keypoint-based representations for the audio-visual scene analysis tasks.

Visual-Audio Fusion Module. We propose a novel attention based audio-visual fusion model. To verify its efficacy, we replace this module with Feature-wise Linear Modulation (FiLM) [40] used in [56]. The comparison results are

[1] The results on real mixture are provided by their authors.

Method	SDR
Ours w/o fusion	9.64
Ours w/o RGB	10.22
Our	10.12

Table 3: Ablated study on SDR metric for mixtures of 2 different instruments .

shown in Table 3. We can find that the proposed audio-visual fusion module brings 0.5dB improvement in term of SDR metric on 2-mix sound source separation.

4.4. Homo-musical Separation

In this section, we conduct experiments on a more challenging task, sound separation when sound is generated by the same instruments.

Experiment Setup We select 5 kinds of musical instruments whose sounds are closely related to body dynamic: trumpet, flute, piano, violin, and cello for evaluation.

Inspired by previous work [56, 38], we also adopt a 2-stage curriculum learning strategy to train the sound seperation model of the same instruments. In particular, we first pre-train the model on multiple instrument separation, then learn to separate the same instrument. We compare our model against SoM [56], since previous appearance based models fail to produce meaningful results in this challenging setting. The results are measured by both automatic SDR scores and human evaluations on AMT.

Results Analysis. Results are shown in Table 4 and Table 5. From these tables, we have three key observations: 1) our proposed model consistently outperforms the SoM system [56] for all five instruments measured by both automatic and human evaluation metrics; 2) The quantitative results on separating violin and cello duets are close (See Table 4). However, we find that the SoM system is quite brittle when testing on the real mixtures. People tend to vote our system more on real mixtures, as shown in Table 5; 3) The SoM provides much inferior results on trumpet, piano, and flute duets compared to our model, since the gap is larger than 3 dB. This is not very surprising since separating duet of these three instruments mainly relies on hand pose movements. It is very hard for the trajectory and optical flow features to capture such fine-grained hand movements. Our approach can overcome this challenge in that we explicit model the body motions by tracking the coordinates changes of hand keypoints. These results further validate the efficacy of body dynamics motions on solving more and harder visual sound separation problems.

Instrument	SoM [56]	Ours
trumpet	1.8	**4.9**
flute	1.5	**5.3**
piano	0.8	**3.8**
violin	6.3	**6.7**
cello	5.4	**6.1**

Table 4: Sound source separation performance on duets of the same instruments under the SDR metric.

Instrument	SoM [56]	Ours
trumpet	18%	82%
flute	14%	86%
piano	30%	70%
violin	26%	74%
cello	28%	72%

Table 5: Human evaluation result for the sound source separation on mixture of the same instruments.

4.5. Visualizations

As a further analysis, we would like to understand how body keypoints matters the sound source separation. Figure 4 visualize the learned attention map of keypoints in the audio-visual fusion module. We observe that our model tends to focus more on hand keypoints when separating guitar and flute sounds, while pays more attention to elbows when separating the cello and violin.

Fig 5 shows qualitative results comparison between our model and the previous state-of-the-art SoM [56] on separating 3 different instruments and 2 same instruments. The first row shows the video frame example, the second row shows the spectrogram of the audio mixture. The third to fifth rows show ground truth masks, masks predicted by SoM, and masks predicted by our method. The sixth to

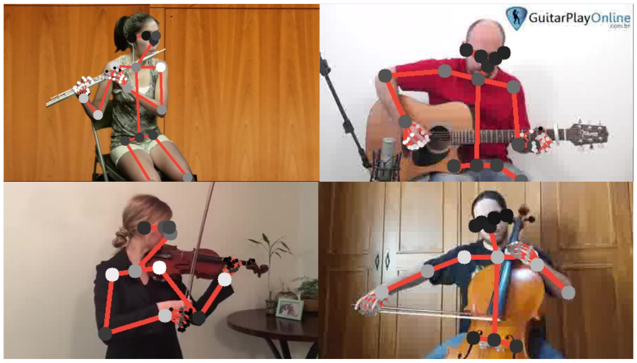

Figure 4: The attention map of body keypoints. Brighter color means higher attention score.

