Crafting Precision in the Wild: Learning Spatial Sound Localization from Unlabelled Egocentric Videos

Anonymous submission

Abstract

Human beings integrate visual, auditory cues, and general knowledge to discern object localization in their environment. However, accurately linking audio signals to corresponding object bounding boxes is a labor-intensive challenge, especially in instances where pinpointing sound sources that are not visible in videos. This challenge is further amplified in the field of computer vision due to the reliance on synthetic, lab-created, or domain-specific datasets, coupled with the scarcity of real-world data and robust methodologies. A significant observation is the correlation between the relative rotation of cameras and the positions of sound sources in egocentric videos. To address these issues, we introduce a novel methodology leveraging approximate camera rotation angles and sound variations in egocentric videos for enhanced sound direction localization. The effectiveness of our methodology is demonstrated through experiments on a unique, in-the-wild dataset sourced from YouTube walking tours, exhibiting its superior performance over several established baseline methods.

Introduction

Our ability to perceive stereo sound enables us to localize objects beyond our line of sight, a crucial skill for applications like robotics and autonomous driving, particularly in challenging conditions such as poor lighting or occluded environments. Stereo audio, leveraging arrival time and sound level differences between spatially separated microphones, provides valuable supplemental information for understanding the movement of surrounding objects.

Despite the abundance of high-quality stereo sound recordings, particularly in consumer phone videos, current methods struggle with accurate sound source localization within them. Challenges arise when dealing with correlated noise or multiple sound sources. Traditional approaches employ hand-crafted features or supervised learning, but acquiring natural labeled data is difficult, leading to reliance on synthetic training data that may not fully represent real-world complexities.

To address the limitations of existing methods and enhance sound localization in diverse and complex real-world scenarios, we propose a novel approach. By using coarse estimates of sound source motion angles as supervision, we train a neural network to precisely determine sound source angles. Given the difficulty of annotating motion directions in wild data, we substitute camera rotation angles as a proxy. Through simulation experiments, we establish that camera rotation can roughly correspond to sound source positions.

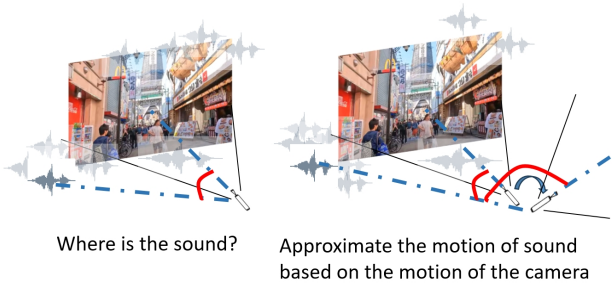

Figure 1: learning sound localization according to the motion of the camera

We curate a dataset comprising city walk videos with stereo sound, identifying segments with significant camera rotation to serve as weak supervision. Training the network on these segments allows it to infer precise sound angles. Experimental results on both simulated and our newly introduced dataset demonstrate superior performance over baseline methods.

In this work, we present a novel approach to the challenging task of localizing objects in unconstrained environments. Our contributions can be summarized as follows:

- To the best of our knowledge, our work represents the first attempt to address the issue of localizing objects in the wild.
- We introduce a method that leverages the approximate motion angles of sound sources to train a neural network, yielding precise localization of sound sources. This approach demonstrates robust performance on both synthetic data and real-world datasets.
- We have curated a dedicated training set and benchmark for sound localization in unconstrained environments, utilizing stereo audio from egocentric YouTube videos. The dataset includes variations in camera rotation, encompassing scenarios with both relatively static and dynamic camera movements. We anticipate that this re-

source will significantly contribute to advancing research in the realm of cross-modal perception.

