Supervising Sound Localization by In-the-wild Egomotion

Anna Min[2] Ziyang Chen[1] Hang Zhao[2,3] Andrew Owens[1]

[1]University of Michigan [2]Tsinghua University [3]Shanghai Qi Zhi Institute

Abstract

We present a method for learning binaural sound localization using egomotion as a supervisory signal. Over the course of a video, the camera's direction to a sound source will change as the camera moves. We train an audio model to predict sound directions that are consistent with visual estimates of camera motion, which we obtain using traditional methods from multi-view geometry. This provides a weak but plentiful form of supervision that we combine with traditional binaural cues. To evaluate this method, we propose a dataset of real-world audio-visual videos with egomotion. We show that our model can successfully learn from real-world data and that it performs well on sound localization tasks.

1. Introduction

Our sense of hearing allows us to perceive events that are out of sight, such as objects that are distant, occluded, or outside of our narrow field of view. Despite the importance of spatial audio perception, existing methods for stereo sound localization often struggle in real-world settings, such as by being limited to synthetic training data [13], specific binaural cues [12], or specific visual object categories [19].

Recent work has addressed this problem using time-delay estimation [12] and jointly learning camera rotation with sound sources [13]. These approaches either rely on training with simulated data or specific hand-chosen binaural cues and thus can only be applied to limited domains. Natural sound recordings vary widely in ways that are difficult to capture through simulation. For example, properties of the equipment itself, such as the geometry of the stereo microphone pair and their frequency responses, are highly varied, and the sounds that they record are often complex mixtures of different sources that make it difficult to extract a learning signal from.

We address the problem of learning stereo sound localization entirely from unlabeled "in the wild" videos. We exploit the fact that real-world audio-visual signals contain moments in which the camera moves while the sound source remains approximately stationary, thereby providing a form of cross-modal supervision. We use well-established multi-view geometry methods to estimate the relative pose of the camera, and then train a sound localization method to make predictions that are consistent with this motion.

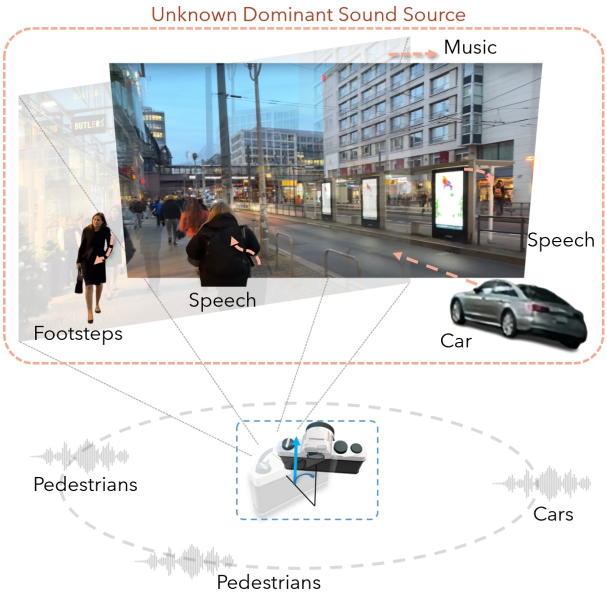

Figure 1. **Supervising sound localization using egomotion from natural video.** We use camera motion as a supervisory signal for stereo sound localization. Our model learns to predict changes in sound direction that correlate with changes in visually indicated camera motion. In contrast to prior self-supervised sound localization methods, our model is trained solely on in-the-wild natural videos, which often contain complex camera motions and noisy mixtures of different component sounds.

Our technical approach exploits the fact that, while obtaining ground truth information from vision is difficult, *approximate* visual information (such as the direction of the camera rotation) are straightforward to estimate and provide constraints that can be used as part of a weakly supervised learning framework. We show that this weak supervision from vision can be combined together with traditional interaural intensity difference cues.

Our method has a number of advantages over prior approaches. In contrast to other work that uses egomotion [13],

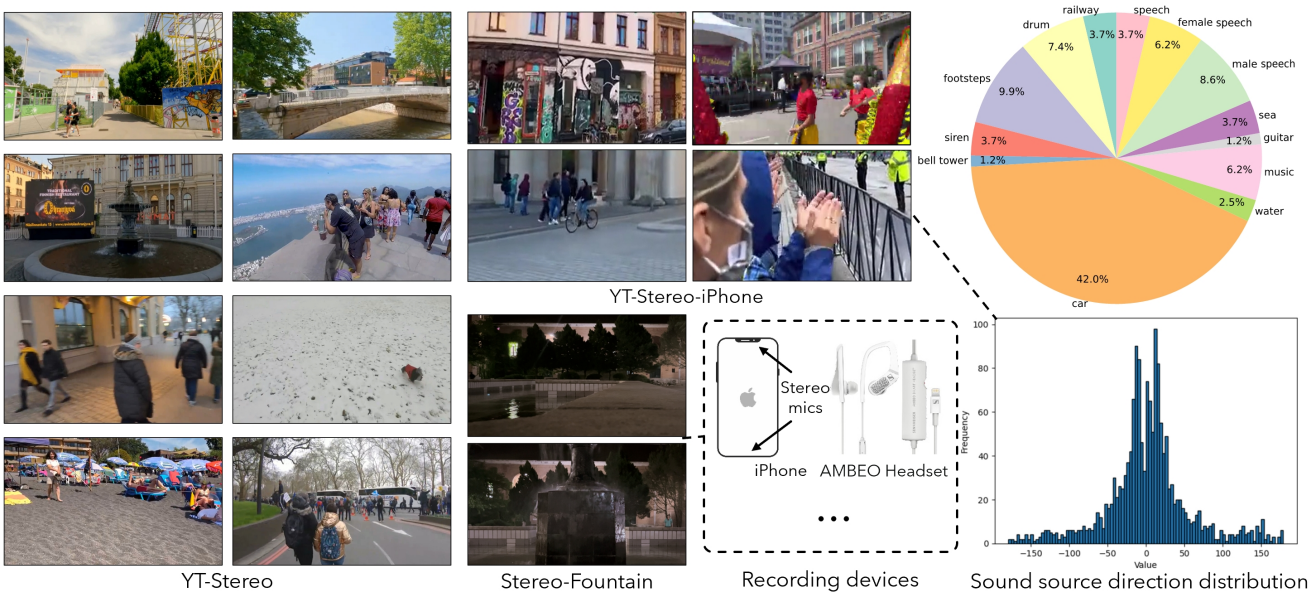

Figure 2. **The StereoWalks dataset.** Example videos from different subsets of our dataset. We focus on in-the-wild audio-visual data sourced from the internet. Our dataset contains YouTube videos recorded with stereo microphones and iPhones, as well as lab-collected in-ear binaural and stereo data. We also show the estimated distribution of different sound source categories in each subset, to help visualize the contents of each dataset.

