LAVSS: Location-Guided Audio-Visual Spatial Audio Separation

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

Abstract

Existing machine learning research has achieved promising results in monaural audio-visual separation (MAVS). However, most MAVS methods purely consider what the sound source is, not where it is located. This can be a problem in VR/AR scenarios, where listeners need to be able to distinguish between similar audio sources located in different directions. To address this limitation, we have generalized MAVS to spatial audio separation and proposed LAVSS: a location-guided audio-visual spatial audio separator. LAVSS is inspired by the correlation between spatial audio and visual location. We introduce the phase difference carried by binaural audio as spatial cues, and we utilize positional representations of sounding objects as additional modality guidance. We also leverage multi-level cross-modal attention to perform visual-positional collaboration with audio features. In addition, we adopt a pretrained monaural separator to transfer knowledge from rich mono sounds to boost spatial audio separation. This exploits the correlation between monaural and binaural channels. Experiments on the FAIR-Play dataset demonstrate the superiority of the proposed LAVSS over existing benchmarks of audio-visual separation. Our project page: https://yyx666660.github.io/LAVSS/.

1. Introduction

Auditory and visual characteristics can convey important semantic and spatial information, which plays a crucial role in audio-visual separation [76]. The well-known cocktail party problem [2] is a classical task of sound source separation [11, 56, 71] and localization [45, 46]. It aims at separating the target source audio from the given audio mixture. A popular line of work for audio-visual separation is to encode visual information as guidance for resolving sound ambiguity from mixed audio sources [66, 73, 77]. For instance, lip motion [14, 30] and facial expression [26] information were
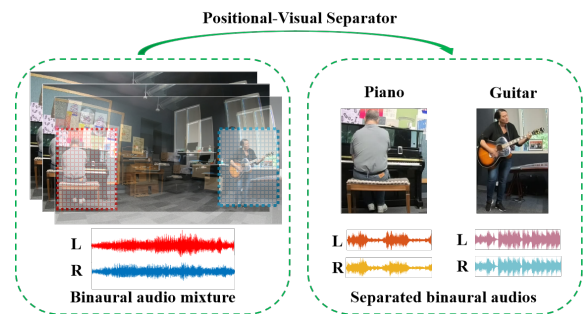

Figure 1. Our LAVSS can separate individual binaural sounds for sounding objects (piano and guitar) from a binaural audio mixture.

applied to separate speech sounds from different speakers. Motion [70, 78] and gesture [9] appearance features were exploited to guide music sound separation. Other methods utilize instrument category [4, 13] or multimodal attention [5, 59, 60] to leverage the association between visual and audio modalities.

Predominant audio-visual separation (AVS) methods have typically been designed for monaural audio-visual separation (MAVS). However, scenarios limited to single-channel audio lack the capacity for perceiving 3D visual scenes accompanied by spatial audio. Although being attempted earlier in [12], researches on *audio-visual spatial audio separation* (AVSS) (see Fig. 1) are highly limited. For example, audiences can discern the orientation of the piano and guitar since they hear the mixed spatial audio with varying acoustic intensities for each ear [12]. Unlike MAVS, AVSS provides listeners with a more immersive perceptual experience, thus making it a novel and challenging task.

Existing spatial audio-visual works have mainly focused on spatial audio generation [12, 29, 39, 67]. This involves converting standard monaural audio into binaural or ambisonic sounds. Sep-stereo [74] regards MAVS as a specific case of binaural audio reconstruction at the cost of artificially rearranging visual information. However, these methods lack sufficient audio-visual modeling and still exhibit a domain gap when it comes to spatial audio separation.

In this paper, we address the audio-visual spatial audio

---

*Corresponding author.

separation task by simultaneously considering **what and where** the sounding object is. In an effort to overcome current limitations, we introduce a new Location-Guided Audio-Visual Spatial Audio Separation (LAVSS) method. We first detect sounding objects to obtain regional visual embeddings (what). Then we encode the spatial location of the sounding objects explicitly. The positional embeddings can be another guidance to reveal the spatial information (where), which benefits separating individual audio in different directions. How does this correspond to audio? Since binaural audio carries spatial information cues, we consider the inter-microphone phase difference (IPD) [63,64], which is commonly used in multi-microphone speech segregation and separation [41, 52, 68]. The IPD information represents the established spatial feature between the left and right channel. We force the network to learn the synchronization and correlation between the spectra-spatial audio feature and the visual-positional representations. Moreover, we propose a multi-scale attention-based fusion network to integrate the visual, positional, and audio features. All constituent modalities work in concert to benefit AVSS.

Additionally, to leverage the correlation between monaural and binaural channels, we employ a pre-trained separator. This aids in the knowledge transfer from rich mono sounds, thereby enhancing spatial audio separation. By utilizing the extensive video data with monaural sounds available in the MUSIC-21 dataset, we accomplish effective pre-training. Experiments on the binaural FAIR-Play dataset can validate the efficacy of LAVSS. It achieves state-of-the-art performance, particularly in scenarios where similar acoustic sources are positioned in different directions.

Our contributions are as follows: i) We put forward a multi-modal framework to address the AVSS task. ii) We take advantage of the correlation between the IPD and positional features, which respectively represent the spatial properties of binaural audio and the explicit location cues of the sounding objects. iii) We pre-train the separator on an external mono dataset to facilitate AVSS network learning by leveraging the correlation between monaural and binaural channels. iv) Experiments demonstrate the superior improvement and generalizability of our LAVSS over state-of-the-art audio-visual separation approaches.

2. Related Work

Audio-Visual Learning Audio-visual learning has gained considerable interest in recent years, with researchers achieving promising results in a variety of fields. These include self-supervised learning [6, 10, 42], audio-visual speech recognition [21,22,30,40,48], visually guided spatial audio generation [12,29,67,74], audio-visual speech and music separation [11, 26, 58, 71] and localization [44, 55, 57, 65, 72], as well as environment acoustics learning [28, 31, 35]. Unlike these prior works, we make the

first attempt to tackle audio-visual spatial audio separation by incorporating visual positional features as an additional modality and employing the cross-modal attention.