Related Work

Audio for spatial perception. Recent studies have delved into leveraging sound for comprehending spatial dimensions. In simulated environments, floor plans were reconstructed by (Purushwalkam et al. 2021) utilizing sound, as detailed in (Chen et al. 2020a). Further, (Chen, Hu, and Owens 2021) utilized environmental ambient sounds to gain insights into scene structures. While (Konno et al. 2022) combined sound localization with visual Structure from Motion (SfM), they did not learn these elements in conjunction. Diverse approaches have been employed to learn representations for spatial audio-visual tasks, such as Yang (Yang, Russell, and Salamon 2020) determining the swapping of stereo channels in videos, and Morgado (Morgado, Li, and Nvasconcelos 2020) addressing spatial alignment challenges. These representations have been instrumental in enhancing localization, up-mixing, and segmentation models. In a different approach, our study focuses on learning camera pose and sound localization using only self-supervision, leading to angular predictions without relying on labeled data. Other researchers have utilized echolocation sounds for representation learning (Gao et al. 2020; Yang et al. 2022b), depth map prediction (Christensen, Hornauer, and Yu 2020; Parida, Srivastava, and Sharma 2021), and camera pose estimation (Yang et al. 2022b) using labeled data. However, our method distinctively co-learns binaural sound localization and camera pose via passive audio sensing, independently of supervised learning.

Acoustic synthesis and spatialization. A variety of studies have delved into the realm of sound synthesis guided by visual cues, as documented in several works (Gan et al. 2020; Ghose and Prevost 2020; Iashin and Rahtu 2021; Du et al. 2023), and audio synthesis directed by textual information (Kreuk et al. 2022; Yang et al. 2022a; Huang et al. 2023). Furthermore, the creation of authentic environmental sounds using visual data has been a subject of research (Chen et al. 2022a; Singh et al. 2021; Chen et al. 2022b; Majumder et al. 2022). The concept of synthesizing binaural sound for a new view, using audio-visual data from an original viewpoint, was introduced in (Chen et al. 2023). Other researchers proposed a neural field for audio-visual integration in real-world environments in their research (Liang et al. 2023). Recent advancements have been made in converting mono audio into spatial audio with the aid of visual indicators (Morgado et al. 2018; Gao and Grauman 2019; Rachavarapu et al. 2021; Xu et al. 2021; Lin and Wang 2021; Zhou et al. 2020; Garg, Gao, and Grauman 2021), as well as through understanding the relative positioning of sound sources and receivers (Richard et al. 2021; Huang et al. 2022). Building on these foundations, our method focuses on learning spatial representations by predicting audio features.

Binaural sound localization. Humans possess the ability to pinpoint the origin of sounds using binaural hearing, a skill highlighted in Rayleigh's research (R.S. 1907). Traditional methods for estimating the location of sound involve calculating interaural time differences using cross-correlation and hand-crafted features (Knapp and Carter 1976), employing factorization techniques (Schmidt 1986), or examining differences in loudness between ears (Rayleigh 1907; Wang and Brown 2006). Researchers adapted self-supervised visual tracking techniques for binaural sound localization in their study (Chen, Fouhey, and Owens 2022). In our approach, we also use self-supervision for direction estimation, but we derive this supervision from cross-modal vision cues instead of relying on correspondence cues. Additionally, our method includes learning to estimate camera rotation visually. In a different approach, (Francl 2022) developed a method to learn sound location representations using contrastive loss, with examples differentiated by the degree of head movements. Other researchers have applied supervised learning with annotated data for sound localization in echoic settings (Adavanne, Politis, and Virtanen 2018; Vecchiotti et al. 2019; Yalta, Nakadai, and Ogata 2017). Contrary to these techniques, our model acquires 3D sound localization skills without the need for labeled data.

