Cyclic Learning for Binaural Audio Generation and Localization

Zhaojian Li, Bin Zhao, and Yuan Yuan*
School of Artificial Intelligence, OPtics and ElectroNics, Northwestern Polytechnical University, China
zj_lee@mail.nwpu.edu.cn, {binzhao111, y.yuan1.ieee}@gmail.com

Abstract

Binaural audio is obtained by simulating the biological structure of human ears, which plays an important role in artificial immersive spaces. A promising approach is to utilize mono audio and corresponding vision to synthesize binaural audio, thereby avoiding expensive binaural audio recording. However, most existing methods directly use the entire scene as a guide, ignoring the correspondence between sounds and sounding objects. In this paper, we advocate generating binaural audio using fine-grained raw waveform and object-level visual information as guidance. Specifically, we propose a Cyclic Locating-and-UPmixing (CLUP) framework that jointly learns visual sounding object localization and binaural audio generation. Visual sounding object localization establishes the correspondence between specific visual objects and sound modalities, which provides object-aware guidance to improve binaural generation performance. Meanwhile, the spatial information contained in the generated binaural audio can further improve the performance of sounding object localization. In this case, visual sounding object localization and binaural audio generation can achieve cyclic learning and benefit from each other. Experimental results demonstrate that on the FAIR-Play benchmark dataset, our method is significantly ahead of the existing baselines in multiple evaluation metrics (STFT↓: 0.787 vs. 0.851, ENV↓: 0.128 vs. 0.134, WAV↓: 5.244 vs. 5.684, SNR↑: 7.546 vs. 7.044).

1. Introduction

Hearing and vision are the most important ways for humans to perceive the world. We are capable of associating different sounds with objects through unconscious learning [19]. Simultaneously, we can easily decode the spatial properties of the sound and locate the sounding object in visual scenes, since the acquired sound is binaural [1, 42]. Therefore, the semantic and spatial information conveyed by hearing plays an important role in visual scene perception.

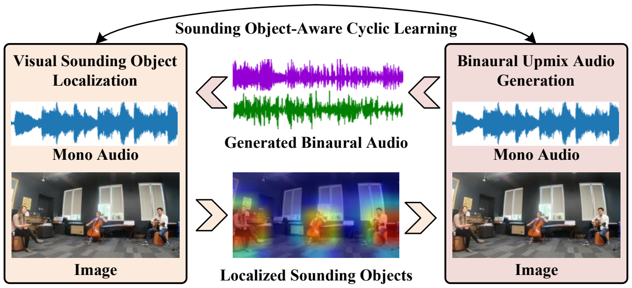
Sounding Object-Aware Cyclic Learning
Visual Sounding Object Localization
Mono Audio
Image
Generated Binaural Audio
Localized Sounding Objects
Binaural Upmix Audio Generation
Mono Audio
Image

Figure 1. Illustration of our core idea on establishing a link between visual sounding object localization and binaural upmix audio generation. Building upon this connection, the two tasks achieve cyclic learning and promote each other.

Visual sounding object localization aims to locate sounding objects in the visual scene. In the case of mono audio, sounding object localization mainly relies on semantic alignment between audio and visual modalities [15, 29, 37]. Binaural audio is recorded through an artificial head prosthesis, which closely simulates human auditory mapping. Compared with mono audio, the rich spatial information contained in binaural audio can further improve sounding object localization [33]. However, the recording of binaural audio requires professional equipments and knowledge, resulting in higher costs and limiting its applications.

The semantic and spatial information contained in visual modality is also beneficial to restore the binaural spatial sense of mono audio [31]. Therefore, it is a promising way to leverage visual modalities to guide the generation model to upmix mono audio into binaural audio [6, 22, 24]. However, existing binaural audio generation methods are not capable of generating high-fidelity results, they cannot well preserve object information and accurately model sounding object-driven generative expressions. Most existing methods [6, 7, 24, 26, 45, 50] directly learn scene drivers to reconstruct the spatial sense of sounds. Although the scene contains semantic and spatial information of the sounding objects, it also brings interference of silent objects and background. These ambiguous information may lead to

[*] Corresponding author

failure of binaural audio generation.

As mentioned above, visual sounding object localization can identify specific sounding objects for binaural audio generation. Meanwhile, the generated binaural audio adds spatial information on the basis of semantics, which can further improve the performance of sounding object localization [14, 17, 38]. Inspired by this point, we make the first attempt in this field, that is, combining the tasks of binaural audio generation with visual sounding object localization and establishing a cyclic learning framework. Specifically, we first exploit the semantic information of mono audio to localize visual sounding objects. Then, we learn binaural audio generation with the help of localized sounding objects. To this end, we propose an object-aware upmixing model to generate binaural audio using mono waveform and explicit objects as guidance. Unlike most spectrogram-based methods, the proposed approach can directly generate binaural waveform without spectrogram conversion.

Obviously, a good localization model is capable of mitigating interference from silent objects and background noise and improving generation results. However, unsupervised visual sounding object localization without the manual annotations is challenging [3, 28, 39]. The model may tend to associate object sounds with backgrounds when similar backgrounds and objects often co-occur [35]. To tackle this issue, we propose an unsupervised sounding object localization model, which consists of two modules, *i.e.*, Object-Scene Awareness (OSA) module and Semantic-Spatial Mining (SSM) module. The OSA module simultaneously computes the gap between visual and audio modalities at the object and scene levels. The object-level loss can locate a more precise object range, while the scene-level loss can well distinguish potential sounding objects from the background. To further utilize binaural audio to improve localization results, we design a SSM module to simultaneously mine the semantic and spatial information contained in binaural audio. This enables cyclic learning between visual sounding object localization and binaural audio generation, improving both the localization and generation performance, as illustrated in Fig. 1. To summarize, our main contributions can be described as follows:

(1) We analyze the relationship between sounding object localization and binaural audio generation, and present a cyclic learning strategy achieving mutual benefits of them within a unified framework.

(2) We propose an object-aware upmixing model to generate binaural audio, meanwhile, design an OSA module and a SSM module to improve the localization performance of binaural audio.