Audio-Visual Source Separation Sound source separation is a crucial part of speech front-end research and music processing. Traditional signal processing methods usually exploit filtering to strengthen source separation [1, 7, 16, 25, 54, 62] and localization [45, 46]. Machine learning methods like end-to-end speech separation [18, 32, 33] aim at performing waveform transformation in the time domain. The well-known cocktail party problem [2] is a classical task of sound source separation [11, 56, 71]. Recently the self-supervised visually guided audio-visual source separation has obtained significant attention [13,56,66,71,73,77]. For one aspect, most works exploit appearance features as visual guidance. From the whole image frame [11, 71] to detected sounding object regions [13, 56], these works focus on how to obtain precise visual features. Other visual appearances such as motion [70], gestures [9,49] are exploited to capture the body movement postures of players. Recent works regard the human and instruments as nodes to build the graph relationships between them [3, 4]. For another aspect, some researchers optimize the architecture of the separation network [56, 66, 77] and try to fuse visual-audio modality in an effective manner [73]. For recent studies, vision transformers [5,38,49,78] and attentions [6,59,60] are widely used in multi-modal collaboration. From 2D to 3D, active sound separation [34, 36] for AR/VR scenarios has become promising future research. However, methods basically conducted for mono audios have limited capabilities with spatial ones in real scenarios. Different from MAVS approaches, we propose to relate spatial cues of audio and sounding objects to resolve AVSS.

Audio-Visual Spatial Audio Generation Audio-visual cross-modality generation aims to generate audio from visual signals [8,10,12,19,24,29,69,75]. For instance, Zhou *et al.* [75] utilize the synchronization of visual cues and encoders to generate natural sound for videos in the wild. Zhou *et al.* [10] and Gao *et al.* [12] adopt a U-Net to encode monaural input and decode binaural counterpart through visual guidance at the bottleneck. Sep-stereo [74] put forward an associative pyramid structure to better fuse audio and visual modalities for generation stereo. Other methods [8,24] generates audio samples conditioned on text inputs, motion key points, and position information [15, 43], respectively. Other works concentrate on 360° audio generation and spatialization. Scene-aware audio [27] can be converted from a single-channel microphone and transformed into spatial audio. Morgado *et al.* [39] take real spatial audio as self-supervision for ambisonic audio generation. Different from these works, our main focus lies in spatial audio separation.
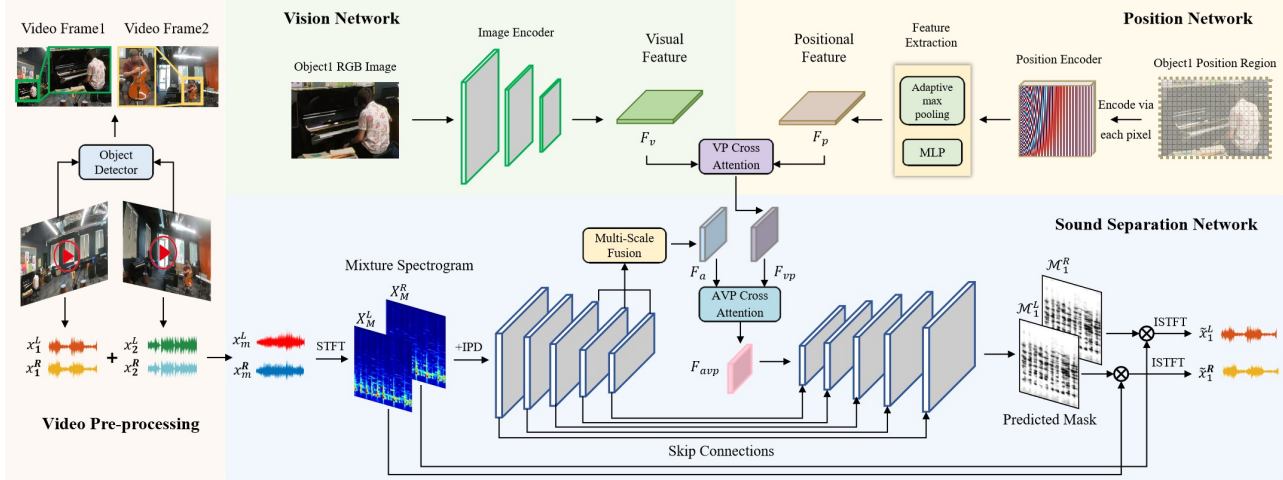
Figure 2. An overview of the proposed architecture. Video pre-processing includes object detection and source mixing. Vision extraction network encodes the visual regions of detected objects; position network simultaneously encodes regional coordination features; VP Cross Attention module aggregates visual and positional representations; sound separation network exploits the fused feature as guidance to separate binaural sounds. Note that all the operations are depicted for only one video (piano), the other remains the same during training.

3. Proposed Method

3.1. Overview

Given an unlabeled video segment V and its corresponding spatial audios $x^L(t)$ and $x^R(t)$, the detected audible objects are defined as $\mathcal{O} = \{O_1, ..., O_N\}$ for each video frame. Our spatial audio separation task aims to separate the individual audio of each sounding object from the mixed audio: $x^L(t) = \sum_{n=1}^{N} x_n^L(t), x^R(t) = \sum_{n=1}^{N} x_n^R(t)$, where $x_n^L(t)$ and $x_n^R(t)$ represent the time signals received at both ears of corresponding object sources.

As depicted in Fig. 2, our LAVSS training architecture consists of four parts: the video pre-processing module, a vision and position network, and a multi-modal sound separation backbone. During video pre-processing, we utilize two sets of solo videos and their synchronized spatial audios $\{V_1, x_1(t)\}, \{V_2, x_2(t)\}$ with sounding objects O_1, O_2 in both videos [56], we artificially mix two binaural sounds: $x_m^L(t) = x_1^L(t) + x_2^L(t), x_m^R(t) = x_1^R(t) + x_2^R(t)$. Then we perform object detection to obtain the object bounding boxes and the corresponding coordinates of the objects. The vision network encodes the detected objects to produce visual features. For the position network, we conduct positional encoding for each pixel in the visual object region. The visual and positional features represent the semantic and spatial information of the sounding object, respectively. Both features are mapped into a common embedding space and performed attention-based fusion.