Camera pose estimation. In our study, we explore camera pose estimation by focusing on image correspondences and addressing an optimization challenge, underpinned by the fundamentals of multi-view geometry (Hartley and Zisserman 2004). We utilize established techniques like structure from motion for refined pose estimation (Schonberger and Frahm 2016) and methods for camera rotation analysis (Brown and Lowe 2007). Our approach deviates from traditional practices by employing neural networks for direct prediction of camera poses, with either standard photographs (Qian, Jin, and Fouhey 2020; Kendall, Grimes, and Cipolla 2015; Melekhov et al. 2017; Jin et al. 2022; Ma et al. 2022) or RGB-D scans as the initial input (Yang, Yan, and Huang 2020; Yang et al. 2019; El Banani, Gao, and Johnson 2021; El Banani and Johnson 2021). We align our research with current studies that emphasize the learning of relative camera poses from constrained perspectives (Jin et al. 2021; Cai et al. 2021), using network architectures that are broadly similar. Unique to our method, we incorporate cross-modal supervision using audio data to determine camera poses, diverging from the usual label-reliant methods. This approach is akin to self-supervised techniques for inferring structure from motion (Zhou et al. 2017; Zou et al. 2020), involving parallel development of models for depth perception and camera pose estimation, adhering to photoconsistency principles. Setting our work apart, we train our model solely with camera rotation supervision .

Audio-visual learning. A variety of research has been directed towards learning integrated representations of audio and visual information, focusing on how they semantically correspond and synchronize over time (Owens and Efros 2018; Xiao et al. 2020; Asano et al. 2020; Morgado, Vasconcelos, and Misra 2021; Owens et al. 2016; Afouras et al. 2022; Mittal et al. 2022). Investigations in this field also extend to areas like pinpointing the origins of sounds in audio-visual contexts (Hu, Chen, and Owens 2022; Mo and Mor-

gado 2022a; Chen et al. 2021; Mo and Morgado 2022b), segregating audio sources (Majumder, Al-Halah, and Grauman 2021; Gao and Grauman 2021; Majumder and Grauman 2022; Tzinis et al. 2022), detection of speaking entities (Afouras et al. 2020; Tao et al. 2021; Alcázar et al. 2021), navigational assistance through sound (Chen et al. 2020a,b; Chen, Al-Halah, and Grauman 2021), and forensic analysis (Zhou and Lim 2021; Haliassos et al. 2022; Feng, Chen, and Owens 2023). Our work diverges by focusing on leveraging audio-visual signals from various angles to explore and learn about geometric aspects.

Method

We employed a basic convolutional neural network that takes multiple audio segments as input and predicts an angle through the network. Subsequently, we compute the angular differences among multiple segments.

There are a total of $n(n-1)/2$ possible combinations of angular differences, and we utilize these differences to supervise the network's training. By comparing the angular differences between two segments, we perform n-class classification to guide our learning process. The purpose of this "comparison" loss function is to predict approximate angular differences.

However, without additional constraints, the solution can become ambiguous and may lead to a meaningless outcome (e.g., predicting zero for all three angles). To prevent this, we introduce an additional loss function known as the "binaural cue loss". This loss is based on interaural intensity differences (IID) and is used to predict whether the sound is to the left or right of the listener, based on which microphone (left or right) records a louder sound.

In the end, our overall loss is a linear combination of the "comparison loss" and the "binaural cue loss," with predefined proportions. This composite loss function ensures that our model can predict angular differences while considering the relative left-right position of the sound relative to the listener, thereby enhancing the model's performance and stability.

Experiments

In this section, we introduce the CityWalk Binaural Dataset, a novel dataset for sound localization, highlighting its distinctive features and origin. To conduct preliminary experiments on our proposed method, we leverage simulated data generated within the SoundSpaces2 (Chen et al. 2022b) framework. Subsequently, we extend the applicability of our method to our CityWalk Dataset and in-the-wild binaural audio scenarios.

Citywalk Dataset

For our study, we curated a new sound localization dataset extracted from binaural sound videos on YouTube, with the majority having an audio sampling rate of 44100Hz. A total of 11,000 videos were collected and segmented into overlapping three-second clips, resulting in a dataset comprising 13,000,000 segments. Due to the inherent noisiness of the raw data, making it challenging to employ sound classification for event detection, we utilized Interaural Level Difference (IID) cues to assist in filtering. Each segment was further divided into five sub-segments, and those with consistent IID cue variations across the sub-segments were retained, resulting in a final set of 500,000 segments.