(3) Experimental results demonstrate that the proposed approach can make sounding object localization and binaural audio generation promote each other and accomplish state-of-the-art performance.

2. Related Work

2.1. Visual Sounding Object Localization

Visual sounding object localization aims to locate the regions in an image corresponding to the sounds. The result of visual sounding object localization is usually obtained by computing the similarity matrix of the audio and visual feature maps, which is expressed as a heat map [25, 36, 51]. In pioneering works, Senocak *et al.* [35] replaced visual-related audio to explore the correspondence between audio and visual modalities to achieve sounding object localization. Hu *et al.* [12] introduced the clustering process into audiovisual learning and used the center distance of audio-visual modalities as a supervision information to sort the audiovisual features. Qian *et al.* [32] presented a multi-task method to perform audiovisual classification and correspondence learning simultaneously, and then used class activation maps [2, 4] to locate category-specific elements. In addition, Hu *et al.* [13] presented a two-stage approach that uses audiovisual semantics learned under a single sound source to help localize multiple sound sources. In recent years, inspired by the localization of human binaural systems, the use of binaural audio for sounding object localization has received widespread attention [33, 44]. Wu *et al.* [44] proposed a BAVNet for binaural localization by extracting and fusing features from images, binaural audio, left channel, and right channel. Kranthi *et al.* [33] introduced an audio localization model to locate the sounding object by extracting the interaural level difference and interaural phase difference [20, 41] of the binaural spectrogram. The binaural audio localization method achieves good localization results. However, they rely heavily on extensive manual annotations.

2.2. Binaural Audio Generation

Traditional binaural audio generation methods are often modeled as a linear invariant system, which can produce plausible auditory perceptions through simple mathematical modeling [5, 8, 46]. However, real sound propagation is non-linear, resulting in traditional methods that are consistently unable to compete with recorded binaural audio [21]. Furthermore, traditional methods rely on personalized head-related transfer function and head tracking, resulting in limited flexibility [9, 16, 43]. In recent years, binaural audio generation using deep learning techniques has received widespread attention [6, 7, 23, 24, 45, 49]. Gao *et al.* [6] proposed a UNet-like framework that converts mono audio to binaural audio by connecting visual modalities into the decoder. Sound source separation and binaural audio generation have similar mixing-separating paradigms [27, 40, 48]. Zhou *et al.* [50] combined them in a comprehensive framework to boost the performance of binaural generation model. Xu *et al.* [45] proposed two mappings

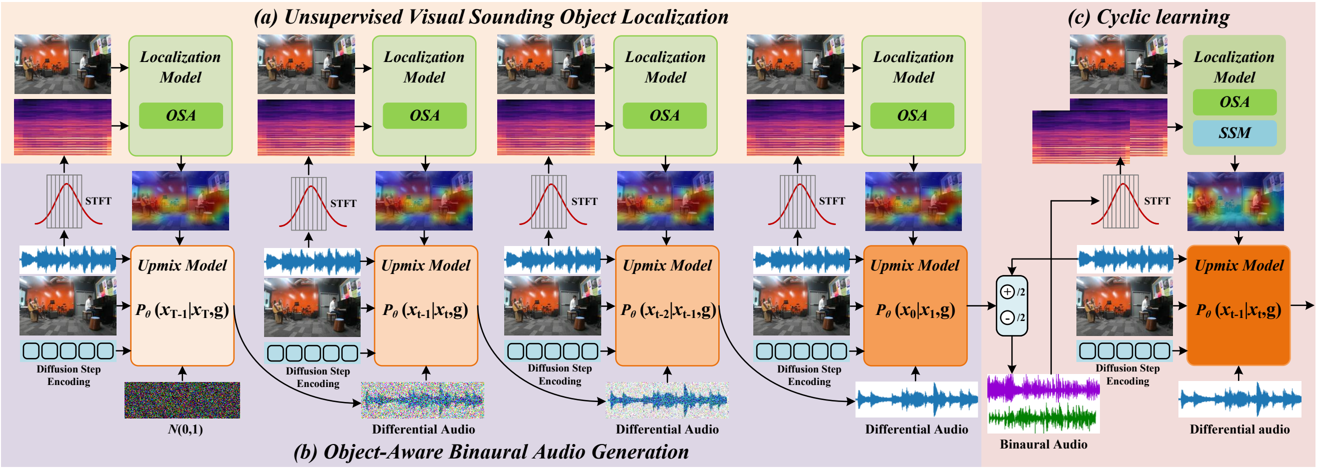
(a) Unsupervised Visual Sounding Object Localization
(c) Cyclic learning
Localization Model
OSA
SSM
STFT
Upmix Model
P_θ (x_{T-1}|x_T,g)
P_θ (x_{t-1}|x_t,g)
P_θ (x_{t-2}|x_{t-1},g)
P_θ (x_0|x_1,g)
Diffusion Step Encoding
N(0,1)
Differential Audio
⊕/2
⊖/2
Binaural Audio
Differential audio
(b) Object-Aware Binaural Audio Generation

Figure 2. Illustration of the cyclic locating-and-upmixing framework. (a) Firstly, the localization model utilizes the semantic information of mono audio to localize sounding objects. (b) Then, the generation model learns object-aware binaural audio generation using the localized sounding objects. (c) Finally, the spatial information in binaural audio is mined to further improve the localization of sounding objects.

to link monaural and binaural as well as visual and spatial locations. Finally, head-related impulse responses [18] are used to generate binaural audio. Li *et al.* [23] proposed a shared vision-guided generative adversarial method to generate binaural audio. The visual modality provides guidance information for the generator while providing reference information for the discriminator. Garg *et al.* [7] presented a multi-task framework to synthesize binaural audio, which is guided by decomposed geometric cues in visual modalities. However, this multi-cue collection is only represented by a shared visual feature without a specific sounding object.

Compared with these methods, our approach clarifies the sounding object-driven diffusion generation paradigm for the first time, which organically combines sounding object localization with binaural audio generation. We significantly surpass existing methods on numerous metrics and show that combining visual sounding object localization and binaural audio generation is a win-win situation.