The binaural audio mixture is transformed into the time-frequency domain and passed to an encoder-decoder sound separation network, which is pre-trained on an external monaural dataset. We creatively introduce the inherent IPD

between the left and right channels for spatial audio separation. The IPD feature and magnitude spectra are concatenated to leverage both spatial and spectral cues of audio, which correspond to the visual and positional features of the object. All features are fused through a multi-scale attention-based fusion module and transformed into time-discrete space. Finally, we obtain the estimated binaural audios $\hat{x}_n^L(t), \hat{x}_n^R(t)$ of individual objects. More details of our LAVSS are provided in the supplementary material.

3.2. Vision-Position Embedding Framework

Vision network In order to precisely localize the audible objects, we choose the widely used detector Faster R-CNN [51] trained on labeled Open Images dataset [23] used in [13,56]. All potential objects $\mathcal{P} = \{P_1, ..., P_N\}$ for each video are detected. Given a video frame V, detections of all objects consist of four items $\{(M_V^n, C_V^n, P_V^n, B_V^n)\}_{n=1}^{N} = $ FRCNN (V), which represent the frame index M, instrument category $C \in \mathcal{C}$, detection confidence probability P and bounding box B for each detected object. Then we screen out one object with the highest confidence score among all detected ones as the audible object for each solo video frame (top two for duet video).

The visual image region for the selected object is of size $3 \times H_b \times W_b$, where H_b, W_b denote the height and width of the detected bounding box. For visual feature extraction, objects are resized and passed to a pre-trained ResNet-18 [20] network. We obtain the visual embedding $F_v \in \mathbb{R}^{C_v \times H \times W}$ before the last fully-connected layer, where H_b', W_b' represent the resized image shape. $H = H_b'/32, W = W_b'/32, C_v = 512$ denote the feature map size and channel dimension of F_v, respectively.

(a) VP Cross-attention module
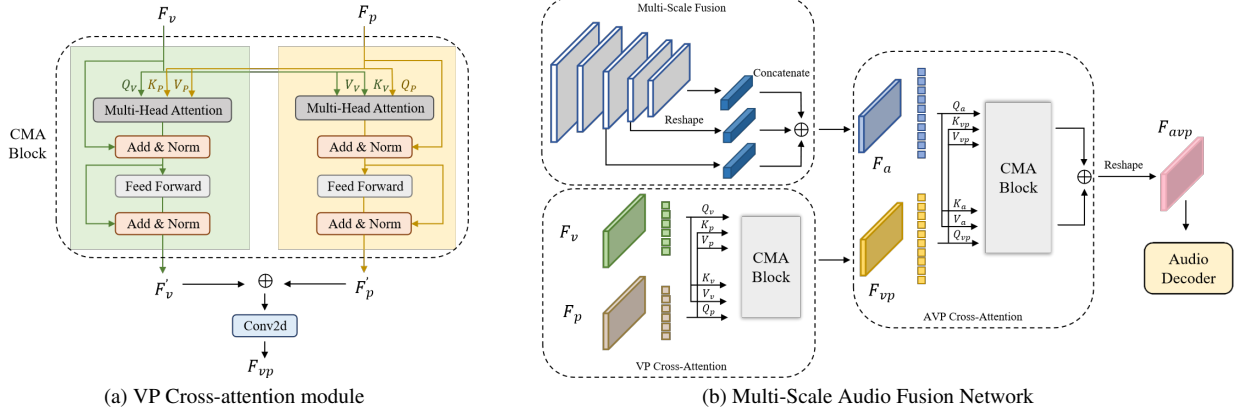(b) Multi-Scale Audio Fusion Network

Figure 3. Two basic blocks of multi-attention modules. (a) VP Cross-Attention, in which the vectors of visual and positional features are integrated through Cross-Modal Attention (CMA) block; (b) The architecture of multi-scale audio fusion network, which consists of the multi-scale fusion, VP/AVP cross-attention modules to introduce the interactions between vision, position and audio modalities.

Position network   Going beyond the general MAVS strategy, one of the critical innovations of our method is specializing in spatial audio separation. Specifically, we leverage positional representations as a new constituent modality and demonstrate the association with spatial distribution embedded in spatial audio. Inspired by the positional encoding in Transformer [61] and NeRF [37], we consider how to encode positional representations of audible object regions into a higher dimensional space. We leverage a 2D positional encoding for spatial coordinates of detected objects, thus forcing our positional network to approximate a higher frequency function and guide spatial audio separation. Here the function $\gamma(\cdot)$ represents a mapping function from low-dimensional space into a higher one,

$$
\begin{aligned}
\gamma(x,y) = \big(&\sin(2^0\pi x), \cos(2^0\pi x), \sin(2^0\pi y), \cos(2^0\pi y), \ldots, \\
&\sin(2^{D-1}\pi x), \cos(2^{D-1}\pi x), \sin(2^{D-1}\pi y), \cos(2^{D-1}\pi y)\big)
\end{aligned}
\tag{1}
$$

This sinusoidal function is applied simultaneously to 2D coordination in $(\mathbf{x}, \mathbf{y})$ (which are normalized to range $[-1, 1]$ [37]) for expanding to higher dimensions via gamma encoding. In our experiments, we set $D = 16$ for $\gamma(x, y)$ to encode each pixel in the detected object region relative to the whole video frames of size 1280×720. Then we obtain a tensor of size $C_e \times H_b \times W_b$ by Eq. (1), where $C_e = 64$ denotes the dimension of the encoded positional embedding. For position feature extraction, the encoded features are performed adaptive max pooling followed by multi-layer perception (MLP). Finally, the positional feature is converted to $F_p \in \mathbb{R}^{C_p \times H \times W}$, where C_p is equal to the vision feature dimension C_v in the previous section.

