LAVSS: Location-Guided Audio-Visual Spatial Audio Separation -Supplementary Material

Yuxin Ye[1], Wenming Yang[1], Yapeng Tian[2]
[1]Shenzhen International Graduate School, Tsinghua University, China
[2]Department of Computer Science, The University of Texas at Dallas, USA
yeyx21@mails.tsinghua.edu.cn, yangelwm@163.com, yapeng.tian@utdallas.com

The supplementary material is organized as follows: Section A explains the discussion and limitation about the proposed method; Section B provides additional qualitative visualization of the source separation; Section C contains extra details of the proposed network architectures; Section D explains the training and evaluation configurations of experiments, and Section E adds extra experiments.

A. Discussions

This section explains the discussions about our motivation, limitation, and future direction.

Our proposed architecture is designed for the separation of binaural channels, which produces predicted masks for both channels. As illustrated in Fig. 1, both the time-discrete signals and spectrograms are obviously different between the left and right channel. Furthermore, we introduce the IPD feature between two channels, which leverages the phase information of input channels and benefits network learning. Consequently, we utilize two channels as input and produce two masks for each channel. Finally, we obtain the separated audios of sounding objects for each channel.

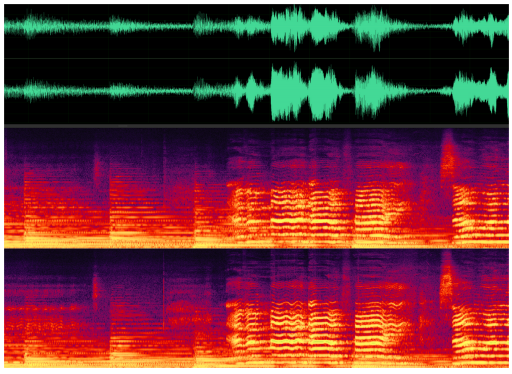

Figure 1. One visualization example of the left and right channel.

In real world videos, there could include scenario noise and background sounds, which are not able to be localized in one certain direction. Our approach is limited to the scenario composed of visible sounding objects and background noises. In addition, all videos we used were recorded in a music room with a 3Dio binaural microphone and a GoPro camera. The position of the performer hardly changes throughout the video. This constraint limits the application for actively moving sound scenarios.

One promising future research direction is to extend our framework to learn audio-visual spatial audio separation in general videos, in which multiple visual objects could coexist and not all of them make sound. Moreover, we prepare to generalize our approach to active sound separation. For another aspect, we will make effort to address comprehensive scenarios containing more than two objects by encoding individual visual and positional features and exploring more effective ways of utilizing the spatial information.

B. Additional Qualitative Results

This section provides the external qualitative results and videos of our proposed separation method. The settings of the experiments are revealed in the main paper.

Fig. 3 presents additional visualization results of separating binaural mixtures using LAVSS from the solo FairPlay datasets. Quantitative experiments in duet videos are not available since the ground truth is unknown. However, we show qualitative video results of our LAVSS method compared to the baselines in the supplementary video. Our source code and pre-trained models will be released.

C. Network Architectures

This section provides additional details of the network structures and implementation.

C.1. Vision-Position Embedding Framework

Video processing and vision network For object detection, we screen out one object with the highest confidence score among all detected ones as the audible ob-

ject. We have verified that the accuracy of selecting the audible objects is up to 90%. For an RGB image of size $3 \times H_b \times W_b$, we perform frame augmentation to resize the image of $3 \times 224 \times 224$. Then we utilize the image encoder of ResNet-18 ($stride = 32$) [2] to extract the visual feature F_v of size $C_v \times H \times W$, where $H = W = 7, C_v = 512$.

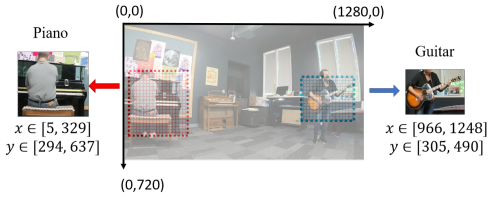

Figure 2. The encoding range of x and y coordinates for objects.

Position network For the detected coordination area of size $H_b \times W_b$, we conduct positional encoding through each pixel of the area and obtain $C_e \times H_b \times W_b$, where C_e equals 64. It is first passed to the adaptive max pooling to reduce the last two dimension size of $H \times W$. Then the multi-layer perception (MLP) is used to extract the positional feature F_p of size $C_p \times H \times W$. The MLP includes two hidden layers of 256 and 512 channels followed by a ReLU activation. Note that C_p is equal to the vision feature vector dimension C_v in the previous section.

