

# Supervising Sound Localization by In-the-wild Egomotion

## Supplementary Material

### A.1. Dataset

We provide more information about our dataset collection procedure.

#### A.1.1. Internet data collection

One of the major components of our dataset is the *YT-stereo* subset, which is obtained by querying YouTube for queries associated with walking tours, such as “walking tour.” We filter out videos shot in portrait mode or with mono audio.

To ensure that the camera’s motion correlates with the relative change of the sound sources, it’s essential that the angular velocity of the sound source is much slower than the camera’s angular velocity. To get stable sound sources as the validation set and test set, which provide a stronger supervision signal, we take a way to filter the dataset for stable sound sources. After the process, 20 hours of videos are selected for the training set.

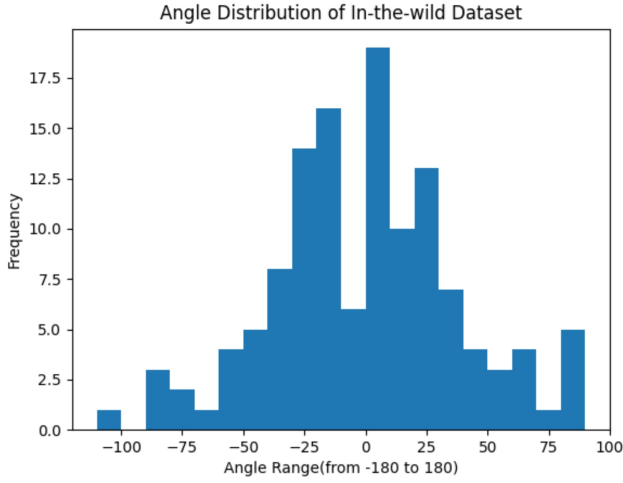


Figure 5. The angle distribution of YT-Stereo-iPhone validation set. The angles range from -180 to 180 without filtering.

#### A.1.2. Dataset Filtering

**YT-Stereo-iPhone.** The training set is unfiltered, while the validation and test sets are filtered using the following methods.

We manually select 100 clips as the validation set from a pool of 1600 filtered clips. To ensure that the motion of the camera correlates with the relative change of the sound sources, it is essential that the angular velocity of the sound source is much slower than the angular velocity of the camera. To get clips of clear camera motion for training, we

filter the dataset with the following steps: 1) We segment the videos into 5-second clips every 3 seconds. 2) We estimate horizontal motion using the Superglue model [47] to get the rotation matrix and translation matrix of the clips. 3) We filter out videos with camera rotation angles of less than 10 degrees.

To get the motion direction of sound sources, we use interaural intensity difference (IID) cues as a pseudo label to predict the moving direction of the sound source following [12, 13]. We follow to estimate whether a certain segment of audio is on the left or right by IID as a pseudo label, which is based on which side sound is louder than the other:  $d = \text{sign}(\log \left| \frac{A_L}{A_R} \right|)$  where  $|A|$  is the magnitude of the spectrogram  $A$ .

To obtain a subset of the videos where the sound source positions are relatively easy for labelers to label, we take a way to filter the dataset for stable sound sources. Subsequently, we undertake curation through the following steps: 1) We segment the videos into 5-second clips every 3 seconds. 2) Estimating horizontal left/right movement using the Superglue model. 3) We calculate the Interaural Intensity Difference (IID) cues of the videos. 4) Due to the large number of videos, for initial filtering, we compute an IID score: every 1 second, we calculate the IID for 2 seconds, arranging the absolute product of IID changes in descending order. 5) We discard videos where the camera rotation direction differs from the IID change direction. 6) We filter out videos with camera rotation angles less than 15 degrees, and the remaining videos are sampled concerning rotation angles to ensure a relatively uniform distribution.

In Fig. 5, we show the angle distribution of the YT-Stereo-iPhone validation set.

**Stereo-Fountain.** One of the authors positioned themselves at various fixed locations near a fountain and recorded using an iPhone 13 Pro. Subsequently, one of the authors added annotations. We split the data for train/test/val randomly.