VP Cross Attention Module   For multi-modal modeling, the VP cross-attention module is implemented to integrate the visual and spatial position embeddings. As illustrated in

Fig. 3 (a), the VP Cross-Attention module is composed of a CMA block and a convolutional layer. For instance, given an input query $M \in \mathbb{R}^{H_m \times W_m \times D}$ and $N \in \mathbb{R}^{H_n \times W_n \times D}$, $CMA(M, N, N)$ performs cross-modal attention over the first and second axes of N, yielding an output tensor of shape $H_m \times W_m \times D$,

$$
\begin{aligned}
\alpha &= LN(MHA(M_Q, N_K, N_V) + M) \\
CMA(M, N, N) &= LN(FFN(\alpha) + \alpha)
\end{aligned}
\tag{2}
$$

where M_Q is the query vector of M, N_K, N_V are key and value vectors of N. MHA, FFN, LN denote the multi-head attention, feed-forward layer, and layer normalization, respectively. The F_v and F_p are first passed to the CMA block. Then the visual-positional feature $F_{vp} \in \mathbb{R}^{C_{vp} \times H \times W}$ can be obtained after a convolutional layer to halve the channel dimension. The core part of the module is given by,

$$
F_{vp} = Conv(CMA(F_v, F_p, F_p) \oplus CMA(F_p, F_v, F_v))
\tag{3}
$$

where $\oplus$ and Conv denote the concatenate operation and point-wise convolution, respectively.

3.3. Multi-modal Sound Source Separation

Audio Embedding Network   We follow the widely used mix-and-separate [71] method and manually mix two video sounds. The time-discrete binaural audio waveform $x_m^L(t), x_m^R(t)$ are first converted to time-frequency spectrograms X_m^L, X_m^R through STFT [17] transform. Several previous MAVS works [13, 56, 71] take only log power spectra as the input of the network. In terms of spatial audio, sound source locations are determined by time differences between the sound sources reaching each ear [12, 50], which can be measured by the inter-microphone phase difference (IPD) between the left and right channels. IPD increases the

feature discrimination of location information and indicates the spatial acoustic characteristics of the room. Alternatively, it reveals the different directions of the same sounding objects. The IPD can be calculated as follows,

$$IPD = cos(\angle X_m^L - \angle X_m^R) \quad (4)$$

where $\angle$ represents the phase angle of the complex spectrogram. One way of utilizing such multi-channel inputs is to feed the network with both log power spectra and IPD features [68]. We concatenate both features and obtain the audio embedding of size $2 \times T \times F$ for each channel, where T and F represent the time and frequency dimensions, respectively. In this manner, the input of the sound separation network contains both the acoustic spectra (what) and spatial cues (where) carried by the binaural audio.

Then a U-Net [53] backbone is used for encoding the composed spectrograms and IPD feature into semantic representations. The architecture is composed of N down- and up-convolutional layers followed by a BatchNorm layer and Leaky ReLU. At the bottleneck, the multi-scale audio fusion network performs multi-modal modeling over the audio, vision, and position features. Note that the sound separation network parameters are shared across left and right channels during training and testing.

Multi-Scale Audio Fusion Network To establish the relationship between the spectra-spatial audio feature and the visual-positional representations, we put forward a multi-scale audio fusion network visualized in Fig. 3 (b). For multi-scale feature fusion, three feature tensors $F_a^{N-i}(N = 7, i = 0, 1, 2)$ extracted by the last three down-sample convolutional layers are reshaped to $C_a \times Q_a^{N-i}$ by multiplying the time and frequency dimension. Then $f_{Concat}(\cdot)$ performs concatenation along the query dimension to generate audio queries $F_a \in \mathbb{R}^{C_a \times Q_a}$,

$$F_a = f_{Concat}(F_a^N, F_a^{N-1}, ..., F_a^{N-i}), i = 0, 1, 2 \quad (5)$$

The audio feature F_a is fed into the AVP cross-attention module to adaptively interact with the visual-positional feature F_{vp}. The output audio embedding $F_{avp} \in \mathbb{R}^{C_a \times \frac{T}{S} \times \frac{F}{S}}$ (S denotes stride of audio feature map) is computed by

$$F_{avp} = f_2(CMA(F_a, F_{vp}, F_{vp}) \oplus f_1(CMA(F_{vp}, F_a, F_a))) \quad (6)$$

where $\oplus$ is the concatenate operation, $f_1(\cdot)$ denotes the one-dimensional convolusion, $f_2(\cdot)$ means dimensional expansion and two-dimensional convolution operation. The feature vector F_{avp} is regarded as guidance for audio separation and passed to the decoder up-sample layers of U-Net. Finally, we obtain the predicted magnitude binary masks $\hat{\mathcal{M}}_n^L, \hat{\mathcal{M}}_n^R$, which are multiplied by the original mixture spectrogram X_m^L, X_m^R to produce the final estimation of output spectrograms. The estimated audios $\hat{x}_n^L(t), \hat{x}_n^R(t)$ are

obtained after ISTFT. More specifically,

$$\hat{x}_n^B(t) = ISTFT(\hat{\mathcal{M}}_n^B \odot X_m^B)$$
$$\mathcal{M}_{gt,n}^B(u,v) = [X_n^B(u,v) \geq X_m^B(u,v)] \quad (7)$$

where $\odot$ denotes element-wise multiplication, (u,v) represents time-frequency dimension, $B \in [L, R], n \in [1, 2]$ (number of the objects). The ground truth of binary masks $\mathcal{M}_{gt,n}^B$ are created by the ratio between the source spectrograms X_n^B and the mixture spectrograms X_m^B.

