Sound Localization from Motion: Jointly Learning Sound Direction and Camera Rotation

Ziyang Chen Shengyi Qian Andrew Owens

University of Michigan

Abstract

The images and sounds that we perceive undergo subtle but geometrically consistent changes as we rotate our heads. In this paper, we use these cues to solve a problem we call Sound Localization from Motion (SLfM): jointly estimating camera rotation and localizing sound sources. We learn to solve these tasks solely through self-supervision. A visual model predicts camera rotation from a pair of images, while an audio model predicts the direction of sound sources from binaural sounds. We train these models to generate predictions that agree with one another. At test time, the models can be deployed independently. To obtain a feature representation that is well-suited to solving this challenging problem, we also propose a method for learning an audio-visual representation through cross-view binauralization*: estimating binaural sound from one view, given images and sound from another. Our model can successfully estimate accurate rotations on both real and synthetic scenes, and localize sound sources with accuracy competitive with state-of-the-art self-supervised approaches. Project site: https://ificl.github.io/SLfM.*

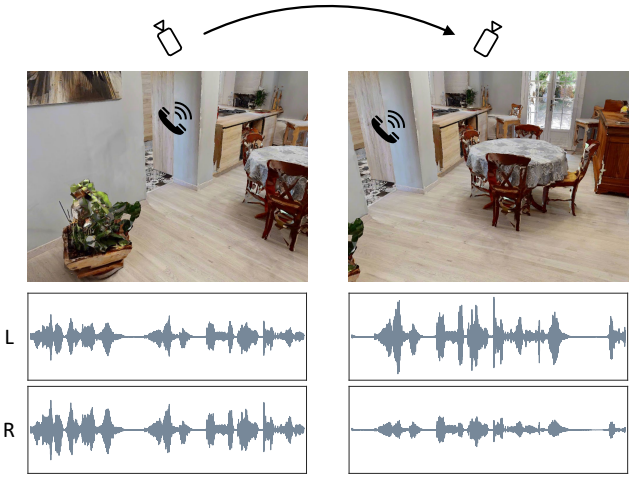


Figure 1: **Images and sounds change in geometrically consistent ways.** For example, when we rotate to the right, a sound source that is initially in front of us becomes louder in our left ear. We use these cues to jointly train models for two tasks: localizing sounds from binaural audio and estimating camera rotation from images. The two models are trained entirely through self-supervision, by learning to produce outputs that agree with one other.

1. Introduction

As you rotate your head, the images and sounds that you perceive change in geometrically consistent ways. For example, after turning to the right, a sound source that was directly in front of you will become louder in your left ear and quieter in your right, while simultaneously the visual scene will move right-to-left across your visual field (Fig. 1).

We hypothesize that these co-occurring audio and visual signals provide "free" supervision that captures geometry, including the motion made by a camera and the direction of sound sources. These are each core problems in machine perception, but are largely studied separately, often using supervised methods that rely on difficult-to-acquire labeled training data, such as annotated sound directions. We take inspiration from self-supervised approaches to structure from motion [101], which learn to estimate 3D structure and camera pose by solving both tasks simultaneously.

Analogously, we propose a problem we call *sound localization from motion* (SLfM): jointly estimating camera rotation from images and the sound direction from binaural audio. By solving both tasks simultaneously, we avoid the need for labeled training data. Our models provide each other with self-supervision: a visual model predicts the rotation angle between pairs of images, while an audio model predicts the azimuth of sound sources. We force their predictions to agree with one another, such that changes in rotation are consistent with changes in sound direction and binaural cues. After training, the models can be deployed independently, without multimodal data at test time.

This is a challenging task that requires perceiving motion in images and binaural cues in audio. Our second contribution is a method for learning representations that are well-suited to this task through *cross-view binauralization*. We train a network to convert mono to binaural sound for one viewpoint, given an audio-visual pair sampled from another

viewpoint. Since the sound source is not necessarily visible in the images, the only way to successfully solve this pretext task is by analyzing the changes in the camera pose and predicting how they affect the sound direction.

All components of our model are entirely self-supervised and are trained solely on unlabeled audio-visual data. Our results suggest that paired audio-visual data provides a useful and complementary signal for learning about geometry. In contrast to other audio or visual self-supervised pose estimation methods, we obtain supervision from abundantly available audio data, thus avoiding the need of 3D ground truth or correspondences between pixels [101, 103] or audio samples [17]. Through experiments, we show:

- Paired audio-visual data provides a supervisory signal for pose estimation tasks.
- We obtain competitive performance with state-of-the-art self-supervised sound localization methods [17].
- We obtain strong rotation estimation performance, and our model generalizes to Stanford2D3D [6] dataset, where it is competitive with classic sparse feature matching methods.
- The features we learn through our pretext task outperform other representations for our downstream tasks.

2. Related Work

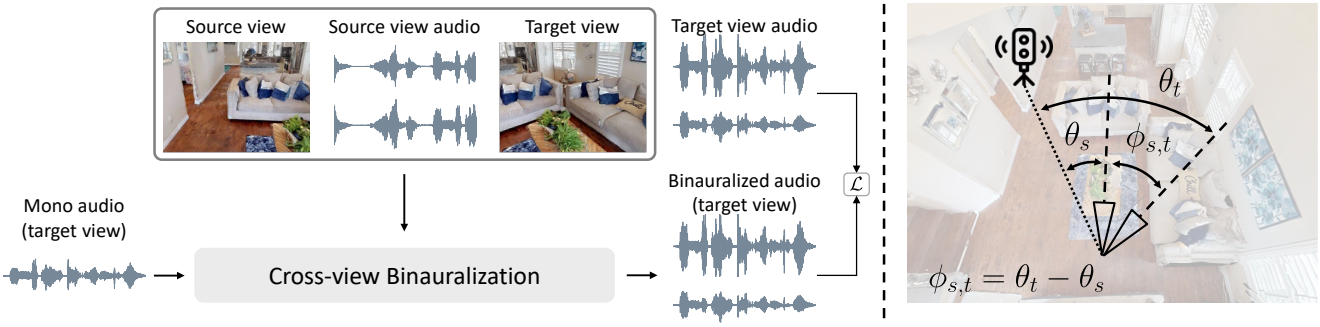
Audio for spatial perception. Recent works have explored the use of sound for spatial understanding. Purushwalkam *et al*. [69] reconstructed floor plans in simulated environments [12]. Chen *et al*. [18] used ambient sounds from environments to learn about scene structures. Konno *et al*. [46] integrated sound localization to visual SfM while do not jointly learn them. Other work learns representations for spatial audio-visual tasks. Yang *et al*. [94] predicted whether stereo channels are swapped in a video, and Morgado *et al*. [61] solved a spatial alignment task. The learned representations are then used to improve localization, up-mixing, and segmentation models. In contrast, we learn camera pose and sound localization solely from self-supervision, obtaining angular predictions without labeled data. Other work uses echolocation sounds to learn representations [27, 93] and predict depth maps [19, 68] and estimate camera poses [93] using labeled data. In contrast, our proposed approach jointly learns binaural sound localization and camera pose through passive audio sensing, without supervision.