3. Methodology

3.1. Problem Definition

Given randomly sampled mono audio A_m of length L with positive image sequences V_p, and negative image sequences V_n, the goal of locating-and-upmixing model is to generate binaural audio $A_b = (A_b^l, A_b^r)$ associated with audiovisual information.

Mono audio is obtained by mixing the left channel audio and right channel audio, while the differential audio is obtained by subtracting:

$$A_m = A_b^l + A_b^r, A_d = A_b^l - A_b^r. \tag{1}$$

Then, our model generates differential audio guided by the localized sounding objects:

$$\hat{A}_d = U(L(A_m, V_p, V_n), A_m, V_p), \tag{2}$$

where L and U represent the localization and upmixing models, respectively. Next, the generated binaural audio is:

$$\hat{A}_b^l = \frac{A_m + \hat{A}_d}{2}, \hat{A}_b^r = \frac{A_m - \hat{A}_d}{2}. \tag{3}$$

The mono audio can be replaced by the generated binaural audio, and Eq. (2) can be reformulated as:

$$\widehat{A}_d = U(L(\hat{A}_b^l, \hat{A}_b^r, V_p, V_n), A_m, V_p). \tag{4}$$

The localization and generation procedures of our model can be illustrated in Fig. 2.

3.2. Visual Sounding Object Localization Model

In this section, we describe the data sampling, object-scene awareness module, and semantic-spatial mining module of the visual sounding object localization procedure in detail, as illustrated in Fig. 3.

Data Sampling: Given several videos $\{(V_i, A_{bi}^l, A_{bi}^r)\}_{i=1}^N$ consisting of image sequences V_i and corresponding binaural audio tracks (A_{bi}^l, A_{bi}^r). The set of positive and negative samples can be expressed as $\{(V_i, V_j, A_{mi}) | A_{mi} = A_{bi}^l + A_{bi}^r, i \neq j; i, j \in N\}$. For simplicity, we use V_p to represent V_i, and V_n to represent V_j.

Object-Scene Awareness Module: We argue that guidance information is typically embodied in the audiovisual correlation of the sounding object during binaural audio generation. Therefore, the OSA module is introduced to refine complex visual scenes into concrete sounding objects. Specifically, we use pre-trained ResNet-18 [10] and VGGish [11] to extract visual and audio features, respectively.

The positive visual features f_{pv} and the negative visual features f_{nv} are combined with audio features f_a^m to obtain positive-negative sample pairs. Then, the visual object feature f'_{pv} and the pseudo-visual object feature f'_{nv} can be obtained through:

$$f'_{pv} = f_{pv} \cdot \sigma\left((f_{pv})^\top \cdot f_a^m\right), \tag{5}$$

and

$$f'_{nv} = f_{nv} \cdot \sigma\left((f_{nv})^\top \cdot f_a^m\right). \tag{6}$$

Next, the distance between positive and negative sample pairs is obtained by

$$(d_+, d_-^{obj}) = (\left\|f'_{pv} - f_a^m\right\|_2, \|f'_{nv} - f_a^m\|_2). \tag{7}$$

Then, the object-level loss can be expressed as:

$$l_{loc}^{obj} = \left\|(D_+, D_-^{obj}) - (0,1)\right\|_2, \tag{8}$$

where, $D_\pm = \frac{\exp(d_\pm)}{\exp(d_+)+\exp(d_-)}$. Here, we refer to visual sounding object localization on negative samples as *pseudo-visual* localization, which allows the model to localize the wrong objects and introduce larger losses. It is similar to trial-and-error learning in reinforcement learning. Furthermore, in addition to the objects in the negative visual sample being irrelevant to the sound, the entire scene is also irrelevant to the sound. Therefore, we globally average pool the spatial visual features f_{nv} to obtain the global visual feature f_{nvg}. The distance between positive and negative sample pairs is obtained by

$$(d_+, d_-^{sce}) = (\left\|f'_{pv} - f_a^m\right\|_2, \|f_{nvg} - f_a^m\|_2). \tag{9}$$

Then, the scene-level loss can be expressed as:

$$l_{loc}^{sce} = \left\|(D_+, D_-^{sce}) - (0,1)\right\|_2. \tag{10}$$

Finally, the total loss of visual sounding object localization in the mono case is:

$$l_{loc} = l_{loc}^{obj} + l_{loc}^{sce}. \tag{11}$$

Semantic-Spatial Mining Module: To simultaneously mine the semantic and spatial information contained in binaural audio, we propose a semantic-spatial mining module. Firstly, we add the left channel features f_a^l and right channel features f_a^r to get the audio features f_a^m in the OSA module. Then, the visual object features f'_{pv} in Eq. (5) can be reformulated as:

$$\begin{aligned} f'_{pv} = f_{pv} \cdot \sigma((f_{pv})^\top \cdot f_a^m \\ +(f_{pv})^\top \cdot \text{Concat}[f_a^l, f_a^r]). \end{aligned} \tag{12}$$

The distance between positive and negative sample pairs is obtained by

$$\begin{aligned} (d_+, d_-^{spa}) = (\left\|f'_{pv} - \text{Concat}[f_a^l, f_a^r]\right\|_2, \\ \left\|f'_{pv} - \text{Concat}[f_a^r, f_a^l]\right\|_2). \end{aligned} \tag{13}$$

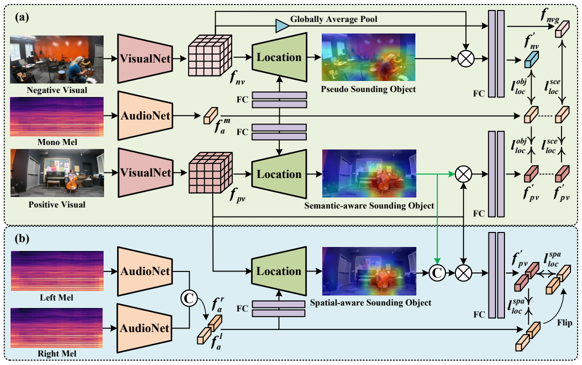

Figure 3. (a) In the mono case, the OSA module locates the sounding object based on audiovisual semantics. (b) In the binaural case, the SSM module further improves the localization of sounding objects by mining spatial information.