Overall learning Objective We optimize our LAVSS framework training objective by jointly minimizing a combination of both frequency and time reconstruction losses. For the frequency domain loss, we measure the linear combination between the L1 and L2 losses over the predicted ratio masks and ground-truth in Eq. (8). Furthermore, we introduce the loss between the target audio $x_n^B(t)$ and reconstructed audio $\hat{x}_n^B(t)$ over the time domain. Formally,

$$\mathcal{L}_{freq} = \sum_{n=1}^{N} \sum_{B \in L, R} \|\hat{\mathcal{M}}_n^B - \mathcal{M}_n^B\|_1 + \alpha\|\hat{\mathcal{M}}_n^B - \mathcal{M}_n^B\|_2$$
$$\mathcal{L}_{time} = \sum_{n=1}^{N} \sum_{B \in L, R} \|\hat{x}_n^B(t) - x_n^B(t)\|_1 \quad (8)$$
$$\mathcal{L}_{binaural} = \mathcal{L}_{freq} + \beta\mathcal{L}_{time}$$

3.4. Transfer learning by external monaural dataset

Due to the complexity of the binaural attributes, the framework designed for spatial audio is complicated for training directly. To alleviate this issue, we choose a widely used mono dataset MIT MUSIC to perform transfer learning for two reasons. First, the binaural FAIR-Play dataset contains much scarce training data due to the recording difficulty. In contrast, the MUSIC dataset includes more instrument categories and videos, which can mitigate the difficulty of AVSS and make the training more robust. Some of the instrument types overlap, which makes the sound separation between similar acoustic characteristics mutually beneficial. Second, considering the relationship between mono and binaural audio, we can transfer knowledge from rich mono sounds to boost spatial audio separation performance. Thus, a pre-trained monaural separator is adopted by training on the MAVS network backbone in [12].

Similar to the training process in Fig. 2, we pre-process the videos in MUSIC dataset. Then we take the monaural mixtures and detected RGB image regions into the U-Net separation and visual network, respectively. Both features are fused by multi-scale attention-based fusion at the bottleneck. Note that the monaural audios do not possess the spatial location information. The IPD and position feature will **not** be considered as input to the network. After training, the separation network can be a good separator for

most mixture audios of different instruments, which simultaneously alleviates network learning for training binaural audios. Finally, we load pre-trained parameters both of the U-Net separation and visual network as initial weights and perform complete position-guided audio-visual separation network training on the FAIR-Play dataset. More details of pre-training are revealed in the supplementary material.

4. Experiment and Results

4.1. Experimental Settings

Datasets In our experiments, both monaural and spatial datasets are used for training. To perform monaural separator pre-training, we use MUSIC dataset [71], which is a commonly used dataset for MAVS. It contains 685 solo and duet videos with 11 instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin, and xylophone. We utilize 520 mono solo videos and split them into train/val/test sets with 468/26/26 for pre-training.

For spatial audio separation, we use the FAIR-Play dataset [12]. The instrument type contains cello, guitar, drum, ukelele, harp, piano, trumpet, upright bass, and banjo. We use 1039 10s solo videos with spatial audio during training and testing. To evaluate the proposed LAVSS model conditioned on the detected object coordination, we randomly split it into train/val/test sets: 728/103/208. Moreover, we evaluate the separation ability of LAVSS for separating multiple sources. We take 418 duet videos to perform testing as illustrated in Fig. 1.

Metrics To measure the quality of separation [47], we adopt the widely-used mir_eval library metrics: Signal-to-Distortion Ratio (SDR) measures both interference and artifacts, Signal-to-Interference Ratio (SIR) measures interference. Higher values indicate a better degree of separation.

Implementation Details We train our LAVSS framework with the implementation of PyTorch. We re-sample the audio at 11025Hz to get approximately 5.9s clip for each video. Then we perform STFT frame length of size 1022 and hop length of 256 [12, 71] to convert the time domain signal into 2D magnitude spectrogram of T, F = 256 after re-sampling to a log-frequency scale. We set the frame rate as 8fps and randomly select one frame per 5.9s video. We resize and crop the detected bounding boxes to 224×224 as the input of the ResNet-18 network. The MLP consists of two layers of 256, 512. All the attention modules are set of 8 heads and 2 decoder layers. In Eq. (8) the α and β are set to 0.5 and 0.25, respectively. We apply Adam optimizer with $\beta_1 = 0.9$ and a weight decay of 1e-4. Since the MAVS and AVSS tasks are mutually related, we need to learn good initial models for AVSS. We start by pre-training

on the MUSIC dataset to train the vision and sound separation network. Secondly, we introduce the IPD feature and co-learn the position network on FAIR-Play initialized with the pre-trained weights. The evaluation details are illustrated in the supplementary material.

4.2. Audio-Visual Sound Separation

Comparison with State-of-the-Art To evaluate the performance of our LAVSS framework on audio-visual sound separation, we compare it to two baselines most related to binaural audio separation and generation: 2.5D Separation [12] and Sep-Stereo [74], and recent state-of-the-art methods: SoP [71], Co-separation [13], and CCoL [56].

Note that five methods are evaluated for fair comparison on the FAIR-Play binaural dataset (including audio preprocessing) as ours. Since those methods are specialized in MAVS, we take the left and right channels into the network separately for training (after pre-training on the MUSIC) and evaluation. The SDR and SIR quantitative analysis are illustrated in Tab. 1. The results show that our LAVSS model outperforms its closest competitor, Sep-Stereo [74], by an obvious superiority of 0.65 dB on SDR and 2.82 dB on SIR for binaural channels. Notably, our LAVSS boosts the SDR and SIR metrics by 0.80dB and 1.23dB compared to the most recent baseline CCoL [56]. The above MAVS methods mainly utilize appearance-based visual information, which cannot generalize to AVSS. In contrast, our LAVSS simultaneously considers what and where the object is, thus demonstrating competence for the AVSS task.

Method	Left Channel		Right Channel		Average	
	SDR↑	SIR↑	SDR↑	SIR↑	SDR↑	SIR↑
SoP [71]	3.98	7.03	3.96	6.99	3.97	7.01
2.5D [12]	4.44	8.20	4.47	8.26	4.45	8.23
Co-Sep [13]	4.61	7.93	4.64	8.00	4.63	7.97
Sep-Stereo [74]	5.27	7.34	5.31	7.40	5.26	7.37
CCoL [56]	5.05	8.89	5.17	9.02	5.11	8.96
LAVSS (Ours)	**5.89**	**10.08**	**5.93**	**10.30**	**5.91**	**10.19**

Table 1. Comparisons of methods for source separation results on FAIR-Play test set. Higher is better for all metrics.

