# Position-Aware Audio-Visual Separation for Spatial Audio

Yuxin Ye, Wenming Yang
Shenzhen International Graduate School
Tsinghua University, China
{yeyx21, yang.wenming}@mails.tsinghua.edu.cn

Yapeng Tian
Department of Computer Science
The University of Texas at Dallas, USA
yapeng.tian@utdallas.com

## 1. Introduction

Auditory and visual characteristics can convey important semantic and spatial information. The well-known cocktail party problem [1] is a classical task of sound source separation. It aims at separating the target source audio from 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 [14, 17, 19].

Predominant audio-visual separation 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 [2], researches on *audio-visual spatial audio separation* (AVSS) (see Fig. 1) are highly limited. Unlike MAVS, AVSS provides listeners with a more immersive perceptual experience, thus making it a challenging task.

Existing spatial audio-visual works have mainly focused on spatial audio generation [2, 15]. This involves converting standard monaural audio into binaural or ambisonic sounds. Sep-stereo [18] 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 separation task by simultaneously considering **what and where** the sounding object is. In an effort to overcome current limitations, we introduce a new position-aware audio-visual separation method for spatial audio. 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? We consider the inter-microphone phase difference (IPD) [13], which 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.

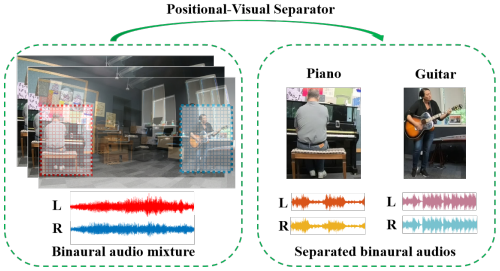



Figure 1. Our approach can separate individual binaural sounds for sounding objects (piano and guitar) from a binaural audio mixture.

Additionally, to leverage the correlation between monaural and binaural channels, we employ a pre-trained separator. 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 our approach.

## 2. Overview of Proposed Approach

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.

As depicted in Fig. 2, our training architecture consists of four parts. 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 [12], 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. Both features are performed attention-based fusion.

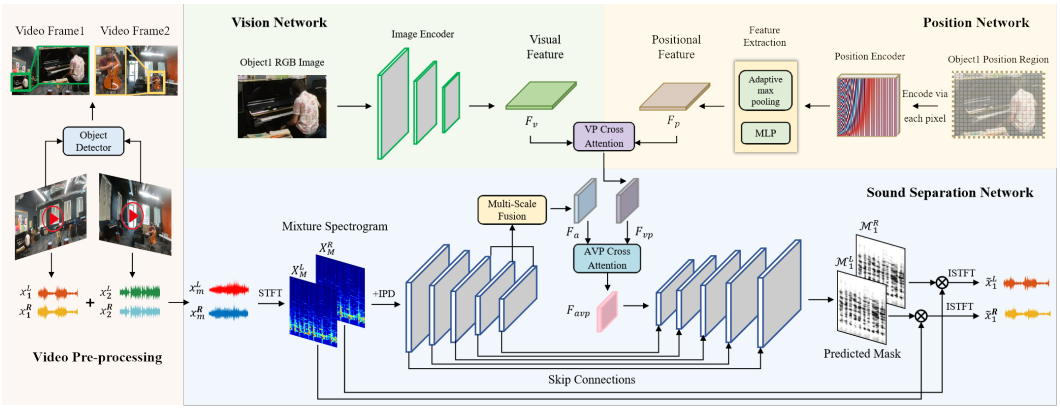



Figure 2. The proposed architecture. Video pre-processing includes object detection and source mixing. Vision extraction network encodes the visual regions of detected objects; position network simultaneously encodes regional coordination features; VP Cross Attention module aggregates visual and positional representations; sound separation network exploits the fused feature as guidance to separate stereo.

The binaural audio mixture is transformed into the time-frequency domain and passed to an encoder-decoder sound separation network. All features are fused through a multi-scale attention-based fusion module. Finally, we obtain the estimated audios $\hat{x}_n^L(t), \hat{x}_n^R(t)$ of individual objects.

