MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

SoundLoc3D: Invisible 3D Sound Source Localization and Classification Using a Multimodal RGB-D Acoustic Camera

He, Yuhang; Shin, Sangyun; Cherian, Anoop; Trigoni, Niki; Markham, Andrew

TR2025-003 January 08, 2025

Abstract

Accurately localizing 3D sound sources and estimating their semantic labels – where the sources may not be visible, but are assumed to lie on the physical surface of objects in the scene – have many real applications, including detecting gas leak and machinery malfunction. The audio-visual weak- correlation in such setting poses new challenges in deriving innovative methods to answer if or how we can use cross- modal information to solve the task. Towards this end, we propose to use an acoustic-camera rig consisting of a pinhole RGB-D camera and a coplanar four-channel microphone array (Mic-Array). By using this rig to record audio-visual signals from multiviews, we can use the cross-modal cues to estimate the sound sources 3D locations. Specifically, our framework SoundLoc3D treats the task as a set prediction problem, each element in the set corresponds to a potential sound source. Given the audio-visual weak-correlation, the set representation is initially learned from a single view mi- crophone array signal, and then refined by actively incorporating physical surface cues revealed from multiview RGB-D images. We demonstrate the efficiency and superiority of SoundLoc3D on large-scale simulated dataset, and further show its robustness to RGB-D measurement inaccuracy and ambient noise interference.

IEEE Winter Conference on Applications of Computer Vision (WACV) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

SoundLoc3D: Invisible 3D Sound Source Localization and Classification Using a Multimodal RGB-D Acoustic Camera

Yuhang He[†‡*] Sangyun Shin[†] Anoop Cherian[¶] Niki Trigoni[†] Andrew Markham[†]
[†]Department of Computer Science, University of Oxford, UK [‡]Microsoft Research, Vancouver
[¶]Mitsubishi Electric Research Labs, Cambridge, MA, US

Abstract

Accurately localizing 3D sound sources and estimating their semantic labels – where the sources may not be visible, but are assumed to lie on the physical surface of objects in the scene – have many real applications, including detecting gas leak and machinery malfunction. The audio-visual weak-correlation in such setting poses new challenges in deriving innovative methods to answer if or how we can use cross-modal information to solve the task. Towards this end, we propose to use an acoustic-camera rig consisting of a pinhole RGB-D camera and a coplanar four-channel microphone array (Mic-Array). By using this rig to record audio-visual signals from multiviews, we can use the cross-modal cues to estimate the sound sources 3D locations. Specifically, our framework SoundLoc3D *treats the task as a set prediction problem, each element in the set corresponds to a potential sound source. Given the audio-visual weak-correlation, the set representation is initially learned from a single view microphone array signal, and then refined by actively incorporating physical surface cues revealed from multiview RGB-D images. We demonstrate the efficiency and superiority of* SoundLoc3D *on large-scale simulated dataset, and further show its robustness to RGB-D measurement inaccuracy and ambient noise interference.*

1. Introduction

The task of 3D sound source localization and classification, which aims at localizing the 3D spatial position of each sound source (either physical position or the direction of arrival (DoA)) and inferring its semantic label (*e.g.*, *a telephone ring*), has numerous applications in a variety of real-world scenarios, including robotics [1,16], audio surveillance [48,61], smart homes [6,58], and augmented/virtual reality (AR/VR) [27,57]. Existing methods for 3D sound source localization and classification can be divided into two main types: 1) methods that are vision-agnostic [7,21,24,26] and thus rely solely on acoustic signals, and 2) methods that use the synergy between acoustic and visual modalities by assuming that the sound source is visually discriminative [37,38,47,63,67]. For example, the sound is heard from a musical instrument that is visually observable.

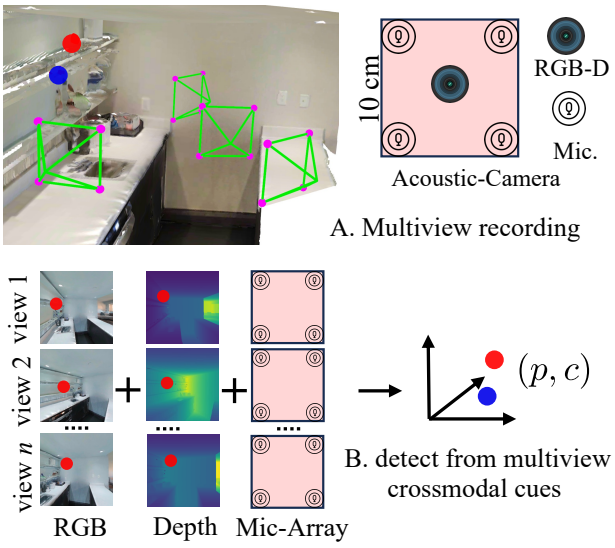


Figure 1. **SoundLoc3D problem setup**: Visually invisible sound sources freely lie on physical object's surface and are emitting sound, A: We use an acoustic-camera to record Mic-Array signal and RGB-D images from multiview. *SoundLoc3D* incorporates multiview crossmodal RGB images, depth maps and Mic-Array signal to jointly localize source position p and semantic label c.