Models	Cello	Drum	Guitar	Harp	Piano	Trumpet
SoP	-2.12	-1.88	-2.69	-1.77	-2.35	-1.78
2.5D-sep	-0.93	0.86	-2.00	-1.49	0.35	-2.55
CCoL	-1.75	-1.48	0.19	-0.55	-0.90	-0.54
Sep-Stereo	-1.32	0.34	1.68	-0.41	1.38	0.73
mix-gt	0.58	0.25	0.67	0.47	1.17	1.51
LAVSS(ours)	**1.53**	**2.67**	**3.72**	**1.83**	**3.13**	**3.12**

Table 2. The average separation results for both channels of the same instrument types from FAIR-Play in terms of SDR.
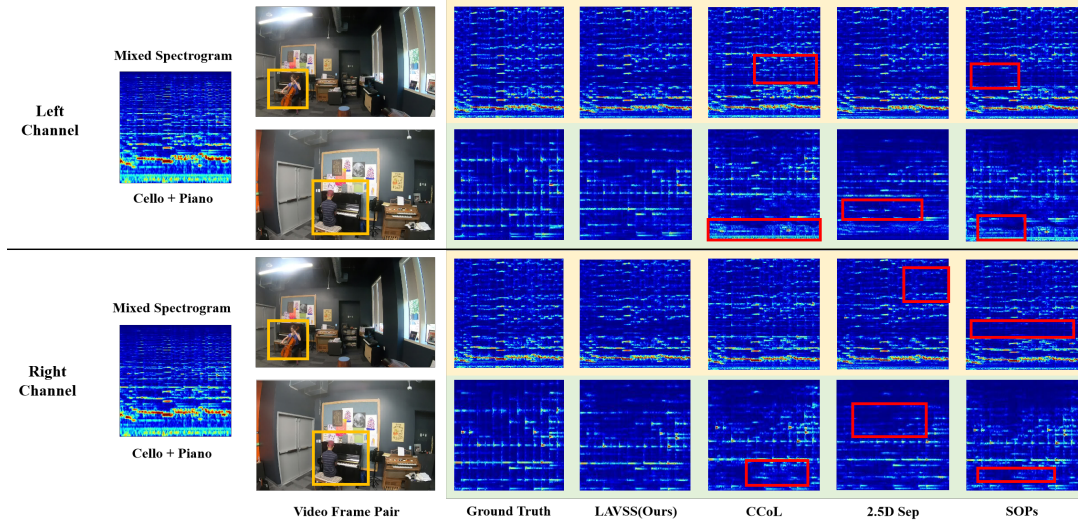
Figure 4. A set of solo separation results on FAIR-Play test set. Predicted spectrograms of SOTA methods and LAVSS are depicted for both channels. Red boxes illustrate the difference between the predicted spectrogram and the ground truth.
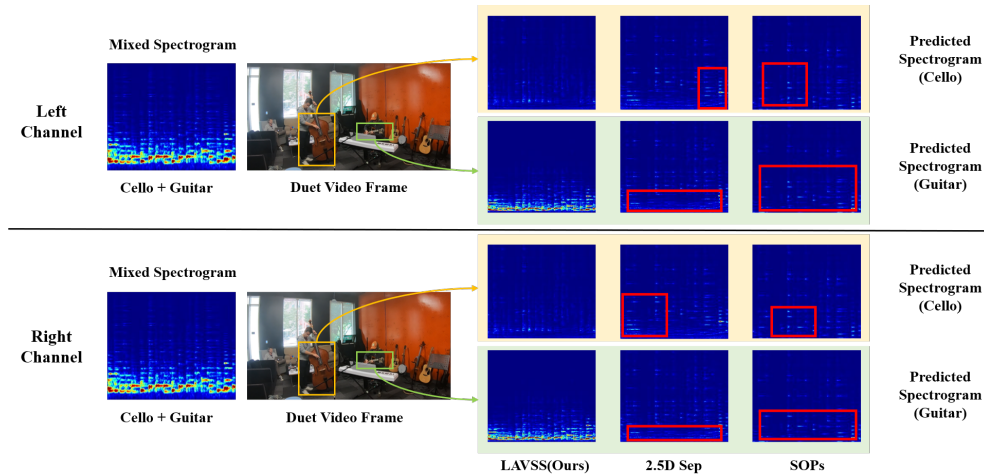

Figure 5. A set of duet separation results on FAIR-Play. Predicted spectrograms (cello and guitar) of SOTA methods and LAVSS are depicted for both channels. Red boxes indicate a comparison of separation ability between LAVSS and benchmarks.

Separating Sources of the Same Type   The source type is one of the critical factors affecting the performance of the separation. When two sounds have similar acoustic properties, separation becomes more complicated. In this case, the appearance features can not provide useful cues regarding similar images, while the location information guidance is particularly critical. Consequently, we select instruments of the same category from the FAIR-Play dataset and compare the separation performance of LAVSS with the MAVS methods. SoP [71] and CCoL [56] are mainly based on appearance guidance, 2.5D-sep [12], and Sep-Stereo [74] are associated with binaural audio generation. Furthermore, we illustrate a comparison result called "mix-gt" to intuitively measure the mix spectrogram with the ground truth.

Table 2 demonstrates the averaged SDR results of both channels for cello, drum, guitar, harp, piano, and trumpet categories. The "mix-gt" baseline shows relatively better results in most cases, which indicates the challenges in monaural appearance-based models for spatial audio separation. Our method outperforms all MAVS baselines for all categories. The CCoL specifies the combinations of instruments selected for different types during training. The Sep-Stereo artificially rearranges the visual images and ignores the original location in the video frame. Fig. 6 shows a case of separating the sound mixture of the same type in different locations. Our method confirms that the relationship between object position and spatial phase cues brings significant improvement in separating similar sources.