Acoustic synthesis and spatialization. Researchers have explored visually-guided sound synthesis [26, 31, 40] and text-guided audio synthesis [47, 92, 38]. Additionally, researchers have investigated generating realistic environmental acoustics using visual information [11, 82, 15, 55]. Chen *et al*. [14] introduced the novel-view acoustic synthesis task, which synthesizes binaural sound at the target view using audio and visual information from a source view. Liang *et al*. [48] proposed an audio-visual neural field in real-world audio-visual scenes. Many recent works have proposed to generate spatial audio from mono audio using visual cues [62, 28, 72, 90, 49, 100, 30], or the relative pose between sound sources and the receiver [76, 39]. Inspired by these works, our feature learning approach learns spatial representations through an audio prediction task.

Binaural sound localization. Humans have the ability to localize sound sources from binaural sound [78]. Traditional approaches estimate interaural time delays via cross-correlation using hand-crafted features [45], factorization methods [80], or loudness differences between ears [75, 87]. Chen *et al*. [17] adapted methods from self-supervised visual tracking to the problem of binaural sound localization. Similarly, we estimate direction through self-supervision. However, we obtain our supervision through cross-modal supervision from vision instead of from correspondence cues. Moreover, we also obtain visual camera rotation estimation through our learning process. Francl [25] learned representations of sound location with a contrastive loss where positive and negative examples are selected based on the extent of head movements. Other works have used supervised learning techniques with labeled data to localize sound sources in reverberant environments [1, 86, 91]. Unlike these methods, our model learns 3D sound localization without labels.

Camera pose estimation. Traditional methods for camera pose estimation are based on finding correspondences between images and then solving an optimization problem using constraints from multi-view geometry [34]. These include structure from motion methods that estimate full pose [81] and camera rotation [8]. Recent methods have directly predicted camera pose using neural networks, including methods that use photos [70, 43, 57, 42, 53] or RGB-D scans [96, 95, 21, 22]. Our setup is similar to work that learns relative camera poses from sparse views [41, 9], and we use their network architectures. However, we learn the camera pose through cross-modal supervision from audio, rather than from labels. Our work is also closely related to methods that learn structure from motion through self-supervision [101, 103], such as by jointly learning models that perform depth and camera pose estimation with photoconsistency constraints. In contrast, our visual model's learning signal comes solely from audio-based supervision, and we jointly learn audio localization.

Audio-visual learning. A number of works have focused on learning multimodal representations for audio and vision, taking into account semantic correspondence and temporal synchronization [65, 89, 7, 63, 66, 2, 58]. Other approaches study audio-visual sound localization [37, 59, 16, 60], source separation [54, 29, 56, 84], active speaker detection [3, 83, 5], navigation [12, 13, 10] and forensics [102, 32, 23]. Our focus, in contrast, is on utilizing multi-view audio-visual signals to learn geometry.



(a) Cross-view audio prediction

(b) Sound localization from motion

Figure 2: **Method overview.** (a) We learn a feature representation by predicting how changes in images lead to changes in sound in a cross-view binauralization pretext task. We convert mono sound to binaural sound at a target viewpoint, after conditioning the model on observations from a source viewpoint. (b) We use the representation to jointly solve two pose estimation tasks: visual rotation estimation and binaural sound localization. We train the visual rotation angle, $\phi_{s,t}$, to be consistent with the difference in predicted sound angles θ_s and θ_t.

3. Sound Localization from Motion

We address the *sound localization from motion* (SLfM) task: predicting the azimuth of a sound source from binaural audio and camera rotation from two images. First, we present a self-supervised representation that can be used to solve this downstream task. Then, we show how the representation can be used to solve the task.

3.1. Learning representation via spatialization

We learn an audio-visual representation that conveys spatial information. To do this, we solve a *cross-view binauralization* task: converting mono sound to stereo for one viewpoint, given an audio-visual pair sampled from another viewpoint (Fig. 2a). In order to successfully solve the task, a model must implicitly estimate the sound direction in the source view and predict how the change in viewpoint will affect the sound in the target view. A key difference between this task and traditional, single-view binauralization [28] is that the sound source *need not be visible* in any of the images. Hence, the task cannot be solved from the target audio-visual pair alone.

We binauralize the sound $\mathbf{a}_t$ at the target view through an audio predictor $\mathcal{F}_\theta$. To make our representation suitable for downstream tasks, we factorize it into visual features $f_v(\mathbf{v}_s, \mathbf{v}_t)$, which are intended to create features relevant to relative pose, and audio features $f_a(\mathbf{a}_s)$, which capture sound localization cues. We predict the binauralized audio $\hat{\mathbf{a}}_t$ at the target view from mono audio $\bar{\mathbf{a}}_t$, the visual change of $(\mathbf{v}_s, \mathbf{v}_t)$ and binaural sound $\mathbf{a}_s$ heard at the source viewpoint:

$$\hat{\mathbf{a}}_t = \mathcal{F}_\theta\left(\bar{\mathbf{a}}_t, f_v(\mathbf{v}_s, \mathbf{v}_t), f_a(\mathbf{a}_s)\right). \tag{1}$$

We represent audio $\mathbf{a}$ as a spectrogram $\mathbf{A}$ using short-time Fourier transform (STFT). Following Gao *et al*. [28], given the mix of stereo audio $\bar{\mathbf{A}}_t = \text{STFT}(\mathbf{a}_t^L + \mathbf{a}_t^R)$, we predict the difference of two channels $\mathbf{A}_t = \text{STFT}(\mathbf{a}_t^L - \mathbf{a}_t^R)$. We optimize the $L1$ loss between predicted spectrogram $\hat{\mathbf{A}}_t$ and ground-truth spectrogram $\mathbf{A}_t$:

$$\mathcal{L}_{\text{pretext}} = ||\hat{\mathbf{A}}_t - \mathbf{A}_t||_1. \tag{2}$$

Multi-view binauralization. Following Zhou *et al*. [101], we improve our representations by binauralizing sounds at N different target viewpoints, using observations from a single source viewpoint s. We hypothesize that jointly solving spatialization problems from a single viewpoint for multiple target viewpoints would require the model to make more accurate predictions of sound source locations, thereby improving the estimation of view changes:

$$\mathcal{L}_{\text{pretext}} = \frac{1}{N}\sum_{i}^{N} ||\mathcal{F}_\theta\left(\bar{\mathbf{A}}_i, f_v(\mathbf{v}_s, \mathbf{v}_i), f_a(\mathbf{a}_s)\right) - \mathbf{A}_i||_1. \tag{3}$$

3.2. Estimating pose and localizing sounds

We now address the problem of learning models for sound localization and pose estimation, using our self-supervised audio-visual features. Given two views, we predict sound directions and relative rotation. We train the model to make these two predictions consistent with one another, while using simple binaural constraints to resolve ambiguities.