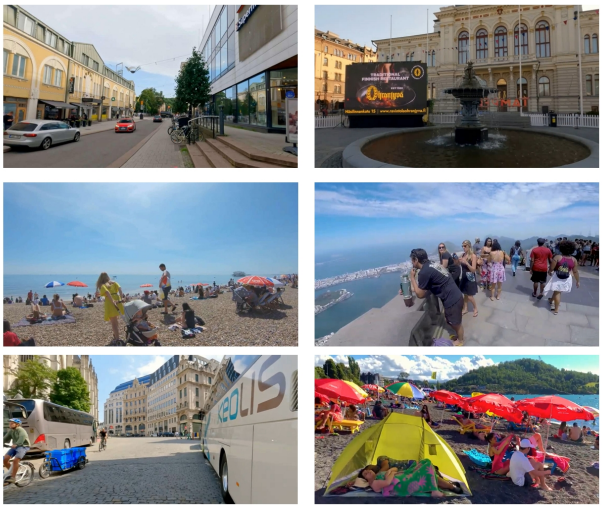

Figure 2: Examples of the scenes in our dataset, most sound sources are cars/water/pedestrians

The segments can be broadly categorized into two types: those where the camera remains static while the source changes position (e.g., moving cars, passing pedestrians), and those where the camera is in motion while the source remains relatively static. The latter case, where the camera is moving, provides weak labels, as estimating camera pose, especially rotation, offers a preliminary understanding of sound changes.

To prioritize segments where camera rotation has a more substantial impact on sound collection compared to the movement of the sound source, we extracted frames from each segment at a frame rate of 5 frames per second (fps). Utilizing the perspective field, we computed the horizontal camera view, and subsequently employed the SuperGlue algorithm to calculate the rotation angles around the vertical axis between consecutive frames. Segments with inconsistent rotation directions (below 80% consistency) were excluded. To mitigate the influence of multiple sound events surpassing the impact of camera rotation, segments with a cumulative vertical axis rotation angle less than 30 were filtered out. Additionally, segments showing discrepancies between the motion direction indicated by IID cues and that computed by the camera were excluded.

In the end, 2,500 segments were retained. Among these, the 500 segments with the highest IID variation product were selected as part of the test set. Furthermore, we identified another 500 segments with minimal rotation (below 10 degrees) and a significant IID variation product, representing instances where the camera motion is limited, and sound events are prominent. This subset was manually annotated,

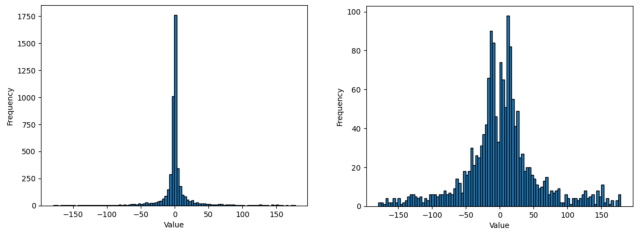

Figure 3: The raw distribution of rotation angles and the modified distribution of rotation angles

categorizing sound angles into eight classes within the [-90, 90] degree range by human annotators.

	Size		Visibility		
	Clips (num)	Duration	Visible	Invisible	Ambiguous
Raw Videos	13,000,000	8000hrs	-	-	-
Training Set	1500	1.25hrs	-	-	-
Validation Set	500	0.42hrs	15%	12%	73%
Test Set	500	0.42hrs	-	-	-

Table 1: Dataset

Experimental Setup

Simulated Dataset We created our dataset, called HM3D-SS, using the SoundSpaces 2.0 platform and 3D scenes from the Habitat-Matterport 3D dataset (HM3D) (Ramakrishnan et al. 2021). This dataset includes realistic 3D environments, photorealistic images, and high-quality spatial audio with natural sound effects like reverberation. It also provides accurate camera positions and sound directions, which are important for evaluation, as we didn't have access to an existing dataset with this information.