it can be trained entirely on real videos instead of simulated environments. It does not require joint training of visual and audio models; instead, it obtains its learning signal from off-the-shelf multi-view geometry methods. This allows our method to handle diverse motions and camera translation (Fig. 1). Moreover, our approach does not require the sound source to be visible [20, 33], nor does it place restrictions on the object category [19].

We also propose a dataset containing in-the-wild stereo sound and egomotion and human-provided sound direction labels (the first such dataset, to the best of our knowledge). Our approach outperforms previous methods trained on simulated data for in-the-wild sound localization. Our contributions are as follows:

- We introduce a dataset, called *StereoWalks*, for learning sound localization from unlabeled audio-visual signals.
- We propose a method for training sound localization methods using weak supervision from egomotion.
- We experimentally evaluate our model in a variety of settings, finding that we obtain better performance than previous methods on our newly proposed sound localization dataset, and obtain performance comparable to other methods on existing real-world benchmarks.

2. Related Work

Audio-visual spatial perception. Recent research has explored the spatial correspondence between sight and sound and used them for scene analysis or reasoning [11, 15, 22, 30, 31, 34, 39, 55–57, 61]. Morgado *et al*. [34] use natural audio-visual alignment to extract representations for practical tasks. Gao *et al*. [20] focus on typical field-of-view videos and binaural audios. Zhou *et al*. [62] unify stereo generation and source separation, blending audio-visual features seamlessly. Xu *et al*. [54] employ spherical harmonic decomposition and head-related impulse response [4] to construct pseudo visual-stereo pairs. In recent developments for 3D audio-visual synthesis, Liang *et al*. [32] implicitly link audio generation with the 3D geometry and material properties of visual environments, facilitating the creation of immersive videos from various camera perspectives. Chen *et al*. [13] jointly learn sound source direction and camera pose via self-supervision from multi-view audio-visual data. In contrast, we estimate camera motion using off-the-shelf multi-view geometry methods, and use this as a supervision signal to train an audio model. Consequently, our approach is able to learn from natural video, whereas [13] required simple, simulated scenes as training data.

Spatial audio datasets. Spatial audio provides sound direction and distance cues, enriching listeners' 3D perception. Many researchers have studied these cues to interpret sound in complex scenes. Many spatial audio and audio-visual datasets have been collected to address challenges in the Direction of Arrival (DOA) problem [1, 17, 25, 41, 42]. Shimada *et al*. [49] collects a real audio-visual dataset of spatial recordings of real scenes (rooms) to study sound event localization and detection tasks. Chen *et al*. [10] multi-modal panoptic dataset with stereo audio for scene understanding. Some researchers also study the sound localization task with synthetic multichannel audio datasets with temporal activa-

tion and ground-truth DOA labels [2, 24, 35, 36, 40, 41, 51]. Many other audio-visual datasets also contain multi-channel audio [32–34, 52, 55, 60] and focus on different tasks. However, many of them are constrained by limited hours and sound type [20, 33] and recorded in lab settings [20, 42, 49]. They also require labor-intensive labeling for annotations. Chen *et al*. [10] employs four directional microphones to collect audio signals, while Grauman *et al*. [23] use Aria glasses to record synchronized egocentric audio-visual data. In contrast to previous datasets, ours is collected from diverse internet videos under an in-the-wild setup, using camera motion signals as free supervision. This approach removes the constraints for specialized hardware like multiple microphones or Aria glasses while capturing more dynamic and chaotic scenarios with complex motion patterns.

Audio-visual learning. Besides spatial correspondence between audio and visual signals, other researchers have explored different topics in audio-visual learning. Some study the deep learning approaches for visual sound localization with audio-visual semantic correspondence [5, 9, 37, 48, 63]. Some works study the temporal alignment between audio and visual streams [8, 18, 28, 37, 50]. Owens *et al*. [37] propose to use self-supervised temporal features for audio-visual scene analysis and apply them to several downstream tasks, *e.g*., action recognition. Many works use visual signals to enhance sound separation [3, 21, 58, 59]. Chen *et al*. [14] explore the visual correspondence between image and spectrogram to create visual spectrograms that that simultaneously look like natural images and sound like natural audio. Differing from those works, we focus on using ego motions from vision to supervise sound localization.

Camera motion estimation. Camera motion estimation is an important topic in 3D vision, focusing on accurately tracking camera movement through an environment. Sarlin *et al*. [47] proposed SuperGlue, a method that matches two sets of local features by jointly finding correspondences and rejecting non-matchable points, significantly improving the accuracy of feature matching. In the domain of visual SLAM (Simultaneous Localization and Mapping), Whelan *et al*. [53] and Raposo *et al*. [44] have integrated multiple camera odometry estimation techniques to achieve robust tracking, ensuring that the system can maintain reliable performance even in challenging environments. Furthermore, Rockwell *et al*. [45] propose methods for estimating camera pose robustly by combining correspondence-based and learning-based methods. Our work uses estimated camera poses as supervision for audio, providing a novel approach to enhance audio-visual learning systems.