We are given images $\mathbf{v}_s$ and $\mathbf{v}_t$ (rotated views recorded at the same position) and learned visual embedding f_v. We predict the scalar rotation angle $\phi_{s,t}$ via the encoder g_v:

$$\phi_{s,t} = g_v\left(f_v(\mathbf{v}_s, \mathbf{v}_t)\right), \; R_{s,t} = \begin{bmatrix} \cos\phi_{s,t} & -\sin\phi_{s,t} \\ \sin\phi_{s,t} & \cos\phi_{s,t} \end{bmatrix}, \tag{4}$$

where $R_{s,t}$ is 2D rotation matrix of $\phi_{s,t}$. Following common practice in indoor scene reconstruction, we give the camera a fixed downward tilt [99, 71, 98] and only estimate azimuth [41]. This is also a common assumption in audio localization [1, 86], since azimuth has strong binaural cues.

Similarly, we predict the azimuths of the sound sources using audio features and the encoder g_a:

$$\theta_i = g_a\left(f_a(\mathbf{a}_i)\right), \quad \mathbf{r}_i = \begin{bmatrix}\cos\theta_i & \sin\theta_i\end{bmatrix}^{\mathsf{T}}, \tag{5}$$

where we represent the azimuth as a vector $\mathbf{r}_i$

Cross-modal geometric consistency. When the camera is rotated, the sound source ought to rotate in the opposite direction (Fig. 2b). For example, a 30° clockwise camera rotation should result in a 30° counterclockwise rotation in sound direction. Such a constraint could be converted into a loss:

$$\mathcal{L}_{\text{rot}} = \|\mathbf{r}_s - R_{s,t}\mathbf{r}_t\|^2. \tag{6}$$

However, a well-known ambiguity exists in binaural sound perception: one cannot generally tell whether a sound is in front of the view, or behind them. To address that, we use permutation invariant training [97] (PIT), and allow the model to use either the predicted sound direction or its reflection about the x axis without penalty. This results in the loss:

$$\mathcal{L}_{\text{geo}} = \min_{\substack{\hat{\mathbf{r}}_s \in \{\mathbf{r}_s, Q\mathbf{r}_s\} \\ \hat{\mathbf{r}}_t \in \{\mathbf{r}_t, Q\mathbf{r}_t\}}} \|\hat{\mathbf{r}}_s - R_{s,t}\hat{\mathbf{r}}_t\|^2, \tag{7}$$

where $Q = \begin{bmatrix}1 & 0 \\ 0 & -1\end{bmatrix}$ reflects the sound direction.

As a consequence of this ambiguity, there are also two possible solutions for the visual rotation model, since the visual rotation matrices can be mirrored about the x axis. For example, one can create a solution with equal loss by multiplying the rotations and sound directions by Q. We discuss this ambiguity in more depth in Sec. 4.4.

Incorporating binaural observations. Without additional constraints, the solution is ambiguous, and may collapse into a trivial solution (*e.g.*, predicting zeros for all three angles).[1] To avoid this, we force the model to agree with a simple binaural cue based on interaural intensity difference (IID). We estimate whether the sound is to left or right of the viewer, based on whether the sound is louder in the left or right microphone: $d = \text{sign}(\log\left|\frac{\mathbf{A}^L}{\mathbf{A}^R}\right|)$, where $|\mathbf{A}|$ is the magnitude of the spectrogram $\mathbf{A}$. We perform this left/right test at each timestep in the spectrogram and then pool via majority voting (see Appendix A.4 for details). We penalize predictions that are inconsistent with these "left or right" observations:

$$\mathcal{L}_{\text{binaural}} = \mathcal{L}_{\text{BCE}}\left(\sin\theta_i, d_i\right), \tag{8}$$

where $\mathcal{L}_{\text{BCE}}$ is binary cross entropy loss.

[1]We note that work in self-supervised structure from motion has similar ambiguity [101], and deals with them by adding analogous constraints, such as photometric consistency.

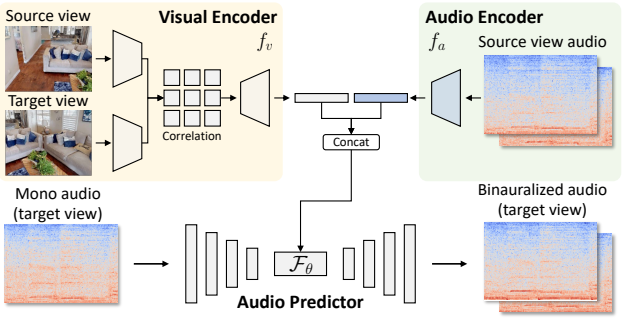


Figure 3: **Cross-view binauralization architecture.** We take a mono spectrogram as input and fuse the audio and visual features from the audio and visual encoder respectively to synthesize the binaural spectrogram at the target viewpoint.

Encouraging symmetry. To help regularize the model, we also add symmetry constraints. For sound localization, swapping the left and right channels of the audio ought to result in a prediction in the opposite direction since the binaural cues are reversed. For rotation estimation, the relative pose between images s and t should invert the pose from t to s. We encourage both constraints via a loss:

$$\mathcal{L}_{\text{sym}} = |\theta + \theta_{\text{flip}}| + |\phi_{s,t} + \phi_{t,s}|, \tag{9}$$

where θ_{flip} is the prediction of sound angle θ using audio with swapped audio channels, and $\phi_{s,t}$ and $\phi_{t,s}$ are the predicted rotations between cameras s and t.

Overall loss. We combine these constraints to obtain an overall loss:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{geo}} + \lambda_2\mathcal{L}_{\text{binaural}} + \lambda_3\mathcal{L}_{\text{sym}}, \tag{10}$$

where λ_1, λ_2 and λ_3 are the weight for geometric, binaural, and symmetric losses, respectively.

4. Experiments

We have introduced a self-supervised method to learn camera pose and sound localization from audio-visual data. In experiments, we first evaluate how well our learned representation captures spatial information. We then evaluate how well our method learns camera pose and sound localization by comparing it with baselines. Finally, we show generalization to indoor panorama images Stanford2D3D [6] and in-the-wild binaural audio [17].

4.1. Implementations

Visual pose encoder. We follow recent pose estimation work [9, 41] and build a Siamese-style visual pose network f_v with ResNet-18 [36] as the backbone. We compute dense 4D correlation volumes between the features from the third residual layer and then encode them by convolution layers.

We resize images to 320×240 and encode a pair of images into 512-d features. We use an MLP g_v to map visual features to 1-d logits for our SLfM models.