## 2.1. Vision-Position Embedding Framework

**Vision network:** In order to precisely localize the audible objects, we choose the widely used detector Faster R-CNN [10] trained on labeled Open Images dataset [6] used in [3, 12]. 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 $\{(M_V^n, C_V^n, P_V^n, B_V^n)\}_{n=1}^N = \text{FRCNN}\ (V)$, 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 visual feature extraction, objects are resized and passed to a pre-trained ResNet-18 [5] network. We obtain the visual embedding $F_v \in \mathbb{R}^{C_v \times H \times W}$ before the last fully-connected layer, where $H = H_b/32, W = W_b/32, C_v = 512$ denote the feature map size and channel dimension of $F_v$, respectively. $H_b, W_b$ represent the resized image shape.

**Position network:** Going beyond the general MAVS strategy, we leverage positional representations as a new constituent modality and demonstrate the association with spatial distribution embedded in spatial audio. Specifically, we leverage a 2D positional encoding for spatial coordinates of detected objects [7], forcing our positional network to approximate a higher frequency function and guide spatial audio separation. The function $\gamma(\cdot)$ represents a mapping function from low-dimensional space into a higher one,

$$\gamma(x, y) = \big(\sin(2^0\pi x), \cos(2^0\pi x), \sin(2^0\pi y), \cos(2^0\pi y), \ldots, \sin(2^{D-1}\pi x), \cos(2^{D-1}\pi x), \sin(2^{D-1}\pi y), \cos(2^{D-1}\pi y)\big) \quad (1)$$

This sinusoidal function is applied simultaneously to 2D coordination in $(\mathbf{x}, \mathbf{y})$ (which are normalized to range $[-1, 1]$ [7]). In our experiments, we set $D = 16$ for $\gamma(x, y)$ to encode each pixel in the detected position region relative to video frames. 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:** 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. $CMA(M, N, N)$ performs cross-modal attention over the first and second axes of $N$,

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

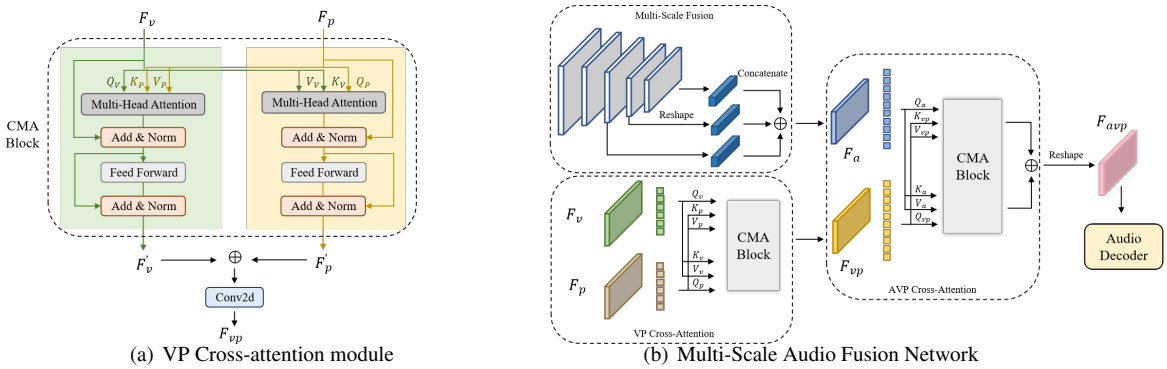
where $M_Q$ is the query vector of $M$, $N_K$, $N_V$ are key and value vectors of $N$. $MHA$, $FFN$, $LN$ denote the multi-head attention, feed-forward layer, and layer normalization, respectively. Then the visual-positional feature $F_{vp} \in \mathbb{R}^{C_{vp} \times H \times W}$ can be obtained after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

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

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

## 2.2. Multi-modal Sound Source Separation

**Audio embedding network:** The time-discrete binaural audio waveform are first converted to time-frequency spectrograms $X_m^L, X_m^R$ through STFT [4] transform. In terms of spatial audio, sound source locations are determined by time differences between the sound sources reaching each ear [2, 9], which can be measured by the inter-microphone phase difference (IPD) between the left and right channels.