The spatial-level loss can be expressed as:

$$l_{loc}^{spa} = \left\|(D_+, D_-^{spa}) - (0,1)\right\|_2. \tag{14}$$

Finally, Eq. (11) can be reformulated in the binaural case as:

$$l_{loc}^* = l_{loc}^{obj} + l_{loc}^{sce} + l_{loc}^{spa}. \tag{15}$$

3.3. Object-Aware Upmix Model

The detailed structure of binaural upmix model is illustrated in Fig. 4. It is built on the diffusion model and is mutually facilitated with the visual sounding object localization model through our cyclic learning strategy. We make two layers of 1D convolution to extract waveform features, the output features are concatenated with visual features and localized object features and then fused through a convolutional layer. In this section, x is used to represent A_d for simplicity.

Diffusion Procedure: Define the waveform distribution as $q(x_0)$, and sample $x_0 \sim q(x_0)$. The diffusion procedure is a fixed-parameter Markov chain, which converts x_0 into the latent x_T in T steps:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{16}$$

At each stage $t \in [1, T]$ of diffusion, in accordance with a variance schedule $\beta = \{\beta_1, \dots, \beta_T\}$, a smidgeon of Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is added to x_{t-1} to produce x_t:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\epsilon). \tag{17}$$

Next, $q(x_t|x_0)$ can be computed by:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\epsilon), \tag{18}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, $\alpha_t := 1 - \beta_t$.

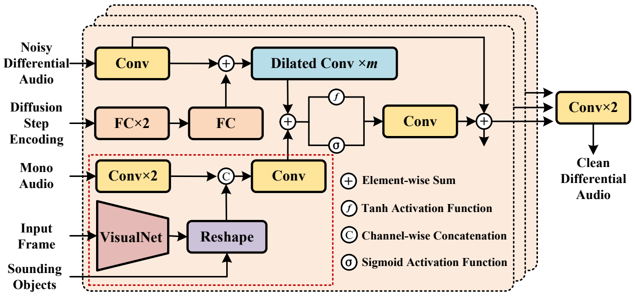

Figure 4. Illustration of binaural upmix model. Conv and FC represent 1D convolutional and fully connected layers, respectively.

Fusion Procedure: The fusion procedure integrates waveform features, visual features, and sounding object features to provide guidance information for the reverse procedure. The length of the waveform feature f_w extracted by 1D dimensional convolution layers is L. The visual features f_{pv} are repeated L times after channel reduction and flattening to obtain new visual features f_{rv}. Then, the sounding object features f'_{pv} are also repeated L times to obtain new features f'_{rv}. Finally, the guidance information g is obtained by fusing these features through a convolutional layer:

$$g = \text{Conv}(\text{Concat}[f_w, f_{rv}, f'_{rv}]) \tag{19}$$

Reverse Procedure: The reverse procedure is a Markov chain with learnable parameters θ from Gaussian noise $p(x_T) \sim \mathcal{N}(0, \mathbf{I})$ to clean waveform x_0. We approximate the reverse distribution $q(x_{t-1}|x_t)$ by a neural network of parameters θ with guidance information g:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, g), \sigma_t^2 \mathbf{I}), \tag{20}$$

where the variance σ_t^2 is predefined as $\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. Then, the full reverse procedure can be described as:

$$p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, g). \tag{21}$$

Training: To learn θ, we minimize a variational bound:

$$\begin{aligned} &\mathbb{E}_{q(x_0)}[-\log p_\theta(x_0)] \geq \\ &E_{q(x_0, x_1, \ldots, x_T)}[\log q(x_{1:T}|x_0) - \log p_\theta(x_{0:T})] =: \mathbb{L}. \end{aligned} \tag{22}$$

Then, we optimize a random term of $\mathbb{L}$ through:

$$\mathbb{L}_{t-1} = \mathbb{E}_q[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t, g)\|^2] + \mathcal{C}, \tag{23}$$

where $\mathcal{C}$ is a constant. Eq. (18) can be re-parameterized as:

$$x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{24}$$

and the parameterization is chosen:

$$\mu_\theta(x_t, t, g) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, g)), \tag{25}$$

Eq. (23) can be simplified to:

$$\mathbb{E}_{x_0, \epsilon, t}\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, g)\|^2. \tag{26}$$

Finally, the outputs of the model is $\epsilon_\theta(\cdot)$.

4. Experiments

4.1. Datasets

We implement experiments on two commonly used binaural audio generation datasets. FAIR-Play dataset [6] is a binaural audio dataset recorded indoors through a dummy head and contains corresponding vision. The dataset contains 1871 videos and provides 10 different splits. We compute the average of 10 splits as the final result. YT-Music dataset [30] is comprised of 397 music videos with different durations collected from YouTube. This dataset contains 360° videos of indoor and outdoor scenes. We follow [6] to transform first-order ambisonic audios into binaural audios.

4.2. Baselines and Metrics

We compare our approach with several state-of-the-art methods, including weakly semi-supervised method: L2BNet[33], autoencoder-based method: MONO2BINAURAL [6], multi-tasking-based methods: APNet [50] and Sep-stereo [50], attention-based methods: Main network [47] and Complete network [47], GAN-based method: SAGM [23]. Mono-Mono represents fake binaural audio created by channel duplication of mono audio. We use STFT Distance, Envelope (ENV) Distance, Wave L2 (WAV$\times 10^{-3}$), Amplitude L2 (AMP), Phase L2 (PHA) [34], and Signal-to-Noise Ratio (SNR) to comprehensively measure the quality of synthetic binaural audio for all methods. See supp. for details.