3. Method

When the camera is in motion, the relative movement between sound sources and the camera can provide supervision for sound variations. Building upon this intuition, we use the relative moving direction as supervision, which can be interpreted as the direction of camera rotation direction and translation. In this section, we first introduce a new spatial audio-visual dataset and then demonstrate how to learn spatial audio information from videos.

3.1. The StereoWalks dataset

Our goal is to learn from in-the-wild audio-visual signals and motion using videos that span a variety of scenes, microphone designs, and cameras. Existing datasets for studying spatial audio are largely recorded in lab settings and primarily contain stationary cameras. To address this, we collect a dataset that we call *StereoWalks*. First, we acquire internet video (from YouTube) with camera motions and stereo sound (*e.g*., walking tours). We obtain a subset that is likely to contain iPhones (we call the larger set *YT-Stereo* and this subset *YT-Stereo-iPhone*). To do this, we search for captions and titles mentioning iPhone versions newer than "12S". We filter the raw videos by automatic means such as excluding those with unstable or minimal camera rotation, obtaining 20 hours of video. Human annotators label the validation and test sets with perceived ground truth angles and categorize the dominant sound source. Our dataset ensures audio-visual synchronization since both signals are collected from the same device (the camera with built-in microphones or connected headphones). This synchronization is maintained throughout the data collection and preprocessing pipeline. For further details, please refer to Sec. 4.1 and the supplementary material.

To study the influence of the recording device and additional cues from human ears, we also record two datasets using a commodity stereo microphone (iPhone 13 Pro) and a commodity in-ear binaural microphone (Sennheiser AMBEO Smart Headset). Since the latter is placed inside an ear, the model has access to additional cues due to the influence of the ear shape on the sound. We label these subsets *Stereo-Fountain* and *Binaural-Fountain* respectively. We show some examples in Fig. 2 and the statistics in Tab. 1.

3.2. Learning audio localization from egomotion

We learn to localize sound sources from stereo audio. Following previous work [12, 13], we only estimate the sound's azimuth angle, since the other degrees of freedom (elevation and distance) are challenging to perceive from audio alone.

Given an audio clip, we randomly select two short segments, $\mathbf{s}_1$ and $\mathbf{s}_2$ from time t and time t' respectively. Since most in-the-wild videos are recorded with stereo audio rather than binaural audio, it remains difficult to predict whether the sound is in front or behind from audio alone if the sound is stationary. We discuss more details in Sec. 4.5 about front-back confusion. We assume all the sound sources are on one side. For each audio clip, we predict an angle

Table 1. **Dataset comparison.** We provide details on dataset length, the proportion of visible sound sources, camera motion types, and sound source types, with each clip representing 5 seconds. Visibility is represented by two numbers: the first indicates the number of sound sources visible for over 4 seconds within the 5-second clip, and the second denotes the number of sound sources visible for less than 4 seconds. IID Binaural acc denotes the accuracy of IID predictions respected by the actual left-right labels.

Dataset	Scene subset	Split	Size		Visibility (%)	Motion Type		IID binaural acc (%)
			Clips (k)	Duration (hr)		Camera	Sound sources	
StereoWalks (ours)	YT-Stereo	Train	14.4	20.0	20 / 50	Rotation&Translation	Unknown	–
		Val	0.1	0.2	10 / 60	Mainly Rotation	Moving	57.5
		Test	0.1	0.2	10 / 60	Mainly Rotation	Moving	57.5
	Stereo-Fountain	Raw	5.0	7.0	10 / 30	Mainly Stationary	✗	76.2
	Binaural-Fountain	Raw	1.4	2.0	10 / 30	Mainly Stationary	✗	98.0
L/R Binaural [12]		Raw	1.8	3.0	–	✗	✗	75.4
Simulated	HM3D-SS 2.0 (ours)	Raw	21.6	30.0	–	Rotation&Translation	Rendered	97.4

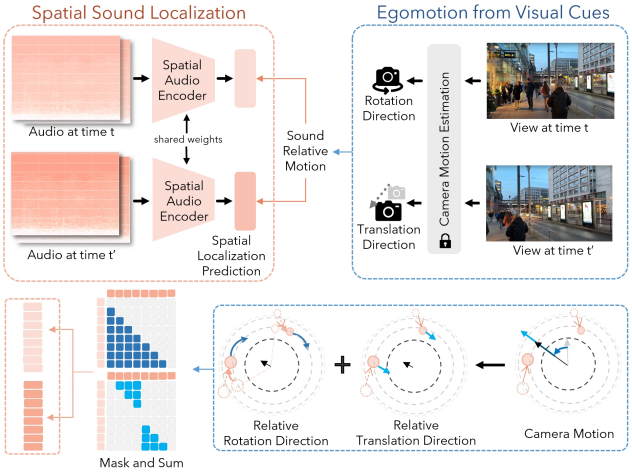

Figure 3. **Method overview.** We use a camera's egomotion to supervise binaural sound localization models. We obtain the rotation and translation of the camera using traditional methods from multi-view geometry. We then train an audio model to predict sound directions that are consistent with visual camera motions, using a dataset that includes in-the-wild walking tours.

$f(\mathbf{s}_t) : \mathbb{R}^{m\times 2} \rightarrow \mathbb{R}^n$, which we represent as a distribution over n angle categories that are uniformly sampled in a grid. The set A_i denotes all the possible angle outputs of s_t, where $\forall a \in A_i, -90° < a < 90°$.

Camera rotation estimation. Predicting the direction that a camera is rotating is easier than obtaining its precise angle (*e.g.*, since the latter may require knowledge of camera intrinsics [26]). We therefore use visual information to predict a binary label d_r that indicates whether the camera is rotating clockwise or counter for a given audio clip. This serves as the pseudo label for training our audio model.

Camera front-back translation estimation. If the position of the sound source changes slowly compared to the position of the camera, the sound source may appear to move in the opposite direction. We estimate a binary label d_t from the visual information, indicating whether the camera is moving forward for a given audio clip. In practice, we use a pre-trained camera pose estimation model to obtain both the translation and rotation labels, d_t and d_r.