While these two settings have been well-explored by audio/speech processing and audio-visual multimodal research communities, respectively, there is a third important setting that is currently under-explored, but reflects many real-world scenarios: when the sound and vision are weakly correlated; for example, when the source is either too small to be visually observed, blended with background in texture and appearance, or has no associated visual form at all. Typical examples include: gas leak from pipes, water dripping, electrical zapping, abnormal sounds of cooling fans in computers, etc., among others. In many cases, the task of accurately detecting the source of such sounds and inferring their semantic labels is important for damage pre-diagnosis.

*Work done while interning at Mitsubishi Electric Research Labs.

Tackling the above task naturally poses three research questions: 1) is cross-modal information useful in a sound-vision weak-correlation setting? and if so 2) what sort of visual cues can be used?, and 3) how to effectively use cross-modal cues within the weak-correlation setting? In the pioneering work of Sound3DVDet [25], the focus is primarily in using cross-modal multiview RGB images towards addressing Q1 and Q2. In this paper, we go further to explore – in addition to multiview RGB images – whether the depth map of the scene can be used to further improve the performance. The motivation for incorporating the depth maps is two-fold. First, alongside the RGB images, the depth maps can be easily collected (up to an accuracy rate) using either direct depth sensors or stereo matching [66]. Second, the depth map provides more direct cue of the object's physical surface than RGB images. The integration of RGB images and depth maps thus could benefit the task by leveraging the well-explored vision-based multiview geometry methods (*e.g.*, S*f*M [46] and feature matching [31, 50]). Thus, the core questions we seek to answer are: 1) *how to* incorporate multiview RGB-D images to solve the 3D sound source localization problem within the audio-visual weak-correlation setting? and 2) *how to* design a framework that is robust to RGB-D measurement inaccuracy and ambient noise interference? Fig 1 shows the problem setup.

To solve this task, we propose *SoundLoc3D* – an effective, unified and scalable framework for visually invisible 3D sound source localization and classification. To ensure *SoundLoc3D* is scalable to handle arbitrary 3D sound sources, we follow [25] to treat this task as a *set prediction* problem [9, 60], each element in the set is a *query* and associated with a potential sound source. Given the audio-visual weak-correlation, *SoundLoc3D* learns an initial set representation from each of the single-view Mic-Array signal and subsequently optimizes the set representation by actively incorporating sound source cues revealed by multi-view RGB-D images and crossview estimation consistency revealed from multiview observations. Specifically, using the cross-modal RGB-D images, we constrain the sound source to lie on object's physical surface by encouraging: 1) **visual appearance consistency** from multiview RGB images in a feature space, and 2) **spatial proximity** of the source informed by multiview depth maps. From the cross-view observation perspective, we encourage 3) **cross-view estimation consistency** of 3D sound source.

To evaluate *SoundLoc3D*, we provide experiments on a large-scale simulated multiview RGB-D and Mic-Array dataset following the setup described in Sound3DVDet [25]. Our experimental results show that: 1) incorporating depth maps improves the performance significantly and 2) *SoundLoc3D* demonstrates robustness to ambient noise interference and inaccurate RGB-D measurements, showing its potential for real world applications.

2. Related Work

Sound Source Detection. There are several prior works focusing on 3D sound source detection purely from microphone array signals [1, 7, 8, 22, 24, 26]. They either detect 3D sound source direction of arrival (DoA) [1, 7, 24, 26] or spatial position $[x, y, z]$ [22]. In their setting, they assume the microphone receivers are stationary while the sound source can freely move around. This is different from our setting where we instead assume the microphones are movable and the number of sound sources may vary. The recent work Sound3DVDet [25] is the most similar work to ours, it involves multiview RGB to assist the localization.

Multiview based Object Detection. Many existing works on multiview based object detection share the core concept proposed in DETR [9], where object proposals are learned with Transformer in the 2D image space. Adhering to Transformer architecture, DETR3D [60] expands the domain to 3D with multiview input for learning sparse object queries. They detect by either detecting in polar coordinates [13] or integrating 2D features into 3D domain [33, 34].

Audio-Visual Multimodal Learning. Audio-visual multimodal learning has received lots of attention in recent years [2, 18, 19, 36, 39], audio-visual dereverberation [11], localization and navigation [20, 43, 47, 52], mono-to-binaural audio generation. Similar to ours, crossmodal RGB image and depth are incorporated for tasks such as dereverberation [11] and mono-to-binaural audio generation [41].

Image Feature Matching. Finding correspondence between images has been a fundamental topic in computer vision. The research can be divided into detector-based and detector-free methods. Detector-based methods make use of detector to find key-points [3, 4, 14, 45, 49, 54, 64]. Detector-free-based methods find denser correspondences [30, 32, 44, 50, 53]. We utilize the image matching features to provide sound source localization cue.