4.3. Implementation Details

We utilize 1s clips of 10s mono audio and a center frame that matched it to train our model. The binaural audios are resampled to 16kHz, while the input frames are randomly cropped to 448×224. The input with a batch size of 12 is fed into the model, and Adam is used to optimize it. The learning rate of localization model is set as 10^{-5}, while that of the generation model is set to 2×10^{-4}. The model terminates training when the training epochs reache 3000. In inference stage, we employ a sliding window of 0.1s to generate the binaural audio.

4.4. Object-aware Binaural Generation

Quantitative Results. We compare our model with other baseline models in Table 1. The proposed model significantly outperforms other baseline models on multiple metrics for both datasets. The audio generated by our method is superior to existing methods in terms of the spectrum

Method	FAIR-Play Dataset						YT-Music Dataset					
	STFT↓	ENV↓	WAV↓	AMP↓	PHA↓	SNR↑	STFT↓	ENV↓	WAV↓	AMP↓	PHA↓	SNR↑
Mono-Mono	1.155	0.153	7.666	0.267	0.592	5.735	1.853	0.184	7.729	0.287	0.622	3.890
L2BNet [33]	1.028	0.148	-	-	-	-	1.816	0.189	-	-	-	-
MONO2BINAURAL [6]	0.959	0.141	6.496	0.252	0.591	6.232	1.346	0.179	6.337	0.246	0.528	5.008
APNet [50]	0.889	0.136	5.758	0.247	0.585	6.972	1.070	0.148	5.805	0.241	0.521	5.542
Sep-stereo [50]	0.879	0.135	6.526	0.256	0.590	6.422	1.051	0.145	6.323	0.272	0.547	4.779
Main network [47]	0.867	0.135	5.750	0.246	0.583	6.985	1.036	0.144	5.944	0.240	0.514	5.573
Complete network [47]	0.856	0.134	5.787	0.247	0.584	6.959	1.023	0.142	6.313	0.261	0.563	4.873
SAGM [23]	0.851	0.134	5.684	0.243	0.570	7.044	0.875	0.126	5.792	0.240	0.510	5.601
Ours	**0.787**	**0.128**	**5.244**	**0.234**	**0.568**	**7.546**	**0.856**	**0.124**	**5.774**	**0.228**	**0.503**	**5.711**

Table 1. Quantitative results of our method on FAIR-Play and YT-Music datasets.

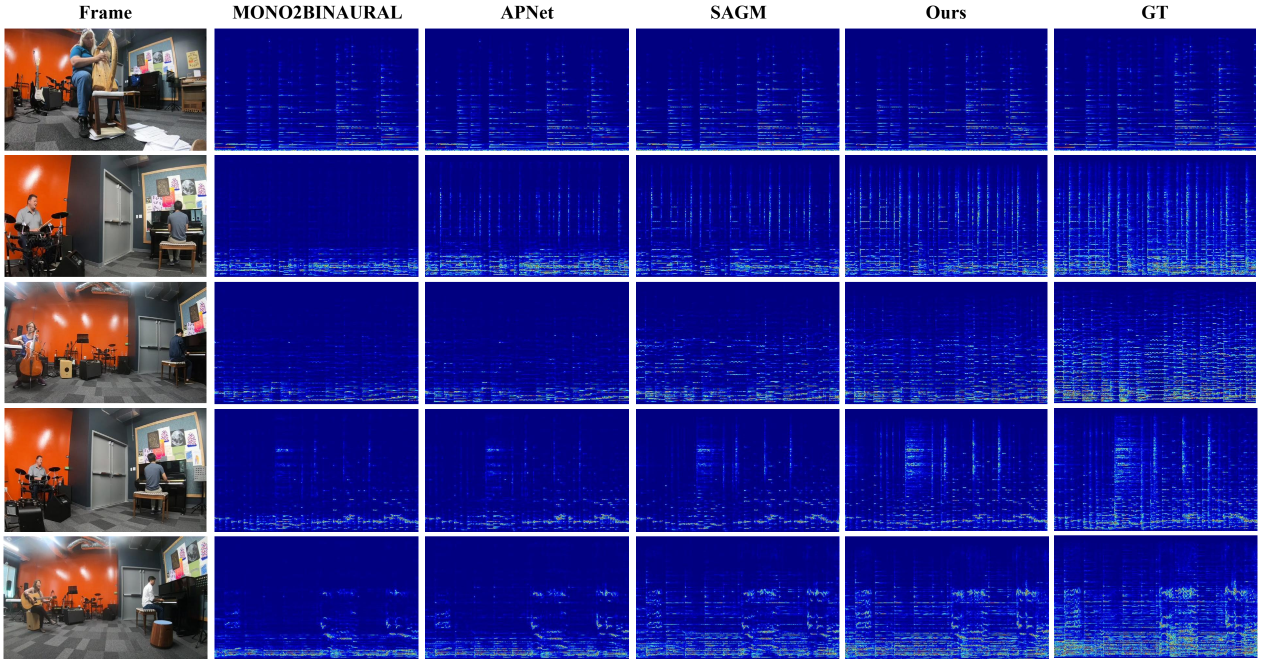

Figure 5. Qualitative results for audio differential spectrograms. The first row represents single-source binaural audio, and rows 2∼5 represent multi-source binaural audio.

(STFT), waveform (WAV) and signal quality (SNR), which shows that the proposed method can generate more realistic binaural audio. Binaural aduio localization is dependent on differences in sound amplitude and phase. The proposed method has the best performance on AMP and PHA. Therefore, our method has better spatial and localization performance. In general, the audio generated by our approach is closer to the recorded binaural audio in terms of data structure and spatial effect.

Qualitative Results. Fig. 5 and Fig. 6 show the qualitative results of binaural audio generation. In Fig. 5, we can see that the spectrogram generated by our approach is closer to the ground truth than other methods. In the case of a single sound source, SAGM seems to be on par with the proposed method. However, SAGM, like other methods, shows fragile time-frequency structure under multiple sound sources. The proposed method has as significant a time-frequency structure as the ground truth. In Fig. 6, we visualize the envelope curve of the binaural signal. It can be seen that our approach is more comparable to the warping of the real waveform envelope. Therefore, the audio generated by our approach has a more realistic spatial sense.