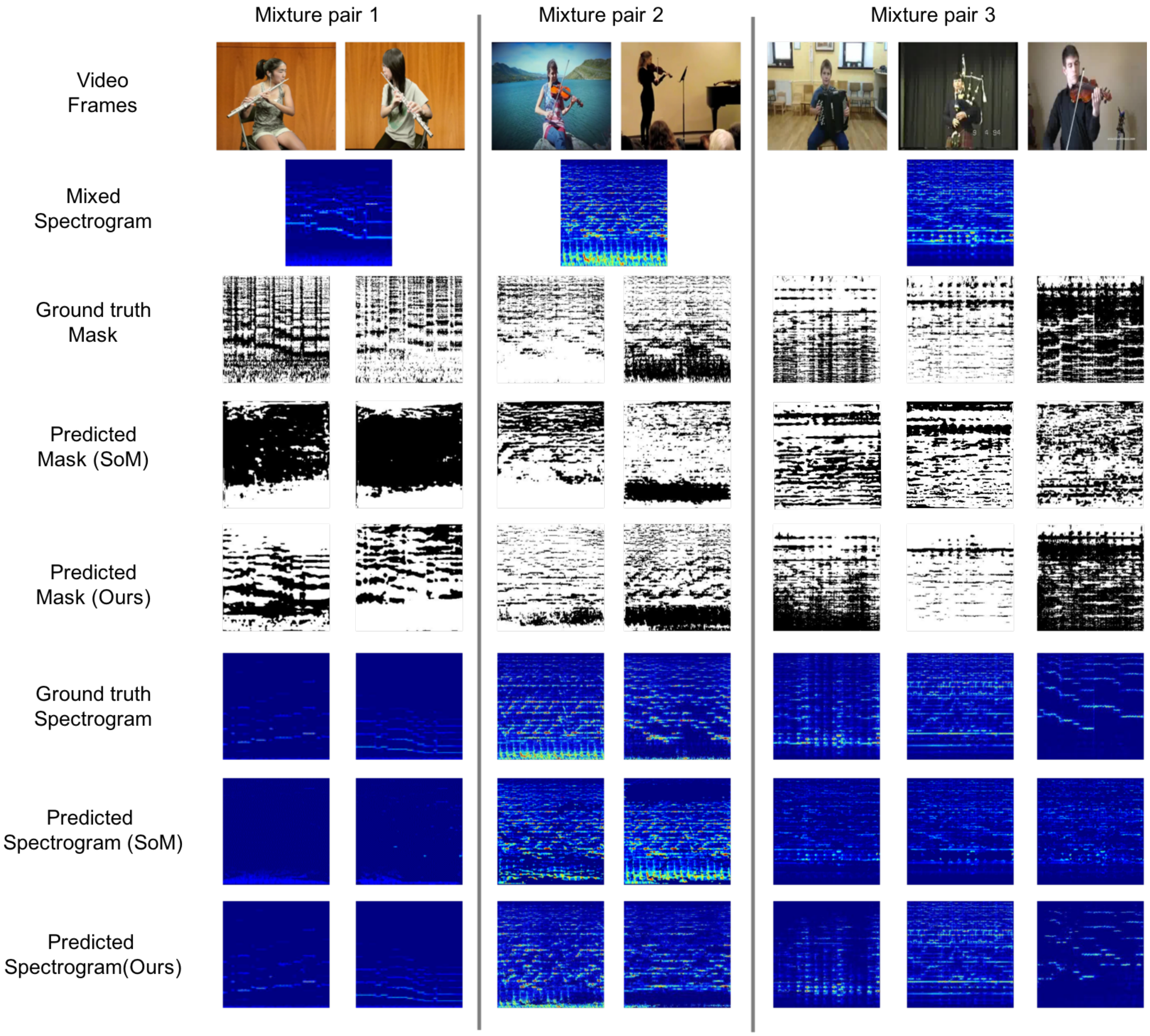

Figure 5: Qualitative results on visual sound separation compared with Sound of Motions (SoM) [56].

eighth rows show the ground truth spectrogram and comparisons of predicted spectrogram after applying masks on the input spectrogram. We can observe that our system produces cleaner sound separation outputs.