(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 multi-scale fusion and VP/AVP cross-attention modules.

The IPD can be calculated as $IPD = cos(\angle X_m^L - \angle X_m^R)$, where $\angle$ represents the phase angle of the complex spectrogram. We concatenate both log power spectra and IPD features and obtain the audio embedding of size $2 \times T \times F$, 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 [11] backbone is used for encoding features into semantic representations. At the bottleneck, the multi-scale audio fusion network performs multi-modal modeling over the audio, vision, and position features.

**Multi-scale audio fusion network:** To fuse the spectra-spatial audio feature and the visual-positional representations, we put forward a multi-scale audio fusion network visualized in Fig. 3 (b). 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.

$$F_a = f_{Concat}(F_a^N, F_a^{N-1}, ..., F_a^{N-i}), i = 0, 1, 2 \quad (4)$$

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,

$$F_{avp} = f_2(CMA(F_a, F_{vp}, F_{vp}) \oplus f_1(CMA(F_{vp}, F_a, F_a))) \quad (5)$$

where $f_1(\cdot)$ 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 upsample layers of U-Net. Finally, the predicted magnitude binary masks $\hat{\mathcal{M}}_n^L, \hat{\mathcal{M}}_n^R$ are multiplied by the original mixture spectrogram $X_m^L, X_m^R$ to produce the final estimation of output spectrograms and estimated audios $\hat{x}_n^L(t), \hat{x}_n^R(t)$.

$$\begin{aligned} \hat{x}_n^B(t) &= ISTFT(\hat{\mathcal{M}}_n^B \odot X_m^B) \\ \mathcal{M}_{gt,n}^B(u,v) &= [X_n^B(u,v) \geq X_m^B(u,v)] \end{aligned} \quad (6)$$

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 framework training objective by minimizing a combination of both frequency and time reconstruction losses. We measure the linear combination between the L1 and L2 losses over the predicted ratio masks and ground-truth. 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,

$$\begin{aligned} \mathcal{L}_{freq} &= \sum_{n=1}^{N} \sum_{B \in L,R} \|\hat{\mathcal{M}}_n^B - \mathcal{M}_n^B\|_1 + \alpha \|\hat{\mathcal{M}}_n^B - \mathcal{M}_n^B\|_2 \\ \mathcal{L}_{time} &= \sum_{n=1}^{N} \sum_{B \in L,R} \|\hat{x}_n^B(t) - x_n^B(t)\|_1 \\ \mathcal{L}_{binaural} &= \mathcal{L}_{freq} + \beta \mathcal{L}_{time} \end{aligned} \quad (7)$$

## 2.3. Transfer Learning by 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 [16] to perform transfer learning. This dataset includes more instrument categories and videos, which can mitigate the difficulty of AVSS. Some of the instrument types overlap, which makes the sound separation between similar acoustic characteristics mutually beneficial. Similar to the training process in Fig. 2, we pre-process the videos in MUSIC dataset. Both audio and visual features are fused 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. 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 binaural dataset.