We followed the settings of (Chen, Qian, and Owens 2023). To create binaural Room Impulse Responses (RIRs), we used 100 scenes from HM3D. In each audio-visual scenario, we randomly placed sound sources in the scene, with the same height 1.5m. We sampled four different rotated viewpoints, each within four meters of the sound source. These rotations ranged from 10° to 90° relative to the source viewpoints. To maintain consistency, we set the height of agents (objects in the scene) to 1.5 meters and applied a fixed downward tilt angle. We then generated binaural RIRs and images based on the positions of agents and sound sources. The binaural audio was created by combining binaural RIRs with mono audio samples from LibriSpeech and Free Music Archive.

Our dataset comprises 6000 audio pairs. 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.

Baseline On the synthetic dataset, we employed angle supervision as a baseline. However, on the in-the-wild dataset, due to the absence of real angle information, we lacked a clear baseline. Therefore, for the task of predicting whether the sound is on the left or right side, we utilized the Interaural Level Difference (IID) cue as a baseline. This cue helps predict whether the sound is located on the left or right side.

Data Source	**Accuracy(prediction == gt label)**
GT supervised synthetic data	86%
IID supervised synthetic data	87%
IID supervised Youtube data	85%
IID cues	77%

Table 2: 2-classification test of Left/Right Prediction on synthetic/in-the-wild data

Experiment Results

On synthetic data, it can be seen that the Mean Absolute Error (MAE) of angle prediction by our method is very close to that of the supervised method, achieving the anticipated results; this holds true for both classification and regression.

When comparing supervised learning approaches, regression appears to yield better results than classification. Regression also demonstrates increased robustness towards front-back confusion, possibly due to additional preprocessing needed for classification. Experimenting with labels denoting 'more left,' 'more right,' and 'indistinguishable' showcased a slight decrease in performance.

models	**MAE loss for angles**	
	[-180,180]	restrain to [-90,90]
baseline		
regression	3.9	2.0
with [0,1] labels for L/R		
regression	6.8	3.8
32 classification	42	3.8
20 classification	42	4
with [0,1] labels for L/R [-1,0,1] labels for more L/R/cannot-distinguish		
regression	11.5	5.2
32 classification	43.0	5.5

Table 3: Prediction on Synthetic Data

Ablation Study

Trunks of rotation angle We conducted experiments on angular difference classification with two scenarios: six categories and four categories. The numerical results indicate that as the number of categories decreases, from six to four, the precision of Mean Absolute Error (MAE) improves. This suggests that in the task of angular difference classification, reducing the number of categories increases the difficulty of the task.

Losses We initially experimented with using only the left and right movement of sound for supervision. However, this approach failed to converge. It was only when we switched to using approximate angle differences for supervision that we began to see better results. Interestingly, removing the Interaural Level Difference (IID) loss from our model did not significantly affect the outcomes, which suggests that our method does not rely heavily on IID cues for sound localization.

A notable issue in the field of Sound Localization is the 'front-back confusion,' where it is challenging to determine the front or back placement of a sound source based

solely on one angle during inference. Our current solution predicts multiple angles and uses the direction of angle changes—whether clockwise or counterclockwise—to infer the front-back orientation.

In the wild data scenarios, sound source localization becomes even more complex due to the audio being captured by 4K cameras with built-in microphones, as is common with egocentric videos on YouTube shot by various individuals. The internal microphones of these cameras complicate the task of discerning whether a sound is coming from in front of or behind the camera. This aspect represents one of the limitations of our method, indicating an area for future improvement in our approach to sound localization in unstructured environments.