Binaural audio encoder. We obtain binaural audio embeddings $f_a(\cdot)$ using ResNet-18 [36] that operates on spectrograms. We covert the two-channel waveform of length L to a spectrogram representation of size $256 \times 256 \times 4$ using short-time Fourier transform (STFT), where we keep both the magnitude and phase of spectrograms. We extract 512-d features of binaural sound with f_a and map them to 1-d logits using an MLP g_a.

Audio prediction model. We adopt the light-weighted audio-visual U-Net [28] to perform binauralization. We feed in spectrograms of size $256 \times 256 \times 2$ and predict the target spectrograms. We concatenate the visual pose features $f_v(\cdot)$ and audio features $f_a(\cdot)$ at the bottleneck of U-Net. We show the architecture of our models in Fig. 3. Please see Appendix A.4 for more implementation details.

4.2. Dataset

Since there is no public multi-view audio-visual dataset with camera poses and sound direction ground truth, we use the SoundSpaces 2.0 platform [15] to create a dataset. Our 3D scenes come from Habitat-Matterport 3D dataset (HM3D) [74], which is a large dataset of real 3D scenes. This setup allows us to have photorealistic images and high-quality spatial audio with real-world acoustics phenomenon (*e.g.*, reverberation), as well as providing the ground-truth camera pose and sound directions that can be used for evaluation. We call this dataset HM3D-SS.

We generate binaural Room Impulse Responses (RIRs) and images with a 60° field of view, using 100 scenes of HM3D [74]. For each audio-visual example, we randomly place sound sources in the scene with a height range of $(0.7, 1.7)$ meters, and sample 4 different rotated viewpoints at one location within 4 meters. The rotations are limited to $(10^\circ, 90^\circ)$ relative to the source viewpoints. We follow the standard practice to set the height to agents to be 1.5m and lock a downward tilt angle [41, 98, 71, 99]. We render the binaural RIRs and images given the position of agents and sound sources. We obtain binaural audio by convolving binaural RIRs with mono audio samples from LibriSpeech [67] and Free Music Archive [20]. To ensure that the evaluation tests the model's pose estimation abilities, rather than its ability to visually localize sound sources, the sound sources are not visible on screen.

We create 50K audio-visual pairs from 200K viewpoints. The audio was rendered with average reverberation of $\text{RT}_{60} = 0.4$s (see Appendix A.4 for details). We divided our data into 81/9/10 scenes for the train/val/test, respectively.

4.3. Evaluating the learned representation

First, we directly evaluate the quality of our learned features for rotation estimation and sound localization via linear probing with labeled data (rather than learning them jointly through self-supervision).

Baselines and ablations. We compare our model with several baselines that use alternative pretext tasks: 1) **AVSA** [61]: it learns spatial cues by training a model to spatially align video and audio clips extracted from different viewing angles. We adapt this model to our dataset and train with 4 different views; 2) **RotNCE** [25]: it applies contrastive learning on the audio from different angles and uses annotations of the agent's rotation to select positive and negative samples, which results in learning audio spatial representation. For baselines, we use the same architecture for feature extractors to ensure fair comparisons.

To determine if we utilize visual and audio features from different views to solve the binauralization task, we also study some variants of our models: 1) **Ours-NoA**: we only provide visual features for the binauralization task; 2) **Ours-NoV**: which only uses audio from the other view to spatialize sounds; 3) **Ours-GTRot**: we provide ground-truth rotation embedding instead of features from visual frames.

Besides the mono-to-binaural task, we also experiment with another objective: predicting the right channel from the left channel. We train our **L2R** model with the same setup

	Model	Audio Loc. Acc (%) ↑	Camera Rot. Acc (%) ↑
LibriSpeech	Random feature	4.8	4.7
	ImageNet [36]+Random	–	56.3
	RotNCE [25]	50.9	–
	AVSA [61]	71.2	6.5
	Ours–NoA	–	9.6
	Ours–NoV	53.9	–
	Ours–GTRot	70.6	–
	Ours–L2R (3 views)	**78.5**	76.1
	Ours (2 views)	74.5	80.0
	Ours (3 views)	75.4	**81.3**
	Supervised	81.5	95.8
FreeMusic	Random feature	6.0	4.7
	ImageNet [36]+Random	–	56.3
	RotNCE [25]	46.3	–
	AVSA [61]	66.5	6.7
	Ours–L2R (3 views)	**72.0**	76.5
	Ours (2 views)	67.5	76.2
	Ours (3 views)	67.5	**81.1**
	Supervised	77.1	95.8

Table 1: **Downstream task performance on HM3D-SS dataset.** We report linear probe performance on the audio localization and camera rotation downstream tasks.

	Model	Audio angle MAE (°) ↓	Camera angle MAE (°) ↓
LibriSpeech	Chance	40.28	29.41
	SIFT [52]	–	12.2
	Ours w/o Reflect	28.08	26.80
	Ours–L2R	3.22	0.80
	Ours	**3.17**	**0.77**
	Ours–Front	4.48	3.08
	Ours–GTRot	1.83	–
	Superglue [79]	–	2.47
	Supervised	1.71	0.46
FreeMusic	Chance	40.81	29.41
	SIFT [52]	–	12.2
	Ours w/o Reflect	28.24	26.81
	Ours–L2R	**3.18**	**0.81**
	Ours	3.37	0.84
	Ours–Front	3.99	2.49
	Ours–GTRot	1.96	–
	Superglue [79]	–	2.47
	Supervised	2.50	0.46

Table 2: **Sound localization from motion results on HM3D-SS.** We evaluate our SLfM models on each modality independently.

as our M2B model. (Please see Appendix A.2 for pretext results.)

Downstream tasks. We assess the quality of spatial representations we learned from our pretext tasks in two downstream tasks: relative camera rotation and 3D sound localization. We formulate them as classification problems, where angles are categorized into 64 bins, and we use accuracy as the evaluation metric. To evaluate the learned features, we freeze them and train a linear classifier on the downstream tasks. We compare the performance of our features with those learned from RotNCE [25], AVSA [61], ImageNet [35], and random features, and report the results in Tab. 1. Our approach outperforms the baselines in both tasks, indicating that we learn better spatial representations. Furthermore, our linear probe models show comparable performance against the supervised method which can be regarded as approximate upper bounds for our models, suggesting that our pretext tasks help learn a useful representation.

Emerging camera pose from audio prompting. To help understand the strong performance of our self-supervised features, we asked whether we could use the cross-view binauralization model *alone* to estimate camera rotation. Inspired by prompting in vision and language models [73], we obtain rough estimates of camera rotation by providing our model with carefully-provided inputs. Given a pair of images $(\mathbf{v}_s, \mathbf{v}_t)$, we create a synthetic binaural audio *prompt*, $\mathbf{a}_s$. We

Figure 4: **Qualitative results on real-world examples.** We show our predictions on Stanford2D3D [6] and In-the-wild audio [17]. Green denotes accurate predictions.

then ask our model to generate the binaural sound $\hat{\mathbf{a}}_t$ for the target viewpoint. By analyzing the IID cues in $\hat{\mathbf{a}}_t$, we can estimate the model's implicitly predicted camera pose. To do this, we find the nearest neighbor of our generated audio $\hat{\mathbf{a}}_t$, using a database of synthetically generated audio with known sound directions. Please refer to Appendix A.1 for more details. Our method achieves the mean absolute error of $\mathbf{9.13}^\circ$ on HM3D-SS dataset where chance is 29.41°. Our approach can also generalize to Stanford2D3D [6] dataset, where we can achieve mean absolute error of $\mathbf{9.93}^\circ$.