3. SoundLoc3D Framework Introduction

3.1. Problem Definition

We assume M sound sources $\mathcal{S} = \{(p_m, c_m)\}_{m=1}^{M}$ arbitrarily lie on some physical objects' surfaces in an enclosed room environment, continuously emitting sound waveforms. The physical objects are commonly seen indoor objects such as chair, wall, and door. Each sound source is associated with a 3D spatial position $p_m \in \mathbb{R}^3$ and a semantic class label $c_m \in C$ ($C = \{c_1, c_2, \cdots, c_k\}$, k is class number). The sound sources are invisible and mutually independent, thus may lie on the surface of any object in the scene, emitting a sound waveform of any sound class. The task is to jointly localize each sound source's 3D spatial position and predict its semantic label.

RGB-D Acoustic-Camera. We base our study on the RGB acoustic camera rig proposed in [25], and further equip it with a depth sensor, thereby advancing it to collect depth

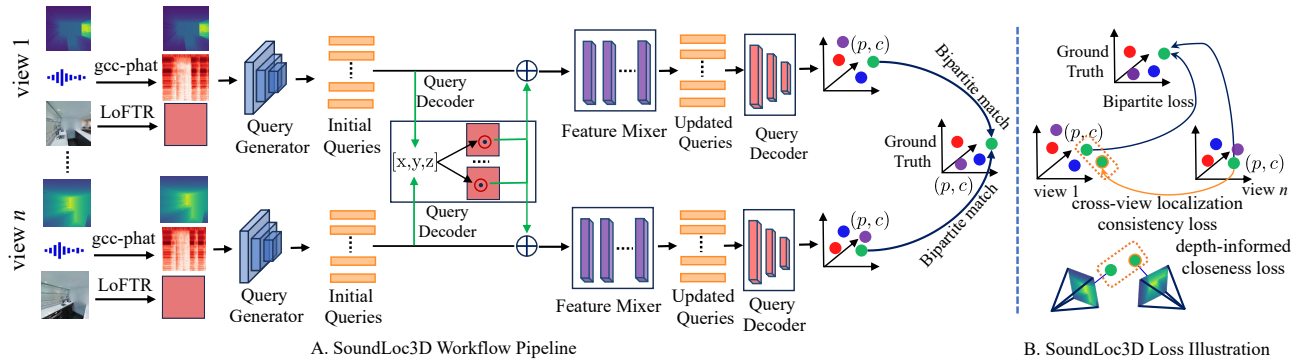


Figure 2. **SoundLoc3D Pipeline**. The RGB image is first pre-processed by a feature matching aware pre-trained model to get an embedding (LoFTR), Mic-Array signal feature is extracted by stacking Log-Mel scale TF and GCC-Phat features. The query generator $\mathcal{G}$ is applied to get the initial queries, which are further fed to query decoder $\mathcal{D}$ to aggregate crossview RGB image informed sound source cues. The queries after aggregation is further optimized by Feature Mixer network $\mathcal{M}$. During training, these queries are matched with ground truth through bipartite matching and the loss considers the discrepancy between prediction and ground truth, depth map informed closeness, and multiview detection consistency. During inference, these optimized queries are simply decoded into sound sources.

maps alongside the RGB image. Specifically, this data acquisition rig consists of a centered pinhole RGB-D camera and four microphones arranged co-planarly at the four corners with 10 cm spacing distance (see Fig. 1 A). The RGB-D camera and the four microphones are pre-calibrated and synchronized so that we are able to use the rig to record the RGB-D image and Mic-Array signal simultaneously from any viewpoint with known camera poses. In this work, we assume that our method has a coarse estimation of the spatial location of the sound sources (*e.g.*, the gas pipes run along the walls in a kitchen and the *leak sound* thus comes from the kitchen wall). We use the RGB-D acoustic-camera to record the acoustic scene [1] from N nearby views, denoted $\{(\mathcal{A}_i, I_i, D_i)|T_i\}_{i=1}^{N}$. For the i-th view, Mic-Array signal is denoted by $\mathcal{A}_i = [a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}]$, RGB image by I_i, depth map by D_i, camera pose by T_i. Our goal is to design a framework $\mathbf{\Omega}$ to jointly localize and classify 3D sound sources from multiview RGB-D and Mic-Array recordings,

$$\mathcal{S} \leftarrow \mathbf{\Omega}(\{(\mathcal{A}_i, I_i, D_i)|T_i\}_{i=1}^{N}). \tag{1}$$

To reflect the real scenario, the framework $\mathbf{\Omega}$ must consider several factors: **c1**, capable of addressing audio-visual weak-correlation, which implies the presence of a 3D sound source is independent on physical object's category; **c2**, accommodate arbitrary number of 3D sound sources locations. and **c3**, robust to multiview RGB-D measurement inaccuracies and ambient noise interference. We show how *SoundLoc3D* is designed to be compliant with all these factors.