4.5. Visual Sounding Object Localization

Quantitative and Qualitative Results. Table 2, Fig. 7, and Fig. 8 show the quantitative and qualitative results of the localization model. In Table 2, we report the audiovisual distance and classification accuracy on the entire testset. As can be observed, the proposed localization method can automatically associate sounds and objects and demonstrate

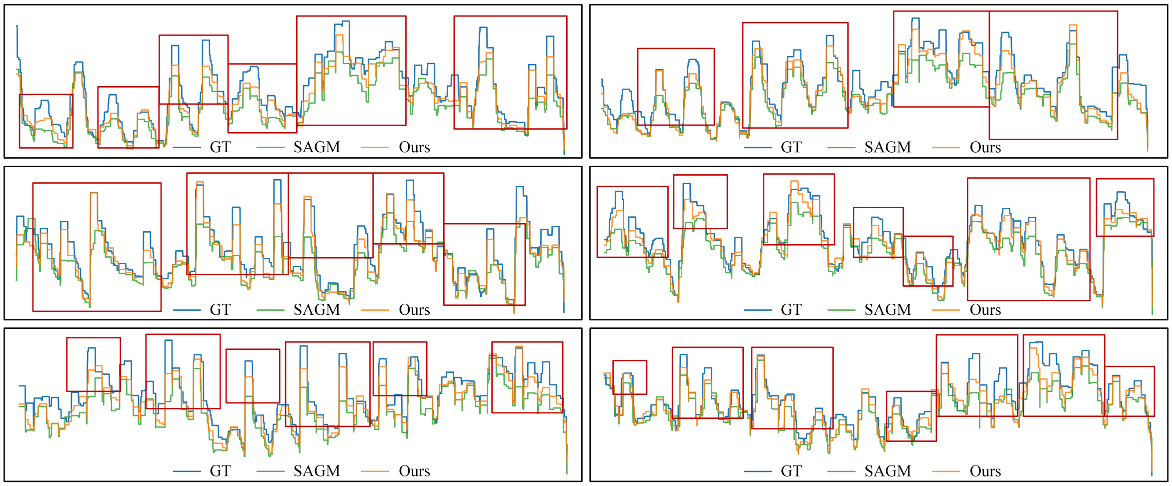

Figure 6. Qualitative results of binaural waveform envelopes. The left and right columns represent the left and right channels, respectively.

Method	Distance↓	Accuracy(%)↑
Object-mismatch	0.065	91.2
Scene-mismatch	0.001	99.9
Object-Scene-mismatch	0.028	96.1

Table 2. The quantitative results of the sounding object localization model on FAIR-Play dataset.

Audio representation type	Distance↓	Accuracy(%)↑
$\{A_m\}$	0.028	96.1
$\{A_d\}$	0.028	96.2
$\{A_l, A_r\}$ w/o SSM	0.024	96.7
$\{A_l, A_r\}$ w/ SSM	0.020	97.4

Table 3. Quantitative results for the types of sounds received by the cyclic framework. In the third row, $f_a^m = f_a^l + f_a^r$.

Method	STFT↓	ENV↓	WAV↓
Audio-Only	1.078	0.148	7.195
Visual-Only	0.786	0.128	5.244
Object-Only	0.828	0.131	5.525
Scene-Only	0.792	0.129	5.283
Visual+Object	0.785	0.128	5.235
Visual+Scene	0.783	0.128	5.212
Ours	0.779	0.128	5.200

Table 4. The ablation results of our approach on the split1 segmentation of FAIR-Play dataset.

excellent sounding object localization. It is worth mentioning that our method achieves nearly 100% accuracy when only counting scene loss l_{loc}^{sce}. This shows that our method can clearly distinguish potential sounding regions from the background. The two row in Fig. 7 intuitively shows the visualization results using scene loss l_{loc}^{sce}. Furthermore, we combine object loss l_{loc}^{obj} and scene loss l_{loc}^{sce} (Full) to achieve coarse-to-fine sounding object localization from potential sounding regions. Fig. 8 shows the sounding object localization and tracking performance of the SSM module. In the second row, the SSM module achieves excellent localization performance by jointly mining semantic and spatial information of binaural audio. In addition, this method has good sounding object tracking performance (the third row).

Cyclic Audio Representation. In order to observe the impact of different audio representations on sounding object localization, we use mono audio, differential audio, and binaural audio as the input of the cyclic framework. $\{A_m\}$ refers to the mixture of left and right channel audio without spatial information. $\{A_d\}$ represents differential audio between left channel audio and right channel audio. $\{A_l, A_r\}$ w/o SSM refers to binaural audio without SSM module. $\{A_l, A_r\}$ w/ SSM is binaural audio with SSM module. Table 3 shows the sounding object localization performance under different audio representations. It can be seen that mono audio performs the worst because it relies solely on sound semantic information for localization. Since binaural audio contains both semantic and spatial information, it has better localization performance. In binaural audio, the proposed SSM module achieves the best localization performance by further mining spatial information.

4.6. Ablation Results

The ablation results of our approach are shown in Table 4. Audio-Only means only mono audio without visual guidance. It can be seen that visual modality is crucial for binaural audio generation. Visual-Only means using the entire visual as guidance without object localization. Object/Scene-

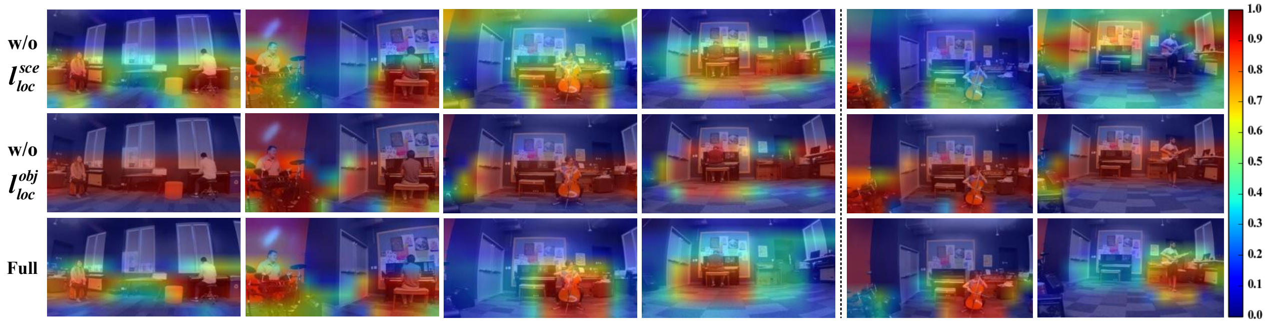

Figure 7. Qualitative results of the OSA module. l_{loc}^{obj} and l_{loc}^{sce} complement each other to achieve better sound source localization.