The data collection steps were as follows: 1) One author selected 15 angles out of 360 degrees around the fountain, using an iPhone and a laser pointer (aligned visually to the fountain) to calculate the angle between the fountain and the phone; 2) The recording device was placed at each corresponding position and recorded for approximately 10 minutes.

**Binaural-Fountain.** One of the authors positioned themselves at various fixed locations near a fountain and recorded using an iPhone 13 Pro and a binaural microphone (Sennheiser AMBEO Smart Headset). Subsequently, annota-

Table 7. **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	Split	Size		Visibility ( % )	Motion Type		IID binaural acc ( % )
			Clips	Duration		Camera	Sound sources	
In the wild dataset	YT-stereo	Raw	13,000k	8.0k hrs	–	Mainly	–	Unknown
		Train	14.6k	20hrs	-	rotation	Unknown	
	YT-Stereo-iPhone	Raw	95k	80hrs	–	–	–	Unknown
		Train	14.6k	20hrs	20/50	Mainly	Unknown	
		Val	0.1k	0.2hrs	10/60	rotation	Moving	
Real-world lab-made		Test	0.1k	0.2hrs	10/60		Moving	
	L/R Binaural[12]	Raw	1.8k	3hrs	-	✗	✗	75.4
	Stereo-Fountain	Raw	1.4k	2hrs	10/30	Mainly Stationary	✗	97.0
	Stereo-Music	Raw	3.6k	5hrs	10/30	✗	✗	70.0
	Binaural-Fountain	Raw	1.4k	2hrs	10/30	Mainly Stationary	✗	98.0
	RWAVS [32]	Raw	2.7k	3.8hrs	Unknown	Rotation&Translation	✗	71.9
	Simulated Dataset	Raw	21.6k	30hrs	-	Rotation&Translation	Rendered	97.4

tions were added by one of the authors. We split the data for train/test/val randomly.

The data collection steps were as follows: 1) One author selected 15 angles out of 360 degrees around the fountain, using an iPhone and a laser pointer (aligned visually to the fountain) to calculate the angle between the fountain and the phone; 2) The author was seated at each corresponding position and recorded for approximately 10 minutes.

**Stereo-Music.** One of the authors positioned the iPhone at a static place in the same room, and one mobile phone was playing music. Subsequently, annotations were added by one of the authors. We split data for train/test/val by mixing all the data and randomly splitting.

**RWAVS [32].** This is the Real-World Audio-Visual Scene (RWAVS) dataset described in [32]. The authors recorded data in both indoor and outdoor environments to represent daily settings. They used a 3Dio Free Space XLR binaural microphone for high-quality stereo audio, a TASCAM DR-60DMKII for recording and storing audio, and a GoPro Max for capturing videos. The devices were mounted together and moved randomly around the environment, unlike the ReplayNVAS [7] dataset, which has a constant environment and recording viewpoint. Data collection for each scene ranged from 10 to 25 minutes.

Table 8. Summary of classification accuracy for RWAVS and Stereo-Music datasets.

Dataset	Classification Task	Accuracy (%)
RWAVS [32]	supervised 4-classification	50
Stereo-Music	supervised 4-classification	28

**Simulated dataset.** Due to SoundSpace 2.0 not supporting material and moving sound currently, we use sound sources at different positions as the moving sound.

We follow [13] to create the simulated dataset. Binaural audio is obtained by convolving binaural RIRs with mono audio samples from LibriSpeech [38].

Our dataset comprises 50,000 audio-visual pairs generated from 200,000 viewpoints. The audio was rendered with an average reverberation of  $RT60 = 0.4s$ . For training, validation, and testing, we divided our data into 81/9/10 scenes, respectively. We utilized approximately 30 hours of synthesized data to ensure fairness in our results.

### A.1.3. Label the dataset

Annotations were added by one of the authors and another participant. This is the questionnaire used to label data.

You will need to label relatively clear audio events whether in the scene or out of the scene. Please use the best headphones you can have. We provide some examples to help you label the other data.