Following [25], we formulate *SoundLoc3D* as a *set prediction* problem, where each element in the set corresponds to a potential sound source with unique and stationary spatial position and semantic label. It is worth noting that treating it as *set prediction* problem can be easily scaled up to handle arbitrary 3D sound sources (**c2**). It also avoids us from performing various time-consuming post-processing (*e.g.*, non-maximum suppression (NMS)). Following the terminology in recent works [9, 60], we call each element in the *set* as a sound source *query*. The initial queries in the set are learned from each single view Mic-Array signal independently (**c1**), which are subsequently optimized by actively incorporating cross-modal sound source cues informed by multiview RGB-D images (see Fig. 3). We incorporate three cross-modal sound source cues: 1) visual appearance consistency on the location of the source from multiview RGB images, 2) proximity of the source to an object surface from the multiview depth maps, and 3) cross-view estimation consistency. Specifically, *SoundLoc3D* consists of three main learnable components $\mathbf{\Omega} = (\mathcal{G}, \mathcal{M}, \mathcal{D})$: query generator $\mathcal{G}$ that is responsible for sound source query generation, a feature mixer $\mathcal{M}$ that efficiently integrates multiview cross-modal RGB-D informed sound source cues and a query decoder $\mathcal{D}$ that decodes a query into its spatial position and semantic label. The overall pipeline is shown in Fig. 2.

3.2. Initial Query Learning from each Single-view Mic-Array Signal

The acoustic-camera records four-channel Mic-Array signal from each single view. The four-channel sound waveforms provide enough cues to estimate a sound source's 3D spatial position and semantic label. The time-frequency representation obtained from short time Fourier transform (STFT) reveals the semantic label and inter-channel phase difference encodes its spatial position. Given one Mic-Array signal $\mathcal{A}_i = [a_{i,1}, a_{i,2}, a_{i,3}, a_{i,4}]$, we follow the practice in [1, 7, 21, 22, 59] to jointly encode the time-frequency representation in log-mel scale for each single channel waveform, as well as the generalized cross-correlation phase trans-

[1] An acoustic scene indicates a localized area containing the physical object and associated sound sources to be localized.

form (GCC-Phat [5]) between each pair of channels. The GCC-Phat feature is widely used to encode inter-channel phase difference [1, 7, 8, 55] as it is relatively insensitive to the ambient noise interference [5]. Given two channel sound waveforms $a_{i,k}$ and $a_{i,l}$ in $\mathcal{A}_i$, the GCC-Phat $f^{k,l}_{\text{gccphat},i}$ is,

$$f^{k,l}_{\text{gccphat},i} = \text{ifft}\left(\frac{F(a_{i,k}) \cdot F^*(a_{i,l})}{|F(a_{i,k})| \cdot |F^*(a_{i,l})|}\right), \quad k \neq l, \tag{2}$$

where $\text{ifft}(\cdot)$ indicates inverse short time Fourier transform, $F(\cdot)$ represents short-time Fourier transform (afterwards transformed to Log-mel scale), and F^* indicates the complex conjugate, for $k, l \in \{1, 2, 3, 4\}$. Given the four-channel sound waveforms from a single view, we can extract 10 2D feature maps by stacking 4 STFT representations in Log-mel scale and 6 GCC-Phat feature maps ($\binom{4}{2} = 6$) together, $f_{\text{mic},i} \in \mathbb{R}^{10 \times H_1 \times W_1}$ (in our case, $H_1 = W_1 = 256$).

The source query generator $\boldsymbol{\mathcal{G}}$ then takes as input the 10-channel feature map f_{mic} to learn the vision-agnostic initial queries $\mathcal{Q}_{\text{init}} \in \mathbb{R}^{q \times d}$ (in our case, $q = 16$, $d = 256$). It is achieved by applying a sequence of 2D convolutions to consecutively reduce the feature map spatial resolution while increasing channel dimension size (2D convolution with *stride* of 2 to halve the spatial size),

$$\mathcal{Q}_{\text{init},i} = \boldsymbol{\mathcal{G}}(f_{\text{mic},i}), \forall i = 1, \cdots, N, \tag{3}$$

where $\mathcal{Q}_{\text{init},i}$ indicates the i-th view initial source queries, each of which corresponds to a potential sound source with specific spatial position expressed in its own camera coordinate system (the i-th view camera's coordinate system) and semantic label. We use the query decoder $\boldsymbol{\mathcal{D}}$ to decode each query representation into its spatial position and class label,

$$(P_{\text{init},i}, C_{\text{init},i}) = \boldsymbol{\mathcal{D}}(\mathcal{Q}_{\text{init},i}), \forall i = 1, \cdots, N. \tag{4}$$

Rather than directly predicting sources from the initial queries (Eqn. 4). We further optimize the queries by incorporating sound source cues from multiview RGB-D images.

3.3. Cross-View Source 3D Position Transform

To incorporate multiview cross-modal source cues, we transform the decoded position in Eqn. 4 that is expressed in its own camera coordinate system, to the coordinate system of another camera (*e.g.*, j-th view, $i \neq j$). This is achieved by applying rigid transformation with known extrinsic projection matrix $T_{j \leftarrow i} \in \mathbb{R}^{4 \times 4}$ that translates and rotates from the i-th view $P_{\text{init},i}$ to j-th view coordinate system,

$$P^T_{\text{init},j \leftarrow i} = T_{j \leftarrow i} \cdot P^T_{\text{init},i}, \quad j \neq i, \tag{5}$$

where $P^T_{\text{init},i} \in \mathbb{R}^4$ is a transposed $P_{\text{init},i}$ in homogeneous coordinates with weight 1. After cross-view 3D position transformation, we can project the decoded sound source's 3D position, expressed in the novel view, to its corresponding RGB image and depth map plane to obtain its 2D projection $[u_x, u_y]_{j \leftarrow i}$ under the perspective projection,

$$[u_x, u_y]_{j \leftarrow i} = K_j \cdot P^T_{\text{init},j \leftarrow i} \tag{6}$$

where $K_j \in \mathbb{R}^{3 \times 4}$ is the intrinsic matrix for j-th view. Please note that $K_i = K_j$ as i-th and j-th views are taken from a single camera. With the obtained cross-view 2D projection $[u_x, u_y]_{j \leftarrow i}$, we are able to extract cross-view RGB-D images informed sound source cues.