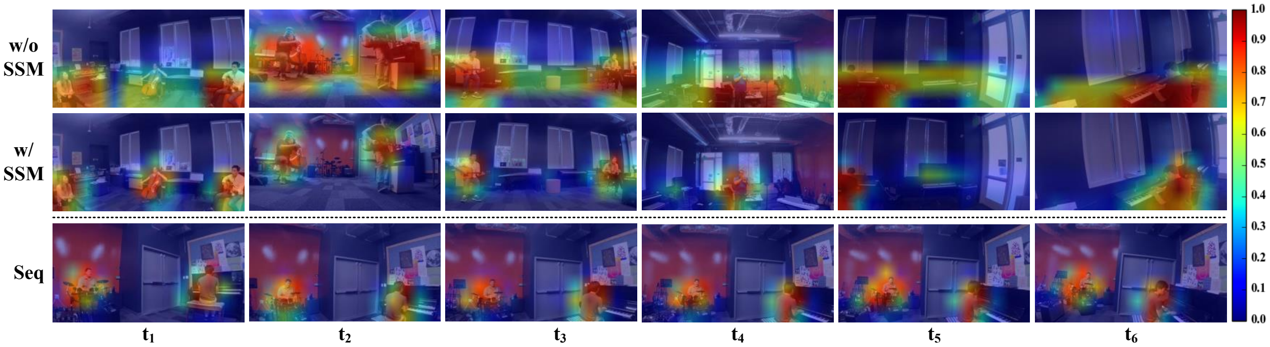

Figure 8. Qualitative and tracking results of the SSM module. The spatial information mined by the SSM module from binaural audio further improves the localization results. At the same time, the proposed method can track moving sounding objects.

Only indicates using localized sounding objects as guidance. It can be seen that the results using only localized sounding objects as guidance are worse than using only full vision. The reason is that the localization model may locate the misplaced object and cause the failure of generation. Scene-Only (l_{loc}^{sce}) is better than Object-Only (l_{loc}^{obj}) because it provides more accurate object information. Visual+Object/Scene represents the fusion of full vision with localized sounding objects. Full vision contains the spatial position information of objects, which can provide supplementary information when localization is insufficient. At the same time, localized sounding objects provide clear object information and avoid interference from irrelevant backgrounds. Therefore, the fusion of localized sounding objects and full vision is necessary. It can be seen that integrated visual guidance is superior to full visual guidance and localized sounding object guidance. This demonstrates the effectiveness of our approach and components.

5. Conclusion

In this paper, we combine visual sounding object localization with binaural audio generation and propose a cyclic learning framework. We introduce visual sounding objects to provide explicit object drivers for binaural audio generation. At the same time, the spatial information contained in binaural audio further improves localization performance. Furthermore, we propose a novel localization model that achieves good audiovisual correlation and localization through joint learning. We demonstrate that our method realizes state-of-the-art synthesis quality under multiple metrics on benchmark datasets, and both tasks benefit from each other through our cyclic learning.

Limitations and Future Work. Currently, our procedure is limited by unsupervised sounding object localization. Sounding object localization improves the performance of the generation model. However, it has to be admitted that the unsupervised method inevitably brings about the loss of object information. Therefore, it is prospective to extend our approach to supervisory scenarios. Apart from that, it is natural to extend our method to bounding box-based object localization. Besides, to alleviate the superstition problem of unsupervised methods, it is also promising to extend our method to semi-supervision.

Acknowledgments. This work was supported in part by the National Science Fund for Distinguished Young Scholars (61825603), the National Key Research and Development Project (2020YFB2103900), the National Natural Science Foundation of China (62106183, 62376222), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Amandine Afonso-Jaco and Brian FG Katz. Spatial knowledge via auditory information for blind individuals: Spatial cognition studies and the use of audio-vr. *Sensors*, 22(13): 4794, 2022. 1

[2] Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 137–146, 2023. 2

[3] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound Localization by Self-Supervised Time Delay Estimation. In *European Conference on Computer Vision*, pages 489–508, 2022. 2

[4] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022. 2

[5] Fabio P Freeland, Luiz WP Biscainho, and Paulo SR Diniz. Efficient HRTF Interpolation in 3D Moving Sound. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, 2002. 2

[6] Ruohan Gao and Kristen Grauman. 2.5D Visual Sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 1, 2, 5, 6

[7] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-Aware Multi-Task Learning for Binaural Audio Generation from Video. *arXiv preprint arXiv:2111.10882*, 2021. 1, 2, 3

[8] Israel D Gebru, Dejan Marković, Alexander Richard, Steven Krenn, Gladstone A Butler, Fernando De la Torre, and Yaser Sheikh. Implicit HRTF Modeling Using Temporal Convolutional Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3385–3389, 2021. 2

[9] William Hartmann. Spatial Relationships Between Interaural Differences in a Room. *The Journal of the Acoustical Society of America*, 152(4):A91–A91, 2022. 2

[10] 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. 3

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN Architectures for Large-Scale Audio Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 131–135, 2017. 3

[12] Di Hu, Feiping Nie, and Xuelong Li. Deep Multimodal Clustering for Unsupervised Audiovisual Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2

[13] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative Sounding Objects Localization via Self-Supervised Audiovisual Matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020. 2