4.4. Evaluating SLfM

We conducted evaluations on our sound localization from motion (SLfM) models on the HM3D-SS dataset. We use the mean absolute error of angle in degrees as evaluation metrics. To avoid sound field ambiguity, we filtered out samples with sound angles outside of $(-90^\circ, 90^\circ)$ for the evaluation set. When training our SLfM models, we use our best-performing features, *i.e.*, pretext tasks trained with 3 views. We freeze learned audio and visual features and only train multi-layer perceptrons on top of them and evaluate them independently on each modality.

Baselines and ablations. For relative camera pose estimation, we compare our models with sparse feature matching using SIFT [52] and SuperGlue [79], followed by rotation fitting. We use those methods to detect key points, run Lowe's ratio test, and use RANSAC with five-point algorithm to recover the camera pose [52, 34, 64, 24]. For the sound localization task, we compared with time delay estimation methods: the popular GCC-PHAT [45] and the recent self-supervised method StereoCRW [17]. As far as we know, there are no other baselines that can estimate poses and sound source directions without labels.

We also compared several variations of our method, including **Ours–Front**, where we filter out the samples with sound sources behind the viewers to remove binaural ambiguity during training, and an oracle model **Ours–GTRot**, which uses ground-truth rotation angles instead.

Results. We show our results in Tab. 2. Our models can predict the azimuths of sound sources, obtaining strong performance without any labels. Without providing reflection invariance (Eq. (7)), the model failed to learn reasonable

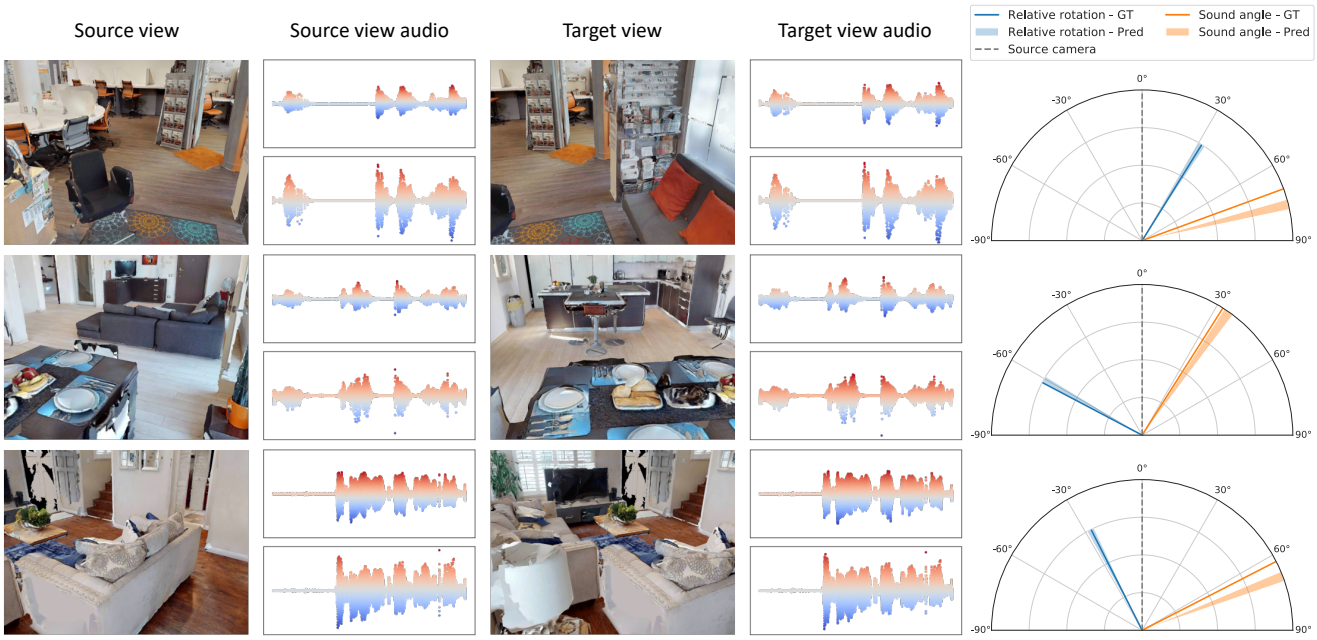


Figure 5: **Qualitative results on HM3D-SS.** Given two images and their corresponding audio at the source viewpoints, our approach can individually predict the relative camera pose and localize sound locations accurately. We visualize our camera rotation prediction with blue bar and our sound angle prediction with orange bar. To highlight the subtle differences in the waveforms, we color code the amplitude. Please see the project webpage for more video results.

In the wild audio [17]	
Model	Acc (%) ↑
GCC-PHAT [45]	77.2
IID [17]	75.4
IID (Ours)	82.1
MonoCLR [17]	**87.4**
StereoCRW [17]	87.2
Ours–L2R	84.9
Ours	84.0

Stanford2D3D [6]		
Model	Rot Err. (°) ↓	
	Mean	Med.
SIFT [52]	16.4	**0.06**
Superglue [79]	5.07	0.07
Ours–L2R	**1.12**	0.71
Ours	1.14	0.67

Table 3: **Evaluation of our SLfM models on the real-world data.** We evaluate our audio localization model on the In-the-wild audio [17] (left) and our camera pose model on the Stanford2D3D [6] (right). *Rot.* denote camera rotation.

geometric due to the audio ambiguity. Our self-supervised model achieves comparable performance against our oracle model (Ours-GTRot) and supervised models, indicating we estimate camera poses and localize sound accurately. We show some qualitative results on HM3D-SS in Fig. 5 with LibriSpeech samples [67].

Generalization to other datasets. We further demonstrate the generalization ability of our models by experimenting with out-of-distribution, real-world data. We evaluate our camera rotation model on Stanford2D3D [6] with real indoor RGB images (Tab. 3). We obtain image pairs by cropping from panoramas. Although our model is trained on renderings of HM3D [74], it obtains strong generalization ability. Compared with rotation estimation based on SIFT [52] and SuperGlue [79], our model is significantly better on mean rotation error. We also report median rotation error to be consistent with prior works [9, 41]. However, SIFT [52] and Superglue [79] have a very low median error. This is likely due to the all-or-nothing nature of feature matching-based approaches, which either produce highly accurate predictions if the matches are correct (especially for "easy" cases with small amounts of rotation) or else produce gross errors.