Conclusion

In summary, the ability to detect and localize sound sources with stereo audio significantly enhances object localization capabilities, which is critical for robotics and autonomous vehicles, especially under conditions of poor visibility or obstruction. Despite the availability of high-quality stereo recordings, accurately pinpointing the origin of sounds remains challenging due to issues such as correlated noise and the presence of multiple sources. Traditional methods have been hampered by the scarcity of natural labeled data, leading to a reliance on synthetic data that may not capture real-world complexities adequately.

Our innovative approach aims to overcome these hurdles by employing a neural network trained with approximate motion angles of sound sources, substituting camera rotation angles for direct annotations. We developed a dataset of city walk videos with prominent camera movements to serve as weak labels, allowing the network to learn precise sound source localization. Our method outperforms existing baselines on both synthetic and real-world data.

References

Adavanne, S.; Politis, A.; and Virtanen, T. 2018. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, 1462–1466. IEEE.

Afouras, T.; Asano, Y. M.; Fagan, F.; Vedaldi, A.; and Metze, F. 2022. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10575–10586.

Afouras, T.; Owens, A.; Chung, J. S.; and Zisserman, A. 2020. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*.

Alcázar, J. L.; Caba, F.; Thabet, A. K.; and Ghanem, B. 2021. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 265–274.

Asano, Y.; Patrick, M.; Rupprecht, C.; and Vedaldi, A. 2020. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33: 4660–4671.

Brown, M.; and Lowe, D. G. 2007. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74: 59–73.

Cai, R.; Hariharan, B.; Snavely, N.; and Averbuch-Elor, H. 2021. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14566–14575.

Chen, C.; Al-Halah, Z.; and Grauman, K. 2021. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15516–15525.

Chen, C.; Gao, R.; Calamia, P.; and Grauman, K. 2022a. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18858–18868.

Chen, C.; Jain, U.; Schissler, C.; Gari, S. V. A.; Al-Halah, Z.; Ithapu, V. K.; Robinson, P.; and Grauman, K. 2020a. 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*, 17–36. Springer.

Chen, C.; Majumder, S.; Al-Halah, Z.; Gao, R.; Ramakrishnan, S. K.; and Grauman, K. 2020b. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*.

Chen, C.; Richard, A.; Shapovalov, R.; Ithapu, V. K.; Neverova, N.; Grauman, K.; and Vedaldi, A. 2023. Novel-View Acoustic Synthesis. *arXiv preprint arXiv:2301.08730*.

Chen, C.; Schissler, C.; Garg, S.; Kobernik, P.; Clegg, A.; Calamia, P.; Batra, D.; Robinson, P. W.; and Grauman, K. 2022b. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *arXiv preprint arXiv:2206.08312*.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16867–16876.

Chen, Z.; Fouhey, D. F.; and Owens, A. 2022. 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*, 489–508. Springer.

Chen, Z.; Hu, X.; and Owens, A. 2021. Structure from silence: Learning scene structure from ambient sound. *arXiv preprint arXiv:2111.05846*.

Chen, Z.; Qian, S.; and Owens, A. 2023. Sound Localization from Motion: Jointly Learning Sound Direction and Camera Rotation.

Christensen, J. H.; Hornauer, S.; and Yu, S. 2020. Batvision with gcc-phat features for better sound to vision predictions. *arXiv preprint arXiv:2006.07995*.

Du, Y.; Chen, Z.; Salamon, J.; Russell, B.; and Owens, A. 2023. Conditional Generation of Audio from Video via Foley Analogies. *Computer Vision and Pattern Recognition (CVPR)*.

El Banani, M.; Gao, L.; and Johnson, J. 2021. Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In *CVPR*.

El Banani, M.; and Johnson, J. 2021. Bootstrap your own correspondences. In *ICCV*.

Feng, C.; Chen, Z.; and Owens, A. 2023. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. *arXiv preprint arXiv:2301.01767*.

Francl, A. 2022. *Modeling and Evaluating Human Sound Localization in the Natural Environment*. Ph.D. thesis, Massachusetts Institute of Technology.