Though the results are remarkable and constitute a noticeable step towards more challenging visual sound separation, our system is still far from perfect. We observed that our method is not resilient against camera viewpoint change and body part occlusions of the musician. We conjecture that unsupervised learning of keypoints from raw images for visual sound separation might be a promising direction to explore for future work.

5. Conclusions and Future Work

In this paper, we show that keypoint-based structured visual representations are powerful for visual sound separation. Extensive evaluations show that, compared to previous appearance and low-level motion-based models, we are able to perform better on audio-visual source separation of different instruments; we can also achieve remarkable results on separating sounds of same instruments (*e.g.* piano, flute, and trumpet), which was impossible before. We hope our work will open up avenues of using structured visual representations for audio-visual scene analysis. In ther future, we plan to extend our approach to more general audio-visual data with more complex human-object interactions.

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] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. *ACM Multimedia*, 2018. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617. IEEE, 2017. 2

[3] Relja Arandjelović and Andrew Zisserman. Objects that sound. *arXiv preprint arXiv:1712.06651*, 2017. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016. 2

[5] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 2

[6] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999. 2

[7] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: driving visual speech with audio. 1997. 2

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 4, 5

[9] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *ICLVASS*, pages 258–266, 2017. 2, 5, 6

[10] Joon Son Chung, Andrew W Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017. 3

[11] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009. 2

[12] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 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. 2, 3

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 4, 5

[14] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015. 2

[15] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. *ICRA*, 2020. 2

[16] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *ICCV*, pages 7053–7062, 2019. 2

[17] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 2, 5, 6

[18] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. *arXiv preprint arXiv:1812.04204*, 2018. 2

[19] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. *ICCV*, 2019. 2, 5, 6

[20] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *CVPR*, pages 3497–3506, 2019. 3

[21] Rolf Inge Godøy and Marc Leman. *Musical gestures: Sound, movement, and meaning*. Routledge, 2010. 1

[22] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[24] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 31–35. IEEE, 2016. 2

[25] John R. Hershey and Javier R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 813–819. MIT Press, 2000. 2

[26] Hamid Izadinia, Imran Saleemi, and Mubarak Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2013. 2

[27] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, pages 1–13, 2019. 3

[28] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017. 3

[29] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 4

[30] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2

[31] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music

performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2018. 2, 5

[32] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal keyless attention fusion for video classification. In *AAAI*, 2018. 2

[33] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *CVPR*, pages 7834–7843, 2018. 2

[34] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009. 2

[35] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *NIPS*, 2018. 2

[36] Amit Moryossef, Yanai Elazar, and Yoav Goldberg. At your fingertips: Automatic piano fingering detection, 2020. 2, 5

[37] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. *arXiv preprint arXiv:1804.00326*, 2018. 2

[38] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *ECCV*, 2018. 2, 3, 7

[39] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816. Springer, 2016. 2

[40] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2017. 6

[41] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014. 5

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[43] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361. IEEE, 2019. 2

[44] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. *arXiv preprint arXiv:1803.03849*, 2018. 2

[45] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *CVPR*, pages 7574–7583, 2018. 3

[46] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 4

[47] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 429–436. Springer, 2015. 2

[48] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003. 2

[49] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017. 3

[50] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93, 2017. 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 4

[52] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007. 2, 5, 6

[53] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[54] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 4

[55] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 241–245. IEEE, 2017. 2

[56] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. *ICCV*, 2019. 2, 5, 6, 7, 8

[57] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3, 5, 6

[58] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. *arXiv preprint arXiv:1712.01393*, 2017. 2