We also evaluate our sound localization model on the in-the-wild binaural audio [17]. We use binary accuracy as the metric for left-or-right direction classification accounting for the fact that microphone baselines are unknown for internet videos. For an apples-to-apples comparison to prior work, we retrained our model using 0.51s length of audio for the binaural audio encoder $f_a(\cdot)$. As shown in Tab. 3, our model obtains similar performance to StereoCRW, a state-of-the-art self-supervised time delay method, suggesting we have a strong capability for sound localization. We show qualitative results in Fig. 4. We also perform both tasks on a self-recorded video (Fig. 6) of a rotating camera and a binaural microphone. We show the mean and standard deviation of predictions in 1.0s windows.

Handling ambiguity. In binaural sound perception, there is a fundamental ambiguity that whether the sound is in front of or behind us. It leads to multiple solutions with equal loss

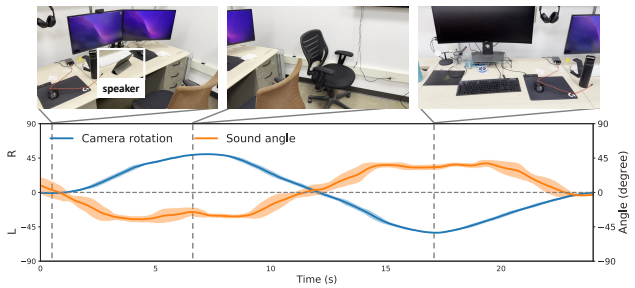


Figure 6: **Real video.** We play music using a speaker and record a video using iPhone with a binaural microphone (different from training examples). We show the predicted camera rotation and sound direction over time. They change smoothly and match our camera motion. Please see our project webpage for video results.

in our model, mirroring sound sources and negating rotation angles with flipped z axis (Fig. 7). These two solutions differ in that a visual rotation angle either indicates a clockwise or counterclockwise rotation. For evaluation, we convert "backwards" counterclockwise predictions to clockwise predictions by simply providing the model with pairs of input frames, then negating the model's outputs if the angle is the opposite of the expected direction[2].

4.5. Generalization to more complex scenarios

We investigate whether our approach generalizes to more complex scenarios, such as with multiple sound sources or when translation is included in camera motion.

Multiple sound sources. We evaluate our representations and SLfM models on more complex scenes containing multiple sound sources. We train our models with two source sources placed in the scenes. One of them is dominant and our sound localization target. We report the results in Tab. 4 and Tab. 5. Our model learns better representations than baselines and achieves accurate predictions of the azimuth of the source and camera pose even in challenging scenarios with multiple sound sources.

[2]Similar to ambiguities in SfM where reconstructions can be reoriented such that the sky is in the positive y direction [34, 33, 4].

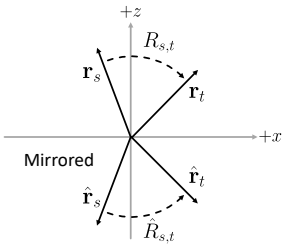


Figure 7: **Mirror ambiguity.** The $+z$ and $+x$ axes represent forward and rightward directions. Both solutions have the same loss.

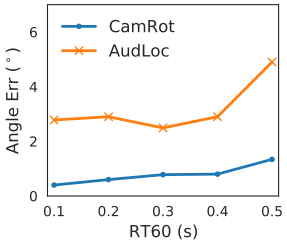


Figure 8: **Robustness to reverberation.** We study the effect of reverberation on our SLfM model.

Model	Multi-source Acc (%) ↑		Small trans. Acc (%) ↑	
	Aud.	Rot.	Aud.	Rot.
Random feature	4.5	4.7	5.7	3.8
ImageNet [36]+Random	–	56.3	–	23.0
RotNCE [25]	35.8	–	–	–
AVSA [61]	55.9	6.7	65.4	6.1
Ours (3 views)	**59.2**	**77.9**	**72.2**	**49.6**
Supervised	74.5	95.8	81.6	63.5

Table 4: **Generalization to more complex scenarios.** We evaluate the audio and visual features learned from more complex scenarios via linear probing. *Aud.* and *Rot.* denote audio localization and camera rotation respectively.

Model	Multi-source MAE (°) ↓		Small trans. MAE (°) ↓	
	Aud.	Rot.	Aud.	Rot.
Chance	40.46	29.41	40.28	29.41
SIFT [52]	–	12.2	–	12.2
Ours	**7.67**	**0.71**	**6.28**	**1.04**
Ours–GTRot	5.81	–	4.24	–
Superglue [79]	–	2.47	–	2.47
Supervised	3.60	0.46	1.71	0.46

Table 5: **SLfM results on more complex scenarios.** We evaluate our model on the version of HM3D-SS with two sound sources. We also evaluate our model trained with small translations on rotation-only examples.

Translation in camera motions. To study how small translations in the camera motion could affect our models, we generate 50K pairs of audio-visual data with both rotation and translation change, limiting the uniformly sampled translation to 0.5 meters. We train our models with LibriSpeech samples [67]. For our linear probing evaluation, we measure the ability of features to handle complex examples by testing on the dataset that has translation. For our SLfM model, we study our ability to learn from noisy data, thus we evaluate it on the rotation-only examples. As the results are shown in Tab. 4 and Tab. 5, we successfully learn useful features and obtain accurate rotation and sound direction predictions despite the presence of translation. Since we jointly learn the audio and visual representations, it can negatively impact the learning of one modality when another one becomes harder.

4.6. Ablation study

Robustness to reverberation. We study how our models perform under the different reverberation configurations. We used SoundSpaces [15] to create audio with average reverberation $\text{RT}_{60} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ while keeping the visual signals the same. We train our model on each set-

Model	Losses			Audio angle	Camera angle
	$\mathcal{L}_{\text{geo}}$	$\mathcal{L}_{\text{binaural}}$	$\mathcal{L}_{\text{sym}}$	MAE (°) ↓	MAE (°) ↓
Ours	✓			37.60	29.20
	✓		✓	37.52	29.17
	✓	✓		3.58	6.99
	✓	✓	✓	**3.17**	**0.77**

Table 6: **Ablation experiments on our SLfM losses.** We evaluate our SLfM models with different combinations of losses.

ting. As shown in Fig. 8, our performance decreases as the level of reverberation increases, where audio becomes more challenging during both training and testing.

Losses for SLfM. We study the necessity of our proposed loss functions in Tab. 6. Our models fail to learn accurate pose estimation without binaural loss or symmetric loss. It highlights the crucial role of these losses.