- 1) Please provide the quality of the video’s audio for scoring.
- 2) Provide the video, with a 5-second audio clip, where are the audio sources located, totaling 16 classes. Or indicate if it is unknown. You only need to annotate the ones that you can distinguish where the sound is. As 90 denotes the left, -90 denotes the right. 0 denotes in front of you. 180/-180 denotes the back of you.
- 3) What are the categories of the audio sources? (If re-

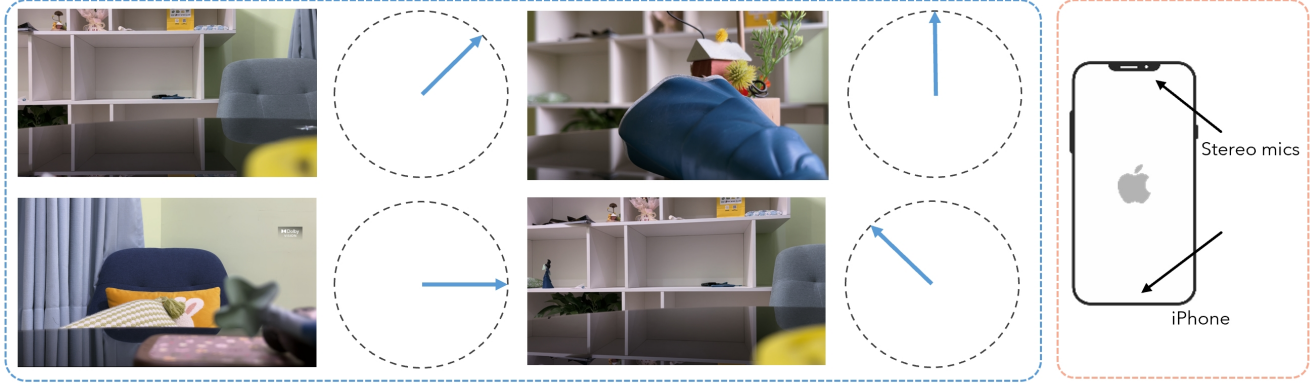


Figure 6. Examples from the *Stereo-Music* subset. We record sound using a stereo microphone in a scene containing music.



Figure 7. The examples from the RWAVS dataset were recorded using binaural microphones in various scenarios where the sound source was a loudspeaker. Above, we present images from the original paper [32].

- peated), please list all of them. Existing categories include car/male speech/female speech/speech/sea/animals
- 4) What is the direction of movement of the audio sources?  
If consistent with question 3, no need to fill it.
  - 5) If the category cannot be indicated by numbers, there's no need to label it. If the quality is too low, just skip it.

#### A.1.4. Additional scenes

To explore the relationships among various recording devices, we employ the identical procedure on a second dataset labeled YT-stereo-iPhone, also sourced from YouTube using keywords ranging from iPhone 12 Pro to iPhone 15 Pro, specifically targeting devices capable of recording spatial audio from both the bottom and top microphones. Besides, we recorded about 2 hours of labeled audio using an iPhone 13 Pro around a fountain as a supplement denoted as Stereo-Fountain and around a music player in one room, which we call *Stereo-Music*, and 2 hours of labeled audio using Sennheiser AMBEO Smart Headset to record binaural audio, which we call *Binaural-Fountain*.

**Music data out of distribution.** For the RWAVS dataset [32], we partitioned each part of the 11 scenes randomly into 80%/10%/10% splits for training, testing, and validation sets, respectively. When applying a supervised approach to the Stereo-Music dataset and RWAVS Dataset,

we observe that training the model solely on each dataset results in limited success. Specifically, the accuracy only reaches 50% for the supervised 4-classification task on the RWAVS Dataset and it reaches only 28% on the Stereo-Music dataset. This outcome highlights the current model's inadequacy when dealing with indoor music data.

## A.2. Experiment Implementations

### A.2.1. Estimate camera motion

For a five-second video, we sampled five frames per second 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. During dataset cleaning, we followed a specific procedure. 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.

### A.2.2. Training details

For training Ours-full, the  $\lambda_1$  is set as 0.9 and  $\lambda_2$  is set as 1. For training Ours-R&T, the  $\lambda_1$  is set as 0.9, and  $\lambda_2$  is set as 0. All models are trained using a learning rate of 0.0001 with the AdamW optimizer. The training schedule followed a

cosine annealing schedule and early stopping. Training one model takes about 2 hours. Due to computational limitations, we train once for each number, using the same seed for every experiment.