Supervising audio localization using sight. The visually estimated camera rotation estimates d_r restrict the pairs of angles that $\mathbf{s}_1$ and $\mathbf{s}_2$ can compatibly be assigned to. We introduce a loss:

$$\mathcal{L}_{\texttt{rot}} = L_{\texttt{ce}}\left(\sum_{(i,j)\in R} f(\mathbf{s}_1)_i f(\mathbf{s}_2)_j, d_r\right), \quad (1)$$

where $L_{\texttt{ce}}$ is cross entropy loss and $R = \{(\alpha,\beta) : \alpha \in A_i, \beta \in A_j, \sin(\beta-\alpha)\cdot d_r > 0\}$ is the set of angle labels that are compatible with.

Similarly, the visually estimated camera front-back translation estimates d_t restrict the pairs of angles that $\mathbf{s}_1$ and $\mathbf{s}_2$ can compatibly be assigned to the audio. We introduce a loss:

$$\mathcal{L}_{\texttt{trans}} = L_{\texttt{ce}}\left(\sum_{(i,j)\in T} f(\mathbf{s}_1)_i f(\mathbf{s}_2)_j, d_t\right), \quad (2)$$

where $L_{\texttt{ce}}$ is cross entropy loss and $T = \{(\alpha,\beta) : \alpha \in A_i, \beta \in A_j, (|\beta|-|\alpha|)\cdot d_t > 0\}$ is the set of angle labels that are compatible with the visual prediction. To prevent the model from collapsing into a trivial solution (for example, predicting all angles to be zero), we exclude cases where the two angles are equal.

Binaural cues. Interaural intensity difference (IID) cues are commonly used in learning-based sound localization works. We use these cues within our model to provide a complementary form of supervision. We follow [13] and use the difference in loudness between the left and right channels to determine which side the sound is on. We penalize predictions that are inconsistent with these predictions:

$$\mathcal{L}_{\texttt{bin}} = L_{\texttt{ce}}\left(\sum_{j\in B} f(\mathbf{s}_i)_j, b_t\right), \quad (3)$$

where B is the set of angle labels that is consistent with the label b_t, which indicates whether the sound is on the left or right side of the camera. We average this loss over the pair of audio clips $\mathbf{s}_1$ and $\mathbf{s}_2$.

Overall loss. Combining these losses together, we minimize:

$$\mathcal{L} = \lambda_1 L_{\texttt{rot}} + (1-\lambda_1) L_{\texttt{trans}} + \lambda_2 L_{\texttt{bin}}, \tag{4}$$

where λ_1 and λ_2 control the relative weight of the two losses. The losses are essentially the sum of individual cross-entropy losses between the prediction and each "allowed angle" after applying the constraints. As shown in Fig. 3, "mask and sum" denotes summing over only the valid labels.

4. Experiments

4.1. Datasets

Real-world dataset. To evaluate our approach with the real-world data, we label a subset of *YT-Stereo-iPhone* and *Stereo-Fountain* as mentioned in Sec. 3.1 and Tab. 1. To ensure stable sound sources for the validation and test sets, we filter the dataset and label the audio sources through several steps. First, we segment each video into 5-second clips with a 3-second overlap. Then, we estimate the horizontal movement (left/right) using the SuperGlue model and calculate Interaural Intensity Difference (IID) cues. We select videos with large changes in IID. For initial filtering, we compute an IID score every second by calculating the IID for the preceding 2 seconds, sorting the absolute product of IID changes in descending order. Videos are then discarded if the camera rotation direction differs from the IID change direction or if the rotation angle is minimal. After filtering, we label each clip with the location of the audio sources and classify the audio type (such as car, male/female speech, sea, or animals).

Simulated dataset. To study how visual cues supervise sound localization, we created a simulated dataset to supplement our in-the-wild data. This dataset contains ground truth sound source locations and egomotion. Following Chen et al. [13], we used the SoundSpaces 2.0 [6] platform to construct a dataset, denoted *HM3D-SS 2.0*, that incorporates more diverse settings of camera motion (rotation and translation) and sound source motion. Specifically, we simulate some moving audio source and include translation in the egomotion as well. We generate the dataset with binaural Room Impulse Responses, using 3D scenes from the Habitat-Matterport 3D [43]. We partitioned the data into train/validation/test sets based on scenes. To create binaural audio, we convolve Room Impulse Responses (RIRs) with mono-channel audio from LibriSpeech [38]. To align with in-the-wild data, we constrained the total rendered audio length to about 30 hours, matching the in-the-wild dataset's scale. The *HM3D-SS* dataset from Chen et al. [13] is a subset of our *HM3D-SS 2.0* dataset, as shown in Setting (1) of Tab. 4. Please see the supplementary material for details.

4.2. Implementation details

For the angle prediction model, we used a ResNet-18 [27] architecture on spectrograms. We transformed the two-channel waveform of length L into a $256 \times 256 \times 4$ spectrogram using a short-time Fourier transform, retaining both magnitude and phase. We extracted features from the spectrograms and mapped them to 1-dimensional logits for n-class classification, setting $n = 32$. For camera motion, we used a pre-trained Superglue model. The training was conducted on an 80GB A100 GPU with a batch size of 300 for 100 epochs. For Ours-full, λ_1 was set to 0.9 and λ_2 to 1; for Ours-R&T, λ_1 was 0.9 and λ_2 was 0. In Tab. 2, we first train on *YT-Stereo* and part of *YT-Stereo-iPhone*, then fine-tune for one epoch on each dataset.

To obtain the camera motion, we employ the Superglue model [46] and utilize the perspective field [29] to predict the horizontal field of view, resulting in the rotation matrix and the translation vector. We sampled 5 frames per second for a 5 sec. video and calculated the rotation matrix and translation vector between each frame. Additionally, we computed the rotation matrix and translation vector between different images at intervals of 3 and 6 frames. For the purposes of cleaning the dataset, for any two time points, we accumulated the rotations calculated at intervals of 3 frames, proportionally adjusting if there were gaps. The same approach was applied to translations.