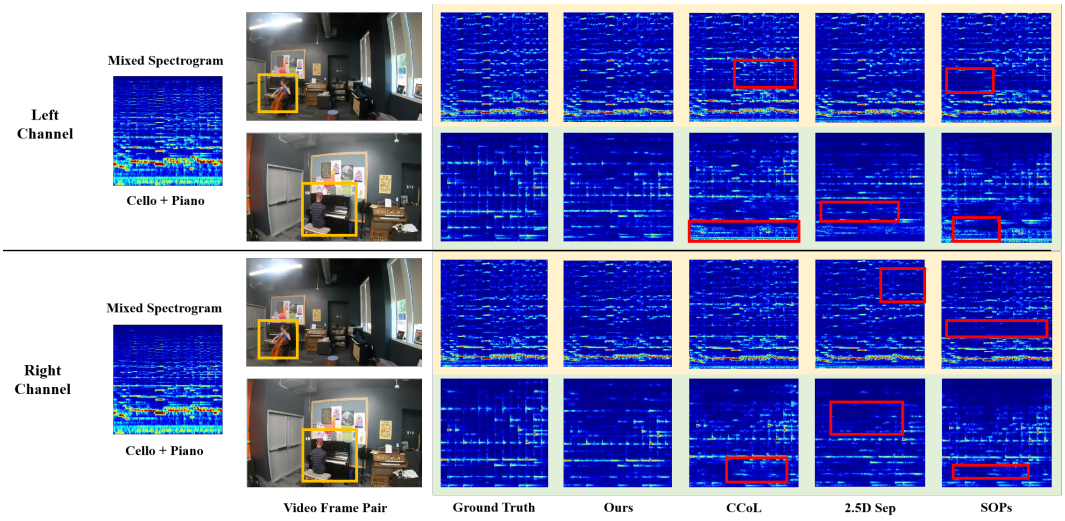



Figure 4. A set of solo separation results on FAIR-Play test set. Predicted spectrograms of SOTA methods and ours are depicted for both channels. Red boxes illustrate the difference between the predicted spectrogram and the ground truth.

## 3. Experiment and Results

In this work, we train and evaluate the proposed position-aware audio-visual separation framework on the FAIR-Play dataset [2]. It contains 1039 10s solo videos with spatial audios. We randomly split it into train/val/test sets: 728/103/208. Moreover, we evaluate the ability for separating multiple sources by testing 418 duet videos as illustrated in Fig. 1. We adopt the widely-used mir_eval [8] library metrics SIR and SDR to measure the quality of separation.

**Comparison with state-of-the-art:** To evaluate the performance of our framework on audio-visual sound separation, we compare it to two baselines most related to binaural audio separation and generation: 2.5D Separation [2] and Sep-Stereo [18], and recent state-of-the-art methods: SoP [16], Co-separation [3], and CCoL [12]. Since those methods are specialized in MAVS, we take the left and right channels into the network separately for training and evaluation. The SDR and SIR quantitative analysis are illustrated in Tab. 1. The results show that our model outperforms its closest competitor, Sep-Stereo [18], by an obvious superiority of 1.0 dB on SDR and 2.85 dB on SIR for binaural channels. The above MAVS methods mainly utilize appearance-based visual information, which cannot generalize to AVSS. In contrast, our approach considers what and where the object is, thus demonstrating competence for the AVSS task. Specifically, both solo and duet video separation performances are illustrated in Fig. 4. Our separated spectrogram is distinctly and completely restored for both channels compared to SoP, 2.5D-sep, and CCoL.

| Method | Left Channel | | Right Channel | | Average | |
|---|---|---|---|---|---|---|
| | SDR↑ | SIR↑ | SDR↑ | SIR↑ | SDR↑ | SIR↑ |
| SoP [16] | 3.34 | 6.45 | 3.29 | 6.42 | 3.31 | 6.43 |
| 2.5D [2] | 3.85 | 7.24 | 3.73 | 7.44 | 3.77 | 7.32 |
| Co-Sep [3] | 4.25 | 7.49 | 4.43 | 7.64 | 4.34 | 7.56 |
| Sep-Stereo [18] | 5.05 | 7.23 | 5.01 | 7.45 | 5.03 | 7.34 |
| CCoL [12] | 4.82 | 8.24 | 4.97 | 8.36 | 4.89 | 8.30 |
| 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.

| 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 |
| | ✗ | ✗ | ✓ | 4.79 | 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 | 9.22 |
| | ✗ | ✗ | ✓ | 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 2. Ablation study of two benchmarks on FAIR-Play test set.

**Ablation studies:** We conduct ablation study to evaluate the effectiveness of IPD, position representation, and monaural transfer learning in our model. Different configurations are conducted for ablation studies on FAIR-Play dataset. We choose SoP and 2.5D-sep as baselines for verifying the versatility on benchmark applications. Tab. 2 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 the model's generalization ability.

**Acknowledgement:** This work was partly supported by the National Natural Science Foundation of China(Nos. 62171251&62311530100) and the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen(Nos.JSGG20211108092812020&CJGJZD20210408092804011).

## References

[1] 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.

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

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, pages 2745–2754, 2021.

[13] 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.

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

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

[16] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.

[17] 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.

[18] 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.

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