3.4. Multiview RGB Informed Sound Source Cue

Due to the audio-visual weak-correlation, we cannot directly detect a 3D sound source from a single RGB image. Multiview RGB images, however, can be exploited to implicitly provide sound source position constraints. Specifically, based on multiview geometry [15, 33, 34, 60], an "on-the-surface" 3D point's projections onto multiview RGB images are matching points that are visually similar, while either "above-the-surface" or "below-the-surface" 3D point's projections are non-matching points and thus less visually similar (see Fig. 3 A) [54, 62, 68]. Depending on this constraint, we can encourage the predicted sound source's spatial position to lie on an object's physical surface and further update the query accordingly. After training, the framework tends to predict query producing the source with matching points projects onto multiview RGB images.

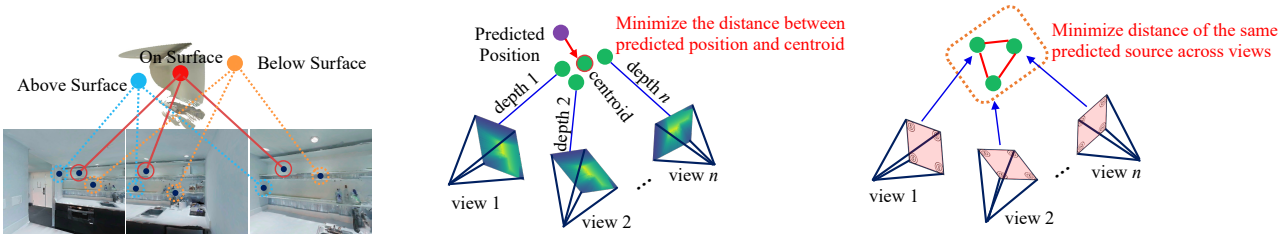
We model crossview visual consistency in feature space. We adopt the pre-trained image matching network LoFTR [50] to directly provide feature embeddings for each sound source projections onto the multiview RGB images. The LoFTR model is specifically trained for image matching in a coarse-to-fine manner. We use only its coarse-level representation (of size $\mathbb{R}^{256 \times 64 \times 64}$). Given a projection pixel point $[u_x, u_y]$, we use bilinear interpolation to obtain its visual appearance embedding $f_{[u_x, u_y]}$,

$$f_{[u_x, u_y]} = \phi_{\text{bilinear}}(\text{LoFTR}(I))_{[u_x, u_y]}, \tag{7}$$

where $\text{LoFTR}(I)$ indicates extracting the LoFTR coarse-level image matching feature representation from the RGB image I. For those invalid projections that are off the image plane, we fill its feature with zeros. Finally, for the initial queries from any single view Mic-Array signal, we aggregate its multiview RGB images-informed sound source cues as:

$$\mathcal{Q}_{\text{init},i} \leftarrow \mathcal{Q}_{\text{init},i} + \frac{1}{N}\sum_{j=0}^{N} f_{[u_x, u_y]_{j \leftarrow i}}; i, j = 1 \cdots, N, \tag{8}$$

where $f_{[u_x, u_y]_{j \leftarrow i}}$ indicates the feature extracted in the j-th view for the query in the i-th view. The updated queries are further fed to $\boldsymbol{\mathcal{M}}$ for further optimization.



A. On-the-Surface Sound Source Visual Consistency B. Depth-informed Source Closeness to Surface C. Localization Consistency across Multiview

Figure 3. **Sound Source Cue from Multiview RGB-D images and Crossview Consistency:** A. While only "on the surface" sound source's projections onto multiview RGB images are guaranteed to be visually similar, either above or below the surface sound sources are much less likely to be visually similar. B. The closer of predicted sound source to the object surface, the smaller of its distance to multiview depth maps informed source position (centroid). C. The same sound source predicted by each single view should be close enough across views.