4.3. Evaluation

Baselines and ablations. We evaluate the performance of sound localization by the MAE of angle prediction and the accuracy of 2-way classification to compare with IID-direct and 8 classification. The accuracy of 2 classification is denoted as "2clf" and that of 8 classification is denoted as "8clf". Since there exists front-back confusion for stereo devices if given only one audio segment, we reflect all the predictions to the front to avoid ambiguity.

We use the following baselines: 1) Chance: Baseline performance by random guessing. 2) IID-direct: Direct prediction using interaural intensity difference (IID) cues, as described in Chen et al. [13]. 3) GTRot: Implementation of the method from Chen et al. [13], which uses "oracle" rotation angle and pseudo binaural cues as supervision. For real-world datasets, we implemented GTRot training using pseudo labels of rotation angles obtained by the SuperGlue model [46] and Perspective Fields [29] for the *YT-Stereo-iPhone* dataset. For the Stereo-Fountain dataset, the angles are labeled by humans for the training set and evaluation set. 4) Supervised: We train a supervised model using ground-truth angles for each dataset, covering 360 degrees. 5) Supervised–RTF: We train a supervised model

Table 2. Comparison with state-of-the-art methods and other self-supervised methods on YT-Stereo and In-the-wild dataset. *Sim.* denote the simulated dataset.

Model	Training Set		Test Set								
			YT-Stereo-iPhone			Stereo-Fountain			Simulated		
	YT-Stereo	Sim.	MAE (°)↓	2clf (%)↑	8clf (%)↑	MAE (°)↓	2clf (%)↑	8clf (%)↑	MAE (°)↓	2clf (%)↑	8clf (%)↑
Chance			55.3	46.7	12.7	62.1	52.0	16.0	39.4	49.0	15.0
IID – direct			–	57.5	–	–	97.0	–	–	97.4	–
GTRot [13]	✓		71.4	54.0	7.1	57.2	97.0	18.4	40.2	48.7	18.7
Ours – IID only [13]	✓		37.3	54.7	28.0	30.1	97.0	**46.0**	88.3	48.7	6.8
Ours – Simulated		✓	73.4	61.7	13.3	**28.5**	85.7	22.3	**9.8**	**98.3**	**63.4**
Ours – Full	✓		**34.0**	**61.7**	**33.3**	29.3	**97.3**	**46.0**	57.2	48.7	10.7

In-the-wild audio [12]	2clf (%)↑
Chance	50.0
IID – direct	75.4
GCC-PHAT	77.2
GTRot [13]	84.9
MonoCLR [12]	87.4
StereoCRW [12]	87.2
Ours – IIDonly	**87.7**
Ours – full	87.5

Table 3. 360-Degree Sound Localization Prediction Results for YT-Stereo-iPhone and Stereo-Fountain Datasets. The YT-Stereo-iPhone dataset includes rotations and translations with mostly moving sound sources, while the Stereo-Fountain dataset has mostly stationary cameras and sound sources.

Model	YT-Stereo-iPhone			Stereo-Fountain		
	MAE (°) ↓	2clf (%) ↑	8clf (%) ↑	MAE (°) ↓	2clf (%) ↑	8clf (%) ↑
Chance	55.3	46.7	12.7	62.1	52.0	16.0
IID-direct	–	57.5	–	–	97.0	–
GTRot [13]	71.4	54.0	7.1	55.1	97.1	18.7
Ours – Simulated	73.4	**61.7**	13.3	28.5	85.7	22.3
Ours – IID only [13]	37.3	54.7	28.0	29.6	97.3	46.0
Ours – R&B	34.5	55.4	**33.4**	28.5	97.3	46.0
Ours – T&B	37.4	53.4	26.8	–	–	–
Ours – Full	**34.0**	**61.7**	31.7	**28.5**	**97.3**	**46.0**
Supervised	–	–	–	33.8	98.0	49.5
Supervised – RTF	–	–	–	27.1	98.0	49.5

using ground truth mapping of the sound source behind the recording device to the front, followed by supervised learning.

For evaluating on the in-the-wild binaural audio benchmark [12], we use binary classification accuracy as the metric. Apart from the baselines above, we compare our model to a state-of-the-art self-supervised binaural prediction method based on time delay prediction and contrastive random walks trained on stereo sounds [12]. MonoCLR denotes the model with instance discrimination trained on mono sounds. GTRot [12] is trained on simulated data HM3D-SS from SoundSpaces 2.0 [6]. StereoCRW and MonoCLR are trained with FMA music samples [16] and FAIR-Play [20]. The IID-direct is trained with the In-the-wild audio of Chen et al. [12].

We investigated several model variants to determine whether our model effectively uses visual cues. 1) **Ours–full**: using L_{rot}, L_{trans} and L_b, 2) **Ours–IIDonly**: employing L_b only, 3) **Ours–R&B**: using L_{rot} and L_b without L_{trans}, 4) **Ours–T&B**: using L_{trans} and L_b without the rotation loss, 5) **Ours–R&T**: using L_{rot} and L_{trans} without the pseudo binaural loss, which is only supervised by the egomotion labels, 5) **Ours–simulated**: training models solely on the simulated dataset *HM3D-SS 2.0* Dataset with three losses.

In GTRot, Supervised, Supervised–RTF, and our models, the output has 32 classification possibilities. We calculate the Mean Absolute Error by summing the product of each possibility and the midpoint of its corresponding interval.

4.4. Results

As shown in Tab. 2, our proposed method, Ours-full, outperforms baselines across various in-the-wild datasets. Tested on the *YT-Stereo-iPhone* dataset, our model achieved the best results, showing advantages over other baselines. We tested our models on both *YT-Stereo-iPhone* and *Stereo-Fountain* datasets, demonstrating robust performance across scenarios. Visualizations are shown in Fig. 4. Our model outperforms the method in [13], which uses stronger supervision.

When considering the sim2real gap, our simulated model has not generalized well to *YT-Stereo-iPhone* and *Stereo-Fountain*. The model trained on real-world data does not generalize well to the simulated dataset. This highlights the necessity of training with ego-motion as supervision on in-the-wild data for spatial sound localization.