The positional embedding of the cropped object region is exactly computed on the **entire image scene**, which illustrates the relative position between the object and video frame of size 1280×720. As depicted in Fig. 2, we encode the position of piano to range $x \in [5, 329]$ and $y \in [294, 637]$ (not from 0 for every cropped object). Thus, the network takes the discriminative position as input.

VP cross attention module We flatten over the last two axes of F_v and F_p and infuse them with a cross-modal attention module. The output vector F'_v and F'_p are concatenated and passed to a 2D convolutional layer to halve the channel dimension. Finally, the visual-positional feature vector F_{vp} is generated of size $C_{vp} \times H \times W$, where C_{vp} equals 512.

C.2. Multi-modal Sound Source Separation

Audio embedding network The audio waveform for both channels is firstly converted to spectrogram representation X_m^L, X_m^R of size $1 \times H_s \times W_s$ using STFT transform, where $H_s = W_s = 256$. In addition, the IPD feature is of the same dimension and concatenated with left and right spectrograms, respectively. Then the audio feature of size $2 \times H_s \times W_s$ is passed to the U-Net encoder, which is composed of 7 down-convolutional layers.

Multi-scale audio fusion network At the bottleneck, the audio feature is of size $512 \times 2 \times 2$. For multi-modal feature fusion, we do not purely utilize the feature extracted after the last down-convolution, since the feature size is too small. Consequently, we perform multi-scale feature fusion for the last three layers by concatenation after flattening the last two dimensions. The multi-scale audio feature F_a is of size $C_a \times Q_a$, where $C_a = 512, Q_a = 84$. The AVP cross-attention module infuses F_a and F_{vp} followed by a concatenation operation and 1×1 convolutional layer. The cross-modal feature F_{avp} is of size $1024 \times 2 \times 2$ and fed into the up-convolutional layers. With the $thresholding(th = 0.5)$ operations, the audio features are converted to binary masks $\mathcal{M}_n^L, \mathcal{M}_n^R$, which are then formulated by element-wise multiplication with the original mixture spectrograms X_m^L, X_m^R. The estimation of separated audio waveforms $\widetilde{x}_n^L(t), \widetilde{x}_n^R(t)$ are obtained after ISTFT.

Signal reconstruction In LAVSS, we leverage the phase of the mixed signal for reconstruction. There are two reasons: 1) We follow the same reconstruction method in 2.5D [1] for binaural audios separation (kindly see 2.5D Supp.C). Prior works on multi-microphone enhancement also multiply T-F masks by the mixed signal to reconstruct microphone arrays at different positions [3]. 2) It is difficult to directly estimate the phase since the variability of slight deviations. Thus, we measure the loss between masks and add time domain loss to alleviate this issue. Complex spectra could be used as an implicit spatial feature to replace the explicit IPD feature. This would be an interesting idea to address the issue in the future.

C.3. Transfer learning by monaural dataset

Details of pre-training During pre-training on the monaural MUSIC dataset, the sound separation network takes a single channel spectrogram input of size $1 \times T \times F$. Then the U-Net weights are re-trained by taking input of size $2 \times T \times F$ after adding the IPD information. In order to handle the mismatch of loading the pre-trained weights, we purely drop the weights of the first up-convolutional layer and keep the remaining network weights of the sound separation and visual network. The experiment results confirms that it has slight impact on the whole training process.

D. Implementation Details

Training optimization We train our LAVSS framework with the implementation of PyTorch and apply Adam optimizer with momentum 0.9, weight decay 1e-4, and batch size 32 for training on 4 NVIDIA 1080 Ti GPUs. For monaural pre-training, we train the vision and sound separation network on MUSIC dataset by using learning rates of 1e-4 and 1e-3. During spatial audio training, we load the pre-trained weights for end-to-end re-training. The learning rate of the vision, position, and sound separation network on the Fair-Play dataset is set to 1e-4, 1e-4, and 5e-4, respectively.

Evaluation The separation performance is only measured by SIR and SDR. The SAR is not informative since it captures only the absence of artifacts. Hence it can be high when the separation result is poor. We evaluate both solo and duet videos. For solo videos, we conduct a "mix-and-separate" strategy (the same as the training process) and evaluate the separation results with the ground truth for both quantitative and qualitative results. For duet videos, we compare the qualitative evaluation of different methods with our LAVSS and inference from the separated sound.

E. Supplementary Experiment Results

Overlap and generalization Due to the overlap of instruments between the MIT MUSIC and FAIR-play datasets, we evaluate the performance of the model for overlapping and distinct instruments with or without pre-training, respectively. The results are shown in Tab. 1. There are only three overlapping instruments between the two datasets: guitar, cello, and trumpet. We compare the model's performance between overlapping instruments and distinct ones in the FAIR-Play dataset. The following results reveal a small difference between the two types. Rich training samples alleviate the difficulty of the separation network, revealing good generalization on distinct instruments.