4.3. Ablation Study and Performace Verification

Ablations of modality configurations We conduct ablation study to evaluate the effectiveness of IPD, position representation, and monaural transfer learning. We choose SoP and 2.5D-sep as baselines for verifying the versatility on benchmark applications. Note that the fusion strategy in Fig. 3 are applied for both baselines. Tab. 3 demonstrates the best scores when all ablation variants are applied, which confirms that the combined setup can be applied to any existing MAVS benchmarks to boost generalization ability.

One of the essential strategies we perform to strengthen AVSS is to explore position representation as a new modality for guidance. Rows 2 and 9 in Tab. 3 overwhelmingly point out the effectiveness of position encoding. Interestingly, we observe that the combination of the position feature and IPD are mutually beneficial since the network learns spatial location from both binaural audio and visual object. As a result, excavating the spatial properties of binaural audio brings 1.42dB and 2.03dB improvement in SDR and SIR, respectively. We also explore the contribution of transfer pre-training on the external mono dataset. "*" denotes without fine-tuning on the FAIR-Play dataset. Rows 4-6 and 11-13 confirm that it definitely brings about 45% overall performance improvement on both metrics.
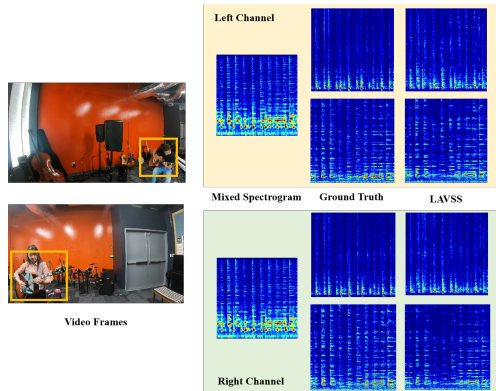

Figure 6. Illustration of the result for separating sound sources of the same type from the FAIR-Play dataset.

Ablations of multi-modal module design Ablation results of multi-scale audio fusion network design on FAIR-Play dataset are shown in Tab. 4. "Tile-Concat" means the vanilla structure of replicating F_{vp} to fit the audio feature F_a at the bottleneck and performing concatenation through channel dimension. "w/o VP/AVP atten." means removing the CMA block. Row 2 and 3 demonstrate that multi-modal fusion based on cross attention promotes stable and improved performance of sound source separation. "w/o multi" indicates that F_a is only composed of tensor extracted by the last down-sample convolutional layer. Notably, multi-scale feature extraction increases the discrimination of audio representations and yields good results.

Baseline Model	Position Guidance	IPD	Monaural Pre-train	Left Channel		Right Channel	
				SDR↑	SIR↑	SDR↑	SIR↑
SoP	✗	✗	✗	3.34	6.45	3.29	6.42
	✓	✗	✗	4.00	7.31	4.02	7.27
	✓	✓	✗	4.32	7.90	4.38	7.86
	✗	✗	✓*	4.22	7.84	4.23	7.88
	✗	✗	✓	4.79	8.36	4.82	8.39
	✗	✓	✓	5.14	8.57	5.15	8.55
	✓	✓	✓	**5.32**	**8.71**	**5.36**	**8.73**
2.5D-sep	✗	✗	✗	3.85	7.24	3.73	7.44
	✓	✗	✗	4.90	8.38	4.82	8.48
	✓	✓	✗	5.27	9.27	5.25	9.28
	✗	✗	✓*	4.67	8.02	4.70	8.03
	✗	✗	✓	5.03	8.56	5.08	8.59
	✗	✓	✓	5.53	9.14	5.59	9.18
	✓	✓	✓	**5.89**	**10.08**	**5.93**	**10.30**

Table 3. Ablation study of two benchmarks on FAIR-Play test set.

Architecture	Left Channel		Right Channel		Average	
	SDR↑	SIR↑	SDR↑	SIR↑	SDR↑	SIR↑
LAVSS (Ours)	**5.89**	**10.08**	**5.93**	**10.30**	**5.91**	**10.19**
w/o VP atten.	5.27	9.34	5.16	9.37	5.22	9.36
w/o AVP atten.	5.04	8.63	5.05	8.65	5.04	8.64
w/o multi.	4.90	8.38	4.82	8.48	4.86	8.43
Tile-Concat	4.83	8.41	4.81	8.38	4.82	8.40

Table 4. Ablations on the design of multi-modal attention module.

Qualitative evaluation Specifically, both solo and duet video separation performances are illustrated in Figs. 4 and 5. Our separated spectrogram is distinctly and completely restored for both channels compared to SoP, 2.5D-sep, and CCoL. For the duet case, the separation results of MAVS hardly show any difference. More qualitative separation results are revealed in the supplementary material.

5. Conclusion

In this work, we present LAVSS, a novel location-guided audio-visual spatial audio separator. We break through the limitation of MAVS methods and put forward AVSS. Our network exploits the synchronization between phase attributes of spatial audio and position embeddings of objects. We leverage location representations of objects and perform fusion with the visual information to consistently guide AVSS. Furthermore, we demonstrate the correlation of monaural and binaural channels by pre-training on external mono dataset for network transfer learning, which outperforms SOTA methods on FAIR-Play. Discussions and future works are provided in the supplementary material.