Gan, C.; Huang, D.; Chen, P.; Tenenbaum, J. B.; and Torralba, A. 2020. 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*, 758–775. Springer.

Gao, R.; Chen, C.; Al-Halah, Z.; Schissler, C.; and Grauman, K. 2020. 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*, 658–676. Springer.

Gao, R.; and Grauman, K. 2019. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 324–333.

Gao, R.; and Grauman, K. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15490–15500. IEEE.

Garg, R.; Gao, R.; and Grauman, K. 2021. Geometry-Aware Multi-Task Learning for Binaural Audio Generation from Video. *arXiv preprint arXiv:2111.10882*.

Ghose, S.; and Prevost, J. J. 2020. Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning. *IEEE Transactions on Multimedia*, 23: 1895–1907.

Haliassos, A.; Mira, R.; Petridis, S.; and Pantic, M. 2022. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14950–14962.

Hartley, R. I.; and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition.

Hu, X.; Chen, Z.; and Owens, A. 2022. Mix and Localize: Localizing Sound Sources in Mixtures. *Computer Vision and Pattern Recognition (CVPR)*.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. *arXiv preprint arXiv:2301.12661*.

Huang, W. C.; Markovic, D.; Richard, A.; Gebru, I. D.; and Menon, A. 2022. End-to-end binaural speech synthesis. *arXiv preprint arXiv:2207.03697*.

Iashin, V.; and Rahtu, E. 2021. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*.

Jin, L.; Qian, S.; Owens, A.; and Fouhey, D. F. 2021. Planar surface reconstruction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12991–13000.

Jin, L.; Zhang, J.; Hold-Geoffroy, Y.; Wang, O.; Matzen, K.; Sticha, M.; and Fouhey, D. F. 2022. Perspective Fields for Single Image Camera Calibration. *arXiv*.

Kendall, A.; Grimes, M.; and Cipolla, R. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*.

Knapp, C.; and Carter, G. 1976. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4): 320–327.

Konno, T.; Nishida, K.; Itoyama, K.; and Nakadai, K. 2022. Audio-Visual SfM towards 4D reconstruction under dynamic scenes.

Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Liang, S.; Huang, C.; Tian, Y.; Kumar, A.; and Xu, C. 2023. AV-NeRF: Learning Neural Fields for Real-World Audio-Visual Scene Synthesis. *arXiv preprint arXiv:2302.02088*.

Lin, Y.-B.; and Wang, Y.-C. F. 2021. Exploiting audio-visual consistency with partial supervision for spatial audio generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2056–2063.

Ma, W.-C.; Yang, A. J.; Wang, S.; Urtasun, R.; and Torralba, A. 2022. Virtual correspondence: Humans as a cue for extreme-view geometry. In *CVPR*.

Majumder, S.; Al-Halah, Z.; and Grauman, K. 2021. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 275–285.

Majumder, S.; Chen, C.; Al-Halah, Z.; and Grauman, K. 2022. Few-shot audio-visual learning of environment acoustics. *arXiv preprint arXiv:2206.04006*.

Majumder, S.; and Grauman, K. 2022. 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*, 551–569. Springer.

Melekhov, I.; Ylioinas, J.; Kannala, J.; and Rahtu, E. 2017. 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*, 675–687. Springer.

Mittal, H.; Morgado, P.; Jain, U.; and Gupta, A. 2022. Learning state-aware visual representations from audible interactions. *arXiv preprint arXiv:2209.13583*.

Mo, S.; and Morgado, P. 2022a. A Closer Look at Weakly-Supervised Audio-Visual Source Localization. *arXiv preprint arXiv:2209.09634*.

Mo, S.; and Morgado, P. 2022b. Localizing visual sounds the easy way. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, 218–234. Springer.

Morgado, P.; Li, Y.; and Nvasconcelos, N. 2020. Learning representations from audio-visual spatial alignment.