3.5. Feature Mixer for Query Optimization

The feature mixer $\mathcal{M}$ is a Transformer encoder network [56]. The updated queries in Eqn. 8 are flattened into tokens and passed through $\mathcal{M}$ for further optimization. The motivation for designing $\mathcal{M}$ as a Transformer encoder is two-fold: 1) The $\mathcal{G}$ learned queries are order-less and thus naturally fits for Transformer-based network architecture as all tokens are kept order-less during learning; 2) The updated queries carrying multiview cross-modal sound source cues can easily be further optimized by inter-query interaction and per-query learning:

$$\mathcal{Q}_{\text{update},i} = \mathcal{M}(\mathcal{Q}_{\text{init},i}), \quad i = 1, \cdots, N \tag{9}$$

Given the updated queries $\mathcal{Q}_{\text{update}}$, we can again apply the query decoder $\mathcal{D}$ to decode each individual query into its corresponding spatial position and semantic class label,

$$(P_{\text{update},i}, C_{\text{update},i}) = \mathcal{D}(\mathcal{Q}_{\text{init},i}), \quad i = 1, \cdots, N \tag{10}$$

For each decoded query in Eqn. 10, we apply bipartite matching [28] to associate it with a ground truth sound source. Since the number of predicted queries is usually larger than the ground truth number, we explicitly append `non-source` categories $\varnothing$ to the ground truth so as to match the query number. After bipartite matching, we compute ℓ_1 loss for spatial position regression and cross-entropy loss for semantic label classification.

$$\mathcal{L}_{\text{bm},i} = \frac{1}{N}\sum_{i=1}^{N} |P_{\text{pred},i} - P_{\text{gt},i}| + \text{CE}(C_{\text{pred},i}, C_{\text{gt},i}), \tag{11}$$

where $|\cdot|$ and $\text{CE}(\cdot)$ indicate the ℓ_1 and the cross-entropy loss, respectively. $P_{\text{gt},i}$ and $C_{\text{gt},i}$ are the matched ground truth sound source spatial position and class label, respectively.

3.6. Multiview Depth Informed Sound Source Cue

Comparing with RGB images, depth map provides more straightforward and direct sound source spatial position cues as each depth map directly indicates the spatial locations of the surface of physical objects. Our key insight is that the query decoder $\mathcal{D}$ decoded sound source position's proximity to an object surface can be directly informed by multiview depth maps: projecting the decoded sound source to one view depth map plane to get the projection point, reversing back along the projection ray of the corresponding depth value distance from the projection point naturally gets that depth map informed sound source position. When projecting this query decoded sound source to multiview depth maps, we can accordingly get multiple such depth map informed sound source positions.

The closer the decoded source position of a query is to the physical object's surface, the more spatially aligned its projection points will be on the multiview depth maps after reprojecting them into 3D space. Conversely, if the source position is farther away, the projection points will be more spatially distant. We thus introduce depth map informed spatial closeness loss to encourage the query-decoded sound source to "march towards" physical object's surface (see Fig. 3 B). In this work, we measure the discrepancy between query-decoded spatial position and the centroid of the multiview depth maps re-projected positions. A loss is incurred if the discrepancy exceeds a pre-defined distance threshold σ (in our case $\sigma = 0.3\ m$),

$$\mathcal{L}_{\text{depth},i} = \max\{||P^{+}_{\text{pred},i} - P_{\text{centroid},i}||_2 - \sigma, 0\} \tag{12}$$

where $P_{\text{centroid},i}$ is the centroid of multiview depth-information sound source positions. $P^{+}_{\text{pred},i}$ indicates the decoded queries matched with meaningful ground truth (not the appended `no-source`). $||\cdot||_2$ is the ℓ_2 distance. It is worth noting that the loss is only incurred iff the discrepancy exceeds the distance threshold σ, which reconciles the depth recording inaccuracies. The multiview depth maps thus update the query to lie closer to the physical surface by directly affecting the loss value.

3.7. Crossview Estimation Consistency

As shown in Eqn. 10, we predict the same sound sources from each single view separately. During training stage, each ground truth sound source is matched with one query from each single view and compute the loss (Eqn. 11) between the

ground truth and each matched query separately. This loss fails to take the detection consistency across multiviews into account, it merely takes the difference between ground truth and each single view detection separately. To enforce the crossview estimation consistency, we explicitly incorporate crossview estimation consistency loss $\mathcal{L}_{\text{crossview}}$ to force the the same sound source estimated from multiviews to be as spatially close as possible (see Fig. 3 C),

$$\mathcal{L}_{\text{crossview}} = \frac{1}{C}\sum_{i=1}^{N}\sum_{j=i}^{N}||P^{+}_{\text{pred},i} - T_{i\leftarrow j}\cdot P^{+}_{pred,j}||_2, \quad i \neq j \tag{13}$$

where $P^{+}_{\text{pred},i}$ and $P^{+}_{\text{pred},j}$ indicate the decoded sound source spatial position in Eqn. 10 for the same ground truth sound source from different views. C indicates the view pair combination number, $C = \binom{N}{2}$.