As shown in Tab. 3, when training on specific datasets, Ours-full achieves comparable best performance on *YT-Stereo-iPhone* and *Stereo-Fountain*, demonstrating the effectiveness of our method. Evaluating on the previous dataset In-the-wild audio, we fine-tuned our model trained on *YT-Stereo*, and it achieved comparable good results.

Table 4. Overlapping sounds make in-the-wild sound localization challenging: Experiments on three simulated datasets where the camera makes small translations in all settings. To better align with the real-world experiments, as in settings (1)(2)(3), the dataset is restricted to 30 hours, while in (4), the dataset is not limited. "Intermittent" refers to sound sources that are silent for half the time. "Overlap" refers to the presence of multiple sound sources.

Model	(1) One Source			(2) Overlap			(3) Overlap Intermittent			(4) More Simulated Data of (3)		
	MAE(°)↓	2clf(%)↑	8clf(%)↑	MAE(°)↓	2clf(%)↑	8clf(%)↑	MAE(°)↓	2clf(%)↑	8clf(%)↑	MAE(°)↓	2clf(%)↑	8clf(%)↑
Chance	39.4	49.0	15.0	39.4	49.0	15.0	39.4	49.0	15.0	39.4	49.0	15.0
IID	-	97.4	-	-	80.1	-	-	80.8	-	-	80.8	-
GTRot [13]	**4.3**	97.5	**84.7**	**22.5**	83.5	33.5	26.9	81.5	32.5	**19.7**	**87.2**	41.0
Ours – IIDonly [13]	22.2	**99.0**	35.7	25.6	78.8	26.1	28.0	78.0	23.7	26.7	79.5	25.7
Ours – Full	9.8	98.3	63.4	23.2	**87.2**	**36.2**	**24.1**	**82.1**	**35.7**	23.7	82.4	**41.2**
Supervised	2.8	98.7	89.8	10.8	91.8	70.2	8.4	94.1	74.6	5.7	96.0	80.2

Table 5. Relationship between ego-translation and relative motion of sources. Experiments were conducted with various types of motion and distances from the camera using the simulated dataset *HMSS-3D 2.0* made by SoundSpaces 2.0 [6]. We report Mean Absolute Error in the unit of degrees (°), and accuracy of 2clf and 8clf in the unit of percentage (%).

Model	(1) Rotation Only			(2) Translation Only			(3) Rotation&Translation			(4) Distant Rotation			(5) Distant R&T			(6) Overlap R&T		
	MAE↓	2clf↑	8clf↑	MAE↓	2clf↑	8clf↑	MAE↓	2clf↑	8clf↑	MAE↓	2clf↑	8clf↑	MAE↓	2clf↑	8clf↑	MAE↓	2clf↑	8clf↑
Chance	39.4	49.0	15.0	39.4	49.0	15.0	39.4	49.0	15.0	39.9	49.2	15.7	39.8	49.1	15.8	39.0	49.5	15.2
IID	-	97.4	-	-	97.4	-	-	97.4	-	-	95.6	-	-	95.1	-	-	73.5	-
GTRot [13]	**4.3**	97.5	**84.7**	-	-	-	4.4	97.5	84.5	**4.5**	97.1	**82.0**	**4.7**	96.7	**86.9**	28.1	79.6	30.1
Ours – IIDonly [13]	22.2	**99.0**	35.7	**22.7**	**99.0**	**35.7**	22.0	**99.1**	35.9	23.1	97.8	35.2	23.0	98.1	35.0	32.5	75.0	25.7
Ours – R&B	9.8	98.3	63.4	-	-	-	**10.1**	98.0	**60.1**	10.7	**98.1**	59.1	11.7	**98.6**	58.7	**25.6**	**81.4**	**34.6**
Ours – T&B	-	-	-	23.2	96.0	**35.7**	23.0	95.7	35.0	-	-	-	22.7	96.7	35.1	30.1	72.5	22.3
Ours – R&T	-	-	-	-	-	-	14.7	96.3	48.3	-	-	-	22.5	95.0	45.7	26.7	79.0	29.2
Ours – Full	9.8	98.3	63.4	23.2	96.0	**35.7**	11.7	98.0	58.7	11.2	**98.1**	56.8	11.9	97.6	55.3	28.6	76.0	28.5
Supervised	2.8	98.7	89.6	2.6	99.1	89.9	2.6	98.9	89.5	2.7	98.8	90.0	2.5	99.0	90.4	12.1	86.0	65.4

4.5. Analysis

Overlapping sound assumption. Previously, based on the hypothesis that co-occurring audio and visual signals offer "free" supervision capable of capturing geometry, including camera motion and sound source direction, Chen *et al*. [13] proposed jointly estimating camera rotation from images and sound direction from binaural audio. Their approach leverages geometric consistency and pseudo-binaural cues. In that work, audio events are mostly singular or the main sound source is significantly louder than secondary sources, making the sound environment relatively ideal. However, in real-world scenes, sounds may overlap and be intermittent, making localization more challenging.

We assume that the overlapping and intermittent sounds contribute to our model's better generalization compared to GTRot, the "oracle" method from [13], despite GTRot having stronger supervision than our model. To test these hypotheses, we conducted experiments in a simulated environment based on SoundSpaces 2.0 [6]. Since most sounds in natural video are overlapping and intermittent sounds, we tested different models under the following settings: (1) *One Source*: Only one sound source in the environment. (2) *Overlap*: A secondary sound source is added at a constant position, with a loudness of 0.7 relative to the main sound source. (3) *Overlap Intermittent*: A secondary sound source is added with 0.7 times the loudness of the intermittent main sound source, where both sources are silent independently for half of the time. (4) *More Simulated Data of (3)*: We generate additional data by pairing room impulse responses with mono-channel audio for convolution, creating more data pairs. In these settings, camera translations are controlled within a small range, where the pose is restricted to a small area. To better align with the real-world experiments, as in settings (1)(2)(3), the dataset is restricted to 30 hours, while in (4), the dataset is not limited.