References

[1] Simon Arberet and Pierre Vandergheynst. Reverberant audio source separation via sparse and low-rank modeling. *IEEE Signal Processing Letters*, 21(4):404–408, 2014. 2

[2] Adelbert W Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000. 1, 2

[3] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. Learning audio-visual dynamics using scene graphs for audio source separation. In *NeurIPS*, 2022. 2

[4] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *ICCV*, pages 1204–1213, 2021. 1, 2

[5] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023. 1, 2

[6] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 3884–3892. ACM, 2020. 2

[7] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation, 2009. 2

[8] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 758–775. Springer, 2020. 2

[9] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020. 1, 2

[10] Chuang Gan, Hang Zhao, Peihao Chen, David D. Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7052–7061. IEEE, 2019. 2

[11] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 1, 2

[12] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 1, 2, 4, 5, 6, 7

[13] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 1, 2, 3, 4, 6

[14] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 1

[15] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*, 2021. 2

[16] Z. Ghahramani. Factorial hidden markov models. *Machine Learning*, 29, 1997. 2

[17] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 4

[18] Rongzhi Gu, Jian Wu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu. End-to-end multi-channel speech separation. *arXiv preprint arXiv:1905.06286*, 2019. 2

[19] Wang-Li Hao, Zhaoxiang Zhang, and He Guan. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6886–6893. AAAI Press, 2018. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[21] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5491–5500. IEEE, 2019. 2

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6231–6241. IEEE, 2019. 2

[23] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 3

[24] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *CoRR*, abs/2209.15352, 2022. 2

[25] Jonathan Le Roux and Emmanuel Vincent. Consistent wiener filtering for audio source separation. *IEEE signal processing letters*, 20(3):217–220, 2012. 2

[26] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *CVPR*, 2021. 1, 2

[27] D. Li, T. R. Langlois, and C. Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 2018. 2

[28] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *CoRR*, abs/2302.02088, 2023. 2

[29] Yan-Bo Lin and Yu-Chiang Frank Wang. Exploiting audio-visual consistency with partial supervision for spatial audio generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2056–2063. AAAI Press, 2021. 1, 2

[30] Rui Lu, Zhiyao Duan, and Changshui Zhang. Listen and look: Audio–visual matching assisted speech source separation. *IEEE Signal Processing Letters*, 25(9):1315–1319, 2018. 1, 2

[31] Andrew Luo, Yilun Du, Michael J. Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, 2022. 2

[32] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *ICASSP*, pages 696–700. IEEE, 2018. 2

[33] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019. 2

[34] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 275–285. IEEE, 2021. 2

[35] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In *NeurIPS*, 2022. 2

[36] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIX*, volume 13699 of *Lecture Notes in Computer Science*, pages 551–569. Springer, 2022. 2

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[38] Juan F. Montesinos, Venkatesh S. Kadandale, and Gloria Haro. Vovit: Low latency graph-based audio-visual voice separation transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697 of *Lecture Notes in Computer Science*, pages 310–326. Springer, 2022. 2

[39] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 1, 2

[40] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8427–8436. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[41] 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

[42] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 801–816. Springer, 2016. 2

[43] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3347–3356, 2022. 2

[44] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, volume 12365 of *Lecture Notes in Computer Science*, pages 292–308. Springer, 2020. 2

[45] Xinyuan Qian, Qi Liu, Jiadong Wang, and Haizhou Li. Three-dimensional speaker localization: Audio-refined visual scaling factor estimation. *IEEE Signal Processing Letters*, 28:1405–1409, 2021. 1, 2

[46] Xinyuan Qian, Qiquan Zhang, Guohui Guan, and Wei Xue. Deep audio-visual beamforming for speaker localization. *IEEE Signal Processing Letters*, 29:1132–1136, 2022. 1, 2

[47] 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 *ISMIR*, 2014. 6

[48] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10483–10492. IEEE, 2022. 2

[49] Tanzila Rahman, Mengyu Yang, and Leonid Sigal. Tribert: Full-body human-centric audio-visual representation learning for visual sound separation. *CoRR*, abs/2110.13412, 2021. 2

[50] Lord Rayleigh. On our perception of the direction of a source of sound. *Proceedings of the Musical Association*, 2:75–84, 1875. 4

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 3

[52] N. Roman, D. L. Wang, and G. J. Brown. Speech segregation based on sound localization. In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, 2001. 2

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5

[54] S. T. Roweis. One microphone source separation. In *International Conference on Neural Information Processing Systems*, 2001. 2

[55] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4358–4366. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[56] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, pages 2745–2754, 2021. 1, 2, 3, 4, 6, 7

[57] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 252–268. Springer, 2018. 2

[58] Thanh-Dat Truong, Chi Nhan Duong, The De Vu, Hoang Anh Pham, Bhiksha Raj, Ngan Le, and Khoa Luu. The right to talk: An audio-visual transformer approach. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1085–1094. IEEE, 2021. 2

[59] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 1, 2

[60] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 368–385. Springer, 2022. 1, 2

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4

[62] T. 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

[63] Zhong-Qiu Wang, Jonathan Le Roux, and John R. Hershey. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 1–5. IEEE, 2018. 2

[64] Zhong-Qiu Wang and DeLiang Wang. Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(2):457–468, 2019. 2

[65] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6291–6299. IEEE, 2019. 2

[66] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *ICCV*, 2019. 1, 2

[67] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *ICCV*, 2021. 1, 2

[68] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, and Fil Alleva. Multi-microphone neural speech separation for farfield multi-talker speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 5739–5743. IEEE, 2018. 2, 5

[69] Wen Zhang and Jie Shao. Multi-attention audio-visual fusion network for audio spatialization. In Wen-Huang Cheng, Mohan S. Kankanhalli, Meng Wang, Wei-Ta Chu, Jiaying Liu, and Marcel Worring, editors, *ICMR '21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021*, pages 394–401. ACM, 2021. 2

[70] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 1, 2

[71] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 1, 2, 4, 6, 7

[72] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016. 2

[73] Dongzhan Zhou, Xinchi Zhou, Di Hu, Hang Zhou, Lei Bai, Ziwei Liu, and Wanli Ouyang. Sepfusion: Finding optimal fusion structures for visual sound separation. In *AAAI*, 2022. 1, 2

[74] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020. 1, 2, 6, 7

[75] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3550–3558. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[76] Hao Zhu, Mandi Luo, Rui Wang, Aihua Zheng, and Ran He. Deep audio-visual learning: A survey. *Int. J. Autom. Comput.*, 18(3):351–376, 2021. 1

[77] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation using cascaded opponent filter network. In *ACCV*, 2020. 1, 2

[78] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1289–1299, 2022. 1, 2