3.8. Training and Inference

The overall pipeline of our proposed *SoundLoc3D* is shown in Fig. 2. Given the collected RGB-D and Mic-Array recordings from multiview, we first extract Log-mel scaled STFT and GCC-Phat feature from four-channel Mic-Array signal (Sec. 3.2) and LoFTR [50] pre-trained image matching model processed RGB image feature embedding (see Sec. 3.4). Afterwards, the learnable query generator $\boldsymbol{\mathcal{G}}$ is applied to learn initial queries representation from each single view Mic-Array signal. Each individual query from one view actively aggregates multiview RGB images informed sound sound cues from all mutliviews by first decoding the query into spatial position and further projecting it to the RGB image plane to collect the corresponding feature. This is achieved by first passing the query feature to the Query Decoder $\boldsymbol{\mathcal{D}}$ to get its decoded spatial position and then projecting the spatial position to the corresponding RGB image plane with relevant camera poses. After merging RGB images informed sound source cues in feature space, the initial queries are fed to feature mixer $\boldsymbol{\mathcal{M}}$ for further optimization. The optimized queries are again fed to Query Decoder $\boldsymbol{\mathcal{D}}$ to be decoded into corresponding spatial position and semantic label. Finally, bipartite matching is applied to match with the corresponding ground truth sound source with one decoded sound source from each view. The overall loss consists of the sum of all of the four aforementioned losses,

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{bm}} + \lambda_2 \cdot \mathcal{L}_{\text{depth}} + \lambda_3 \cdot \mathcal{L}_{\text{crossview}}, \tag{14}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the loss weight and they are all set as 1.0. During training, we adopt the deep supervision strategy [23, 29] to jointly supervise the initial queries in Eqn. 4 and updated queries in Eqn. 10 with the loss expressed in Eqn. 14 (which means P_{pred} is replaced by P_{init} and P_{update} separately). During test, we get the query predictions from each single frame and evaluate against ground truth separately. Finally, we add the evaluation result from each single view together. We do not explicitly merge the predictions from multiviews because predictions from different views can be different (*e.g.*, as it is *set prediction*, the predicted sound source number from different views may vary).

Table 1. Inference time and param. size. Inference time is tested on Intel Core i9-7920X CPU by averaging 100 independent tests.

Method	Inference	Param.	Method	Inference	Param.
SoundDet [26]	1.25 s	13 M	EIN-v2 [7]	2.20 s	26 M
SoundDoA [24]	2.10 s	27 M	SALSA [40]	1.77 s	11.6 M
SALSA-Lite [51]	1.37 s	7.9 M	SELDNet [1]	1.40 s	0.7 M
Sound3DVDet [25]	2.77 s	19.8 M	***SoundLoc3D***	**1.50 s**	**3.8 M**

4. Experiments

Dataset Creation: We follow Sound3DVDet [25] data creation pipeline to create a large-scale synthetic dataset using the SoundSpaces 2.0 [12] and Matterport3D scenes [10]. Specifically, we employ five sound source classes: *telephone-ring*, *siren*, *alarm*, *fireplace* and *horn-beeps* and six physical objects: *wall*, *chair*, *table*, *door*, *ceiling*, and *cabinet*. For any given room scene, we first randomly select a set of those six physical objects. For each object, we independently place n ($1 \leq n \leq 10$) sound sources on its surface (by ensuring any two sources are at least 0.3 m apart), each of which isotropically emits sound waveform. Such an object and the placed multiple sound sources are called an acoustic scene. The collected acoustic scenes show, 1) large visual variability even for the same object like chair; 2) various sound source number; 3) various sound source class. These variabilities force all methods to enhance their generalization capability.

During multiview recording, we put the acoustic-camera approximately $3m$ away from the acoustic scene and ensure all sound sources are not visually blocked in any view. Mic-Array sampling frequency is 21k Hz. To test the visual discriminativeness of the RGB images and their impact on the performance, we divide the scenes into two main categories based on sound source projections' onto the multiview images: 1) texture-homogeneous acoustic scene in which the sound source projections lie on homogeneous textured areas (*e.g.*, textureless wall, table), 2) texture-discriminative acoustic scene where the projections localize on texture discriminative area (*e.g.*, corner, edge). We created 5,000/1250 acoustic scenes for training and test respectively, after filtering views without any depth map. The acoustic scene variabilities discussed above guarantee the training and test sets exhibit enough visual and acoustic difference.

Evaluation Metrics: Following [25], we adopt three metrics: mean average precision (mAP), mean average recall (mAR), mean average localization error (mALE). Given the predicted sound source set and ground truth for a particular class, we apply the bipartite matching algorithm [28] to assign each detected sound source to one ground truth sound source. After the assignment, a detected sound source is a true positive if it is within a distance threshold (we adopt three distance thresholds: $[0.5\ m, 0.8\ m, 1.2\ m]$) with its

Table 2. Quantitative results across all six object categories and five sound classes (left), result on the texture-homogeneous versus texture-discriminative (right). We do not report standard deviation due to space limit (all ≤ 0.010).

Methods	Overall Result			Texture Homogeneous			Texture Discriminative		
	mAP	mAR	mALE	mAP	mAR	mALE	mAP	mAR	mALE
SELDNet [1]	0.103	0.501	0.923	0.107	0.532	0.910	0.100	0.528	0.934
EIN-v2 [7]	0.113	0.607	0.878	0.112	0.620	0.882	0.116	0.600	0.862
SoundDoA [24]	0.212	0.762	0.800	0.225	0.773	0.821	0.224	0.748	0.819
SALSA [40]	0.147	0.722	0.793	0.146	0.722	0.791	0.147	0.723	0.794
SALSA-Lite [51]	0.130	0.712	0.810	0.131	0.710	0.811	0.130	0.713	0.811
SoundDet [26]	0.120	0.674	0.823	0.121	0.675	0.822	0.120	0.674	0.823
Sound3DVDet [25]	0.309	0.998	0.586	0.308	0.997	0.584	0.296	0.992	0.589
SoundLoc3D	**0.518**	**0.999**	**0.320**	**0.517**	**0.997**	**0.312**	**0.519**	**0.997**	**0.301**

Table 3. Ablation study on view number. Standard deviation ≤ 0.005.