Configuration	Type	Left Channel		Right Channel		Average	
		SDR↑	SIR↑	SDR↑	SIR↑	SDR↑	SIR↑
Only pre-train	Overlap	5.13	8.63	5.19	8.66	5.16	8.65
	Distinct	4.99	8.54	5.02	8.58	5.01	8.56
Final model	Overlap	5.94	11.79	6.17	11.76	6.06	11.78
	Distinct	5.79	10.06	5.84	10.14	5.82	10.10

Table 1. Comparisons of separation results for overlap and distinct instruments with or without pre-training.

References

[1] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[3] Tomohiro Nakatani, Rintaro Ikeshita, Keisuke Kinoshita, Hiroshi Sawada, and Shoko Araki. Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6129–6133. IEEE, 2021. 2

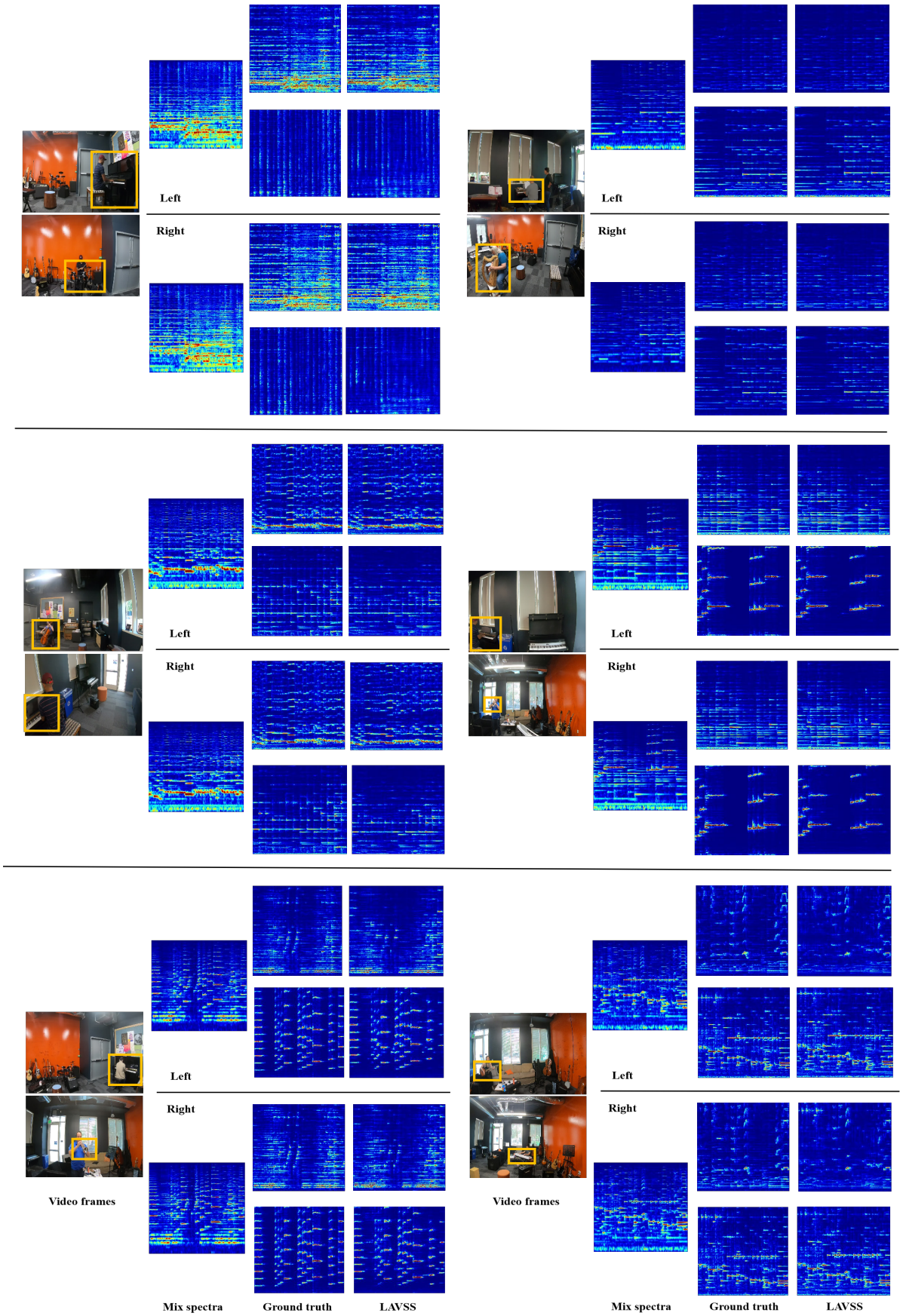

Figure 3. Examples visualization of the sound source separation performances for left and right channel using LAVSS with mixtures of two different sources from Fair-Play dataset.