Advances in Neural Information Processing Systems, 33: 4733–4744.

Morgado, P.; Vasconcelos, N.; Langlois, T.; and Wang, O. 2018. Self-supervised generation of spatial audio for 360 video. *arXiv preprint arXiv:1809.02587*.

Morgado, P.; Vasconcelos, N.; and Misra, I. 2021. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12475–12486.

Owens, A.; and Efros, A. A. 2018. Audio-visual Scene Analysis with Self-Supervised Multisensory Features. *European Conference on Computer Vision (ECCV)*.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *Computer Vision and Pattern Recognition (CVPR)*.

Parida, K. K.; Srivastava, S.; and Sharma, G. 2021. Beyond image to depth: Improving depth prediction using echoes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8268–8277.

Purushwalkam, S.; Gari, S. V. A.; Ithapu, V. K.; Schissler, C.; Robinson, P.; Gupta, A.; and Grauman, K. 2021. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1183–1192.

Qian, S.; Jin, L.; and Fouhey, D. F. 2020. Associative3d: Volumetric reconstruction from sparse views. In *ECCV*.

Rachavarapu, K. K.; Sundaresha, V.; Rajagopalan, A.; et al. 2021. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1930–1939.

Ramakrishnan, S. K.; Gokaslan, A.; Wijmans, E.; Maksymets, O.; Clegg, A.; Turner, J.; Undersander, E.; Galuba, W.; Westbury, A.; Chang, A. X.; Savva, M.; Zhao, Y.; and Batra, D. 2021. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. arXiv:2109.08238.

Rayleigh, L. 1907. XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74): 214–232.

Richard, A.; Markovic, D.; Gebru, I. D.; Krenn, S.; Butler, G. A.; Torre, F.; and Sheikh, Y. 2021. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*.

R.S., L. R. O. P. 1907. XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*.

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

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Singh, N.; Mentch, J.; Ng, J.; Beveridge, M.; and Drori, I. 2021. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 286–295.

Tao, R.; Pan, Z.; Das, R. K.; Qian, X.; Shou, M. Z.; and Li, H. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3927–3935.

Tzinis, E.; Wisdom, S.; Remez, T.; and Hershey, J. R. 2022. 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*, 368–385. Springer.

Vecchiotti, P.; Ma, N.; Squartini, S.; and Brown, G. J. 2019. End-to-end binaural sound localisation from the raw waveform. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 451–455. IEEE.

Wang, D.; and Brown, G. J. 2006. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press.

Xiao, F.; Lee, Y. J.; Grauman, K.; Malik, J.; and Feichtenhofer, C. 2020. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*.

Xu, X.; Zhou, H.; Liu, Z.; Dai, B.; Wang, X.; and Lin, D. 2021. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15485–15494.

Yalta, N.; Nakadai, K.; and Ogata, T. 2017. Sound source localization using deep learning models. *Journal of Robotics and Mechatronics*, 29(1): 37–48.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2022a. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*.

Yang, K.; Firman, M.; Brachmann, E.; and Godard, C. 2022b. 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*, 271–291. Springer.

Yang, K.; Russell, B.; and Salamon, J. 2020. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9932–9941.

Yang, Z.; Pan, J. Z.; Luo, L.; Zhou, X.; Grauman, K.; and Huang, Q. 2019. Extreme relative pose estimation for rgb-d scans via scene completion. In *CVPR*.

Yang, Z.; Yan, S.; and Huang, Q. 2020. Extreme relative pose network under hybrid representations. In *CVPR*.

Zhou, H.; Xu, X.; Lin, D.; Wang, X.; and Liu, Z. 2020. Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*.

Zhou, Y.; and Lim, S.-N. 2021. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14800–14809.

Zou, Y.; Ji, P.; Tran, Q.-H.; Huang, J.-B.; and Chandraker, M. 2020. 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*, 710–727. Springer.