As shown in Tab. 4, with limited data, our proposed method surpasses previous methods in settings (2) *Overlap* and settings (3) *Overlap Intermittent* while in settings (1) *One Source* GTRot has the best performance. This outcome suggests that using egomotion direction instead of angle as supervision can generalize better to more complicated scenarios. Regarding settings (4) *More Simulated Data of (3)*, although GTRot has stronger supervision than Ours-full, they show comparable performance, demonstrating our method's effectiveness.

Rotation and translation. In the internet walking tour videos (the *YT-Stereo* subset), since most sounds are distant, the absolute translation of the sound sources greatly affects the relative translation between the sound source and the camera. However, in terms of angular velocity, the rotation speed of the camera itself is much greater than the rotational speed of the sound source relative to the camera. Therefore, we can disregard the relative translation. From experimental results, it can be observed that removing the translation loss from the model yields performance on par with models that

Table 6. Evaluation of sound localization on the *Stereo-Fountain* and *Binaural-Fountain* datasets, showing improved accuracy in front/back localization with binaural recordings.

Dataset	MAE ↓	Left/Right Acc ↑	Front/Back Acc ↑
Stereo-Fountain	33.8	98.0%	51.0%
Binaural-Fountain	**27.8**	**99.0%**	**69.3%**

include translation loss on the *YT-Stereo* dataset.

To study how camera rotation and translation affect sound localization and identify which loss functions are dominant in different scenarios, we conducted experiments in a simulated environment with the following settings as shown in Tab. 5: (1) *Rotation Only*: Only camera rotation with a single sound source. (2) *Translation Only*: Only camera translation with front-back translation in the range of 0–2m. (3) *Rotation&Translation*: Both camera rotation and front-back translation, with translation in the range of 0–2m. (4) *Distant Rotation*: A sound source 5 meters away from the camera, otherwise identical to setting (1). (5) *Distant R&T*: A sound source 5 meters away from the camera and within 7 meters, otherwise identical to setting (3). (6) *Overlap R&T*: Two sound sources, with the addition of a secondary sound source with a constant position and 0.7 loudness to the main sound source, otherwise identical to setting (3) *Rotation&Translation*. In all settings, the dominant sound sources have their movement within 0.5 meters.

Without relying on pseudo-binaural cues, our method—guided solely by camera ego-motion—achieves similar performance in setting (3) *Rotation&Translation*, where the camera translates within 0–2 meters. This suggests that for nearby sound sources, small camera translations provide sufficient cues for sound localization. However, in setting (5) *Distant R&T*, a new challenge arises as sound sources are positioned further away, making camera shifts less impactful on sound positions and resulting in minimal angle changes.

Models trained with both rotation and pseudo-binaural cues effectively capture distant sound localization cues. In setting (6) *Overlap R&T*, with overlapping sound sources and complex motion, our ego-motion-based approach continues to perform comparably to the best results. This highlights the method's adaptability, using camera movements to accurately localize sounds across various challenging conditions.

Front-Back ambiguity. To explore how different recording devices handle front-back sound localization, we conducted an experiment to predict supervised 360° sound localization on *Stereo-Fountain* and *Binaural-Fountain*, both recorded beside the same fountain using the microphones of an iPhone 13 Pro and a Sennheiser AMBEO Smart Headset, respectively. The recordings allowed us to compare how well each setup managed 360° sound localization.

As shown in Tab. 6, the model obtained higher accuracy on *Binaural-Fountain* than *Stereo-Fountain*. The latter dataset was recorded on an iPhone. Its stereo microphone may have less variation in the audio captured from behind versus in front of the device. In contrast, the in-ear Sennheiser AMBEO Smart Headset offered a clearer distinction, which may be due to the extra cues provided by the wearer's ear or head.

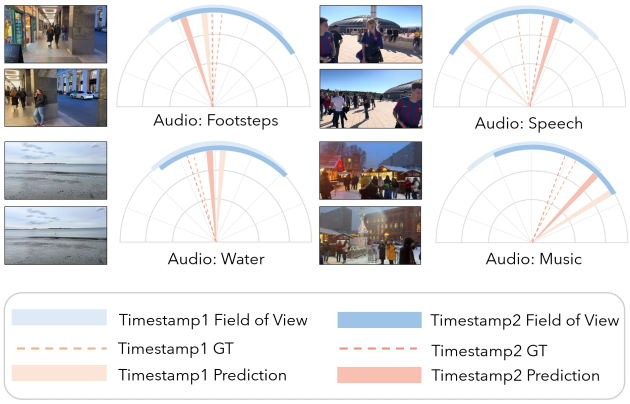

Figure 4. Visualizations of results in the internet walking tour videos (the *YT-Stereo* subset). We show our predictions and ground-truth annotations with angles and audio class labels.

5. Conclusion

We propose a method and dataset for in-the-wild sound localization. Our method is trained without labeled data. Instead, it obtains its supervision from visual camera motion and binaural audio cues. Through our experiments on in-the-wild and simulated data, we find that this egomotion supervision aids sound localization.

Limitations. While our work introduces the first in-the-wild sound localization training dataset and corresponding evaluation sets, the challenges of low-quality stereo audio and prevalent overlapping sounds complicate our data-cleaning process. Filtering valid clips from numerous unlabeled monocular videos captured in the wild may result in overlooking brief sound effects.

Future work. We anticipate that future research will address these limitations by incorporating more visual and contextual data. We see our work as a step toward creating sound localization methods that work in challenging real-world conditions, and a step toward multimodal methods that learn geometric information from unlabeled data.