5. Conclusion

In this paper, we proposed the *sound localization from motion* (SLfM) problem, and provided a self-supervised method for solving it. We also presented a method for learning audio-visual features that convey sound directions and camera rotation, which we show are well-suited to solving the SLfM task. Despite learning our models solely from unlabeled audio-visual data, we obtain strong performance on a variety of benchmarks, including rotation estimation on the Stanford2D3D [6] dataset and "in the wild" sound direction estimation [17]. Our results suggest that the subtle correlations between sights and binaural sounds that result from rotational motion provide a useful (and previously unused) learning signal. We see our work as opening new directions in self-supervised geometry estimation and feature learning that use sound as a complementary source of supervision. We will release code, data, and models upon acceptance.

Limitations and Broader Impacts. Our work has several limitations. First, while we evaluate our models on real images and sounds, we train on data from simulators, due to a lack of available relevant data. We note that this is common practice in visual 3D reconstruction [41, 88, 50]. Second, we assume that both the 3D scene and sound sources are stationary. Third, we do not evaluate our model on extreme viewpoint changes [53], which requires reasoning about images that have little or no overlap, and is largely solved using supervised methods.

Acknowledgements. We would like to thank Ang Cao, Tiange Luo, and Linyi Jin for their helpful discussion. We thank Changan Chen for his help with SoundSpaces 2.0. This work was funded in part by DARPA Semafor and Sony. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466. IEEE, 2018. 2, 3

[2] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10586, 2022. 2

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020. 2

[4] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 8

[5] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 2

[6] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 4, 6, 7, 9

[7] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020. 2

[8] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74:59–73, 2007. 2

[9] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14566–14575, 2021. 2, 4, 7

[10] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021. 2

[11] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 2

[12] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 2

[13] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*, 2020. 2

[14] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. *arXiv preprint arXiv:2301.08730*, 2023. 2

[15] 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 preprint arXiv:2206.08312*, 2022. 2, 5, 8, 14, 15

[16] 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. 2

[17] 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. 2, 4, 6, 7, 9, 14

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

[19] 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

[20] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016. 5

[21] Mohamed El Banani, Luya Gao, and Justin Johnson. Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In *CVPR*, 2021. 2

[22] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *ICCV*, 2021. 2

[23] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. *arXiv preprint arXiv:2301.01767*, 2023. 2

[24] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 6, 14

[25] Andrew Francl. *Modeling and Evaluating Human Sound Localization in the Natural Environment*. PhD thesis, Massachusetts Institute of Technology, 2022. 2, 5, 6, 8

[26] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020. 2

[27] 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

[28] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2, 3, 5, 15

[29] 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. 2

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

[31] Sanchita Ghose and John Jeffrey Prevost. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia*, 23:1895–1907, 2020. 2

[32] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022. 2

[33] R. I. Hartley. Chirality. *International Journal of Computer Vision (IJCV)*, 1998. 8

[34] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 2, 6, 8

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, 2015. 6

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5, 8

[37] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[38] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023. 2

[39] Wen Chin Huang, Dejan Markovic, Alexander Richard, Israel Dejene Gebru, and Anjali Menon. End-to-end binaural speech synthesis. *arXiv preprint arXiv:2207.03697*, 2022. 2

[40] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021. 2

[41] Linyi Jin, Shengyi Qian, Andrew Owens, and David F Fouhey. Planar surface reconstruction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12991–13000, 2021. 2, 3, 4, 5, 7, 9

[42] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. *arXiv*, 2022. 2

[43] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 2

[44] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2015. 15

[45] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976. 2, 6, 7

[46] Takashi Konno, Kenji Nishida, Katsutoshi Itoyama, and Kazuhiro Nakadai. Audio-visual sfm towards 4d reconstruction under dynamic scenes. 2022. 2

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

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

[49] 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*, volume 35, pages 2056–2063, 2021. 2

[50] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. Neurmips: Neural mixture of planar experts for view synthesis. In *CVPR*, 2022. 9

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 15

[52] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 6, 7, 8

[53] Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *CVPR*, 2022. 2, 9

[54] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 275–285, 2021. 2

[55] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *arXiv preprint arXiv:2206.04006*, 2022. 2

[56] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 551–569. Springer, 2022. 2

[57] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, pages 675–687. Springer, 2017. 2

[58] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. *arXiv preprint arXiv:2209.13583*, 2022. 2

[59] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *arXiv preprint arXiv:2209.09634*, 2022. 2

[60] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 218–234. Springer, 2022. 2

[61] 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, 5, 6, 8

[62] 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

[63] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2

[64] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004. 6

[65] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 2

[66] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[67] 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, 7, 8, 14, 15

[68] 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

[69] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192, 2021. 2

[70] Shengyi Qian, Linyi Jin, and David F Fouhey. Associative3d: Volumetric reconstruction from sparse views. In *ECCV*, 2020. 2

[71] Shengyi Qian, Alexander Kirillov, Nikhila Ravi, Devendra Singh Chaplot, Justin Johnson, David F Fouhey, and Georgia Gkioxari. Recognizing scenes from novel viewpoints. *arXiv preprint arXiv:2112.01520*, 2021. 3, 5

[72] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1930–1939, 2021. 2

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[74] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 5, 7

[75] Lord Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907. 2

[76] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021. 2

[77] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 15

[78] Lord Rayleigh O.M. Pres. R.S. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1907. 2

[79] 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. 6, 7, 8

[80] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986. 2

[81] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[82] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 286–295, 2021. 2

[83] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 2

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

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 15

[86] Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Guy J Brown. End-to-end binaural sound localisation from the raw waveform. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455. IEEE, 2019. 2, 3

[87] DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006. 2

[88] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *ECCV*, 2022. 9

[89] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[90] 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

[91] Nelson Yalta, Kazuhiro Nakadai, and Tetsuya Ogata. Sound source localization using deep learning models. *Journal of Robotics and Mechatronics*, 29(1):37–48, 2017. 2

[92] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022. 2

[93] 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. 2

[94] 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

[95] Zhenpei Yang, Jeffrey Z Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, and Qixing Huang. Extreme relative pose estimation for rgb-d scans via scene completion. In *CVPR*, 2019. 2

[96] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In *CVPR*, 2020. 2

[97] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for

speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017. 4

[98] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *CVPR*, 2017. 3, 5

[99] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3, 5

[100] 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

[101] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1, 2, 3, 4

[102] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021. 2

[103] Yuliang Zou, Pan Ji, Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV*, pages 710–727. Springer, 2020. 2

A.1. Camera pose from audio prompting

We illustrate our prompting idea in Fig. 9. To create our audio prompts, we simulate 181 binaural RIRs at different angles from $[-90°, 90°]$ without reverberation using SoundSpaces [15] and render with audio signals from LibriSpeech [67]. We use the sound with an angle of $0°$ as the input prompt $\mathbf{a}_s$ (the source view audio) and mix it into mono audio as the input at the target viewpoint. We calculate the interaural intensity difference (IID) cues for the audio prompts $\mathbf{a}_i$ and generated audio $\hat{a}_t$. We use L1 distance between IID cues to find the nearest neighbors:

$$\arg\min_{\mathbf{A}_i} \left| \log_{10} \frac{\hat{\mathbf{A}}_t^L}{\hat{\mathbf{A}}_t^R} - \log_{10} \frac{\mathbf{A}_i^L}{\mathbf{A}_i^R} \right|, \tag{11}$$

where $\mathbf{A}_i = \text{STFT}(\mathbf{a}_i)$. We use ground truth annotations of sound directions from the nearest prompts to predict the camera rotation angles. We first obtain rotation prediction votes from 1024 audio prompts and use a RANSAC-like mode estimation [24, 17] to get the final prediction.