[14] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and Localize: Localizing Sound Sources in Mixtures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022. 2

[15] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric Audio-Visual Object Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22910–22921, 2023. 1

[16] Vivek Jayaram, Ira Kemelmacher-Shlizerman, and Steven M Seitz. HRTF Estimation in the Wild. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 1–9, 2023. 2

[17] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric Deep Multi-Channel Audio-Visual Active Speaker Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 2

[18] Grady Kestler, Shahrokh Yadegari, and David Nahamoo. Head Related Impulse Response Interpolation and Extrapolation Using Deep Belief Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 266–270, 2019. 3

[19] Sid Kouider, Bria Long, Lorna Le Stanc, Sylvain Charron, Anne-Caroline Fievet, Leonardo S Barbosa, and Sofie V Gelskov. Neural dynamics of prediction and surprise in infants. *Nature Communications*, 6(1):8537, 2015. 1

[20] Gyeong-Tae Lee, Sang-Min Choi, Byeong-Yun Ko, and Yong-Hwa Park. HRTF Measurement for Accurate Sound Localization Cues. *arXiv preprint arXiv:2203.03166*, 2022. 2

[21] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, et al. BinauralGrad: A Two-Stage Conditional Diffusion Probabilistic Model for Binaural Audio Synthesis. *Advances in Neural Information Processing Systems*, 35: 23689–23700, 2022. 2

[22] Sijia Li, Shiguang Liu, and Dinesh Manocha. Binaural audio generation via multi-task learning. *ACM Transactions on Graphics*, 40(6):1–13, 2021. 1

[23] Zhaojian Li, Bin Zhao, and Yuan Yuan. Cross-modal Generative Model for Visual-Guided Binaural Stereo Generation. *arXiv preprint arXiv:2311.07630*, 2023. 2, 3, 5, 6

[24] Yan-Bo Lin and Yu-Chiang Frank Wang. Exploiting Audio-Visual Consistency with Partial Supervision for Spatial Audio Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2056–2063, 2021. 1, 2

[25] Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang. Induction Network: Audio-Visual Modality Gap-Bridging for Self-Supervised Sound Source Localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4042–4052, 2023. 2

[26] Francesc Lluís, Vasileios Chatziioannou, and Alex Hofmann. Points2Sound: From Mono to Binaural Audio Using 3D Point Cloud Scenes. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):1–15, 2022. 1

[27] Sagnik Majumder and Kristen Grauman. Active Audio-Visual Separation of Dynamic Sound Sources. In *Proceedings of the European Conference on Computer Vision*, pages 551–569, 2022. 2

[28] Shentong Mo and Pedro Morgado. A Closer Look at Weakly-Supervised Audio-Visual Source Localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022. 2

[29] Shentong Mo and Pedro Morgado. Localizing Visual Sounds the Easy Way. In *Proceedings of the European Conference on Computer Vision*, pages 218–234, 2022. 1

[30] Pedro Morgado, Nuno Vasconcelos, Timothy R. Langlois, and Oliver Wang. Self-Supervised Generation of Spatial Audio for 360° Video. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems*, pages 360–370, 2018. 5

[31] 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 Winter Conference on Applications of Computer Vision*, pages 3347–3356, 2022. 1

[32] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple Sound Sources Localization from Coarse to Fine. In *Proceedings of the European Conference on Computer Vision*, pages 292–308, 2020. 2

[33] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to Binauralize: Audio Spatialization from Visual Sound Source Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1930–1939, 2021. 1, 2, 5, 6

[34] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando De la Torre, and Yaser Sheikh. Neural Synthesis of Binaural Speech From Mono Audio. In *Proceedings of the International Conference on Learning Representations*, 2021. 5

[35] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1605–1619, 2019. 2

[36] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less Can Be More: Sound Source Localization with a Classification Model. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 3308–3317, 2022. 2

[37] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7777–7787, 2023. 1

[38] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022. 2

[39] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 2

[40] Reuben Tan, Arijit Ray, Andrea Burns, Bryan A Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, and Kate Saenko. Language-Guided Audio-Visual Source Separation via Trimodal Consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10575–10584, 2023. 2

[41] Tan-Hsu Tan, Yu-Tang Lin, Yang-Lang Chang, and Mohammad Alkhaleefah. Sound Source Localization Using a Convolutional Neural Network and Regression Model. *Sensors*, 21(23):8031, 2021. 2

[42] Matteo Tomasetti and Luca Turchet. Playing With Others Using Headphones: Musicians Prefer Binaural Audio With Head Tracking Over Stereo. *IEEE Transactions on Human-Machine Systems*, 2023. 1

[43] Xiang Wu, Dumidu S Talagala, Wen Zhang, and Thushara D Abhayapala. Individualized Interaural Feature Learning and Personalized Binaural Localization Model. *Applied Sciences*, 9(13):2682, 2019. 2

[44] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. Binaural Audio-Visual Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2961–2968, 2021. 2

[45] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually Informed Binaural Audio Generation without Binaural Audios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 1, 2

[46] Navid H Zandi, Awny M El-Mohandes, and Rong Zheng. Individualizing Head-Related Transfer Functions for Binaural Acoustic Applications. In *Proceedings of the IEEE International Conference on Information Processing in Sensor Networks*, pages 105–117, 2022. 2

[47] Wen Zhang and Jie Shao. Multi-Attention Audio-Visual Fusion Network for Audio Spatialization. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 394–401, 2021. 5, 6

[48] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The Sound of Pixels. In *Proceedings of the European Conference on Computer Vision*, pages 570–586, 2018. 2

[49] Tao Zheng, Sunny Verma, and Wei Liu. Interpretable Binaural Ratio for Visually Guided Binaural Audio Generation. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2022. 2

[50] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation. In *Proceedings of the European Conference on Computer Vision*, pages 52–69, 2020. 1, 2, 5, 6

[51] Lingyu Zhu and Esa Rahtu. Visually Guided Sound Source Separation and Localization Using Self-Supervised Motion Representations. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1289–1299, 2022. 2