Acknowledgements. This work was completed as part of an internship by Anna Min at University of Michigan. We would like to thank Zafar Rafii, Ethan Manilow, and Ruohan Gao for their helpful feedback. We thank John Chu for providing samples he recorded on YouTube for preliminary experiments.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018. 2

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection. *arXiv preprint arXiv:1905.08546*, 2019. 3

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *European Conference on Computer Vision (ECCV)*, 2020. 3

[4] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 99–102. IEEE, 2001. 2

[5] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 3

[6] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *arXiv*, 2022. 5, 6, 7

[7] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6419, 2023. 2

[8] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronisation in the wild. *arXiv preprint arXiv:2112.04432*, 2021. 3

[9] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021. 3

[10] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+ x: A panoptic multi-modal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19373–19382, 2024. 2, 3

[11] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. *arXiv preprint arXiv:2111.05846*, 2021. 2

[12] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 489–508. Springer, 2022. 1, 3, 4, 6, 2

[13] Ziyang Chen, Shengyi Qian, and Andrew Owens. Sound localization from motion: Jointly learning sound direction and camera rotation. *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 4, 5, 6, 7

[14] Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas. *Neural Information Processing Systems (NeurIPS)*, 2024. 3

[15] Jesper Haahr Christensen, Sascha Hornauer, and Stella Yu. Batvision with gcc-phat features for better sound to vision predictions. *arXiv preprint arXiv:2006.07995*, 2020. 2

[16] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016. 6

[17] Christine Evers, Heinrich W Löllmann, Heinrich Mellmann, Alexander Schmidt, Hendrik Barfuss, Patrick A Naylor, and Walter Kellermann. The locata challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1620–1643, 2020. 2

[18] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. *arXiv preprint arXiv:2301.01767*, 2023. 3

[19] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7053–7062, 2019. 1, 2

[20] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 6

[21] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 3

[22] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 658–676. Springer, 2020. 2

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas

Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024. 3

[24] Eric Guizzo, Riccardo F Gramaccioni, Saeid Jamili, Christian Marinoni, Edoardo Massaro, Claudia Medaglia, Giuseppe Nachira, Leonardo Nucciarelli, Ludovica Paglialunga, Marco Pennese, et al. L3das21 challenge: Machine learning for 3d audio signal processing. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2021. 3

[25] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3das22 challenge: Learning 3d audio sources in a real office environment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9186–9190. IEEE, 2022. 2

[26] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 5

[28] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. *arXiv preprint arXiv:2401.16423*, 2024. 3

[29] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. *CVPR*, 2023. 5

[30] Takashi Konno, Kenji Nishida, Katsutoshi Itoyama, and Kazuhiro Nakadai. Audio-visual sfm towards 4d reconstruction under dynamic scenes. *CVPR 2020 workshop on Sight and Sound*, 2022. 2

[31] Tingle Li, Renhao Wang, Po-Yao Huang, Andrew Owens, and Gopala Anumanchipalli. Self-supervised audio-visual soundscape stylization. In *European Conference on Computer Vision*, pages 20–40. Springer, 2025. 2

[32] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[33] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *arXiv preprint arXiv:1809.02587*, 2018. 2, 3

[34] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 2, 3

[35] Kento Nagatomo, Masahiro Yasuda, Kohei Yatabe, Shoichiro Saito, and Yasuhiro Oikawa. Wearable seld dataset: Dataset for sound event localization and detection using wearable devices around head. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 156–160. IEEE, 2022. 3

[36] Shutong Niu, Jun Du, Qing Wang, Li Chai, Huaxin Wu, Zhaoxu Nian, Lei Sun, Yi Fang, Jia Pan, and Chin-Hui Lee. An experimental study on sound event localization and detection under realistic testing conditions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[37] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 3

[38] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 5, 2

[39] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8268–8277, 2021. 2

[40] Archontis Politis, Sharath Adavanne, and Tuomas Virtanen. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. *arXiv preprint arXiv:2006.01919*, 2020. 3

[41] Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. *arXiv preprint arXiv:2106.06999*, 2021. 2, 3

[42] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv preprint arXiv:2206.01948*, 2022. 2, 3

[43] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 5

[44] Carolina Raposo and Joao P Barreto. match: Monocular vslam and piecewise planar reconstruction using fast plane correspondences. In *European conference on computer vision*, pages 380–395. Springer, 2016. 3

[45] Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, and David F. Fouhey. Far: Flexible, accurate and robust 6dof relative camera pose estimation. In *CVPR*, 2024. 3

[46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 5

[47] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 3, 1

[48] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 3

[49] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[50] Jiatian Sun, Longxiulin Deng, Triantafyllos Afouras, Andrew Owens, and Abe Davis. Eventfulness for interactive video alignment. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10, 2023. 3

[51] Qing Wang, Jun Du, Zhaoxu Nian, Shutong Niu, Li Chai, Huaxin Wu, Jia Pan, and Chin-Hui Lee. Loss function design for dnn-based sound event localization and detection on low-resource realistic data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[52] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee. A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1251–1264, 2023. 3

[53] Thomas Whelan, Hordur Johannsson, Michael Kaess, John J Leonard, and John McDonald. Robust real-time visual odometry for dense rgb-d mapping. In *2013 IEEE International Conference on Robotics and Automation*, pages 5724–5731. IEEE, 2013. 3

[54] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 2

[55] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2, 3

[56] Karren Yang, Michael Firman, Eric Brachmann, and Clément Godard. Camera pose estimation and localization with active audio sensing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 271–291. Springer, 2022.

[57] Heeseung Yun, Ruohan Gao, Ishwarya Ananthabhotla, Anurag Kumar, Jacob Donley, Chao Li, Gunhee Kim, Vamsi Krishna Ithapu, and Calvin Murdock. Spherical world-locking for audio-visual localization in egocentric videos. In *European Conference on Computer Vision*, pages 256–274. Springer, 2025. 2

[58] 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 (ECCV)*, pages 570–586, 2018. 3

[59] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 3

[60] Wenru Zheng, Ryota Yoshihashi, Rei Kawakami, Ikuro Sato, and Asako Kanezaki. Multi event localization by audio-visual fusion with omnidirectional camera and microphone array. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2573, 2023. 3

[61] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*, 2024. 2

[62] 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 (ECCV)*, 2020. 2

[63] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 3