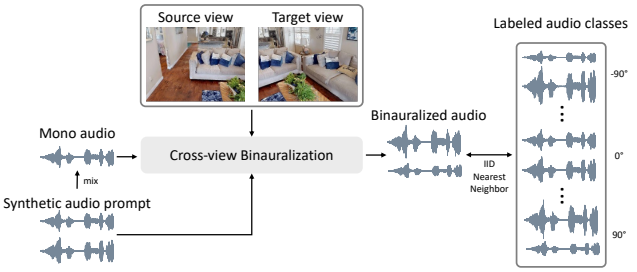


Figure 9: **Estimating camera pose from audio prompting.** We estimate camera rotation by providing our cross-view binauralization model with synthetically generated audio prompts. Given the sound that it predicts, we infer the camera angle. We do this by finding the nearest neighbor (using IID cues) to a database of synthetic sounds, each paired with their corresponding angle.

A.2. Additional experimental results

Evaluating pretext task. We also evaluate the performance of our model on the pretext task, which involves binauralizing sound at a novel microphone pose using sound from a different viewpoint and visual cues from both views as references. We use the STFT distance between the predicted and ground-truth spectrogram to measure the audio reconstruction performance. As the results are shown in Tab. 7, our model that incorporates both visual and audio features as input performs the best and is comparable to the model that receives ground truth rotation angles as inputs. This suggests that our model effectively uses the spatial information in both visual and audio signals to solve binauralization tasks, and encourages the network to learn useful representations. Moreover, the results show that training with more viewpoints improves the performance of the pretext task.

	Model	Input features $\mathcal{V}$	Input features $\mathcal{A}$	STFT distance ↓
Mono2Binaural	Random	✓	✓	0.368
	Ours (2 views)	✓		0.206
			✓	0.207
		✓	✓	0.161
	Ours–GTRot		✓	0.130
	Ours (3 views)	✓	✓	0.131
				0.125

Table 7: **Reconstruct performance of cross-view binauralization pretext task.** We report the STFT distance performance of variants of our models with different input features on HM3D-SS dataset with LibriSpeech samples [67]. $\mathcal{V}$ and A mean visual and audio features, respectively.

SLfM without pretraining. We further demonstrate the important role of the features learned from our cross-view binaural pretext task by training our SLfM model with random features. We show results in Tab. 8. We can see that the models perform better using our feature representations, which emphasizes the significance of our pretext task. Our SLfM model finetuned from random features achieves accurate predictions, highlighting that our proposed method successfully leverages the geometrically consistent changes between visual and audio signals.

Model	Init. feature	Audio angle MAE (°) ↓	Camera angle MAE (°) ↓
Ours	Random (freeze)	36.51	29.26
Ours	Random (finetune)	3.92	1.32
Ours	M2B (freeze)	3.17	0.77
Ours	M2B (finetune)	**2.77**	**0.76**

Table 8: **SLfM results with different features.** We evaluate our SLfM models trained with different feature initialization on HM3D-SS.

A.3. Ablation study

Robustness to reverberation. We also evaluate our representation under the influence of reverberation. We report linear probe performance on downstream tasks with average reverberation $\text{RT}_{60} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. As shown in Fig. 10, the results indicate a decrease in downstream performances as the level of reverberation increases, where audio becomes more challenging during both training and testing.

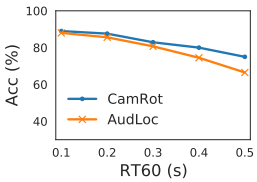


Figure 10: **Robustness to reverberation.** We study the effect of reverberation on our pretext model. Chance performance is 1.5%.

Audio prediction network. We study how audio prediction architectures will influence representation learning from our proposed pretext task. We adapt the U-Net architecture with cross-attention modules for conditional feature inputs [77, 85] and compare the pretext and downstream performance with U-Net [28] we used for our main experiments. We train our models on the HM3D-SS dataset with a single sound source presented in the scenes and use LibriSpeech signals [67]. We report results in Tab. 9. Interestingly, we found that ATTN U-Net can reconstruct better sounds for the pretext task while it does not learn the features as well as the 2.5D U-Net [28]. We hypothesize that a more complex network may transfer the representation learning inside of the prediction networks rather than the feature extractors.

Model	Pretext ↓	Downstream Acc (%) ↑	
	STFT Dist.	AudLoc.	CamRot.
ATTN U-Net [77, 85]	**0.128**	68.0	75.3
2.5D U-Net [28]	0.130	**74.5**	**80.0**

Table 9: **Audio prediction model ablation study.** We evaluate both pretext and downstream performance on the HM3D-SS with LibriSpeech samples [67].

A.4. Implementation details

SLfM model. We use separate multi-layer perceptrons g_v and g_a (*i.e.*, FC ($512 \rightarrow 256$)–ReLU–FC ($256 \rightarrow 1$) layers) to predict scalar rotation and sound angles.

Hyperparameters. For all experiments, we re-sample the audio to 16kHz and use 2.55s audio for the binauralization task. For pretext training, we use the AdamW optimizer [44, 51] with a learning rate of 10^{-4}, a cosine decay learning rate scheduler, a batch size of 96, and early stopping. During downstream tasks, we change the learning rate to 10^{-3} for linear probing experiments. To train our self-supervised pose estimation model, we set the weights λ_1, λ_2, and λ_3 to be $5, 1, 1$ for geometric, binaural, and symmetric losses respectively. For more complex scenarios (Sec. 4.5), we set the weights λ_1, λ_2, and λ_3 to be $3, 1, 1$ to avoid the geometric loss from dominating.

IID cues. We describe our implementation of predicting sound on the left or right using IID cues in detail here: we first compute the magnitude spectrogram $|\mathbf{A}|$ from the binaural waveform $\mathbf{a}$ and sum the magnitude over the frequency axis. Next, we calculate the log ratio between the left and right channels for each time frame. After this, we take the sign of log ratios and convert them into either +1 or -1. We sum over the votes and take the sign of it for final outputs.

Dataset. Due to the fact that SoundSpaces 2.0 [15] does not support material configuration at the current time, we obtain binaural RIRs with different reverberation levels by scaling the indirect RIRs and add them up with direct RIRs. We render binaural sounds with random audio samples as augmentation during training.