View Number	mAP (↑)	mAR (↑)	mALE (↓)
1 view	0.412	0.870	0.520
2 views	0.491	0.923	0.479
4 views	0.516	0.997	0.320
6 views	0.522	0.999	0.318
8 views	0.521	0.999	0.309

Table 4. Ablation study on model architecture variants. Standard deviation ≤ 0.03.

Methods	mAP (↑)	mAR (↑)	mALE (↓)
SL3D_noRGB	0.498	0.944	0.510
SL3D_noDepth	0.472	0.910	0.457
SL3D_noCVC	0.501	0.948	0.389
SL3D_noRGBD	0.328	0.732	0.810
SoundLoc3D	**0.518**	**0.999**	**0.320**

assigned ground truth, otherwise it is treated as a false positive. Afterwards, we compute mAP, mAR and mALE (refer to [25]). Higher mAP and mAR and lower mALE indicate better performance.

Comparison Methods: We compare with 1) six most recent Mic-Array signal based sound source localization and detection baselines: SELDNet [1], EIN-v2 [7] and SoundDoA [24], SoundDet [26], SALSA [40], SALSA-Lite [51]. SELDNet has been used as baseline against various methods, it combines CNN and GRU [17] to infer sound sources; EIN-v2 [7] and SoundDoA [24] are two more recent works; they adopt Transformer [56] and permutation invariant training [65] to infer the location of sound sources. SALSA [40], SALSA-Lite [51] propose to extract log-linear spectrograms and normalized principal eigenvector to represent Mic-Array data. 2) one multimodal method Sound3DVDet [25], which is most relevant to our setting. The comparison on parameter size and inference time is given in Table 1, where we can see *SoundLoc3D* is lightweight and efficient.

Implementation Details: Our framework is implemented in PyTorch [42] and the source code is provided with the supplementary materials. For training the models, we use the AdamW optimizer [35] with a learning rate of 0.0001. Each model is trained for 100 epochs. We train the models three times independently to report the mean and variance for each metric separately. For the Mic-Array signal, the sampling frequency is 21 kHz and we record 1 second data points. The converted time frequency map is of size 256×256 with $\text{n\_fft} = 511$ and $\text{hop\_len} = 78$.

4.1. Experiment Results

The quantitative results are given in Table 2, we can see that *SoundLoc3D* outperforms all the seven comparing methods by a large margin. Comparing with the Mic-Array based best-performing SoundDoA [24], *SoundLoc3D* shows a gain of 0.30 in mAP, 0.23 in mAR and 0.48 in mALE with much smaller network size. Given that most of these methods have larger model sizes (Table 1), the efficacy of *SoundLoc3D* is prominent (even without vision, SL3D_noRGBD in ablation study). *SoundLoc3D* also outperforms Sound3DVDet [25] significantly with much smaller model size.

Further, all methods achieve higher mAR than mAP, suggesting that treating sound source localization and detection as a *set prediction* is capable of estimating all potential sound sources. The overall quantitative performance in terms of texture difference is given in Table 2 right. We observe that *SoundLoc3D* achieves state-of-the-art performance on both texture homogeneous and discriminative scenes, while the other six Mic-Array only methods show no difference because they do not explicitly leverage vision in their methods. *SoundLoc3D* also outperforms Sound3DVDet [25] significantly, showing the potential of depth map in localizing sources. The qualitative comparison is in Fig. 4.

4.2. Ablation Studies

1. Does RGB-D help? 1) We remove RGB based part in our framework (Sec. 3.4 and only feed the initial queries to feature mixer, *SL3D_noRGB*; 2) To test the impact of depth maps, we remove the depth map informed loss (Eqn. 12) and call this version *SL3D_noDepth*; 3) We remove both RGB and depth maps to test cross-modal supervision. From Table 4, we see that removing either RGB images or depth maps results in reduced performance, removing depth maps leads to larger performance drop than RGB images. In *SL3D_noRGBD*, we observe the largest performance drop. It thus shows, in audio-visual weak-correlation, multiview cross-modal visual information can still be use to significantly assist this task.

2. Does Crossview Estimation Consistency help? We remove the crossview consistency loss introduced in Sec. 3.7 (variant *SL3D_noCVC* in Table 4). The dropped performance confirms the importance of crossview consistency.

3. Microphone Number Impact. To understand the impact of microphone array number, we collect another three datasets in which all sound sources lie on "wall" object surface (each dataset includes 800 acoustic scenes for training, and 200 for test). The three datasets are identical except that number of microphones used to record the acoustic scene (we vary the microphone number from 4, 6 to 8). It is worth noting that a minimum of 4 microphones are needed to localize a 3D sound source. In our implementation, all microphones are arranged on a circular plane with a radius 9 cm to the camera center. The results in Table. 5 show that more microphones can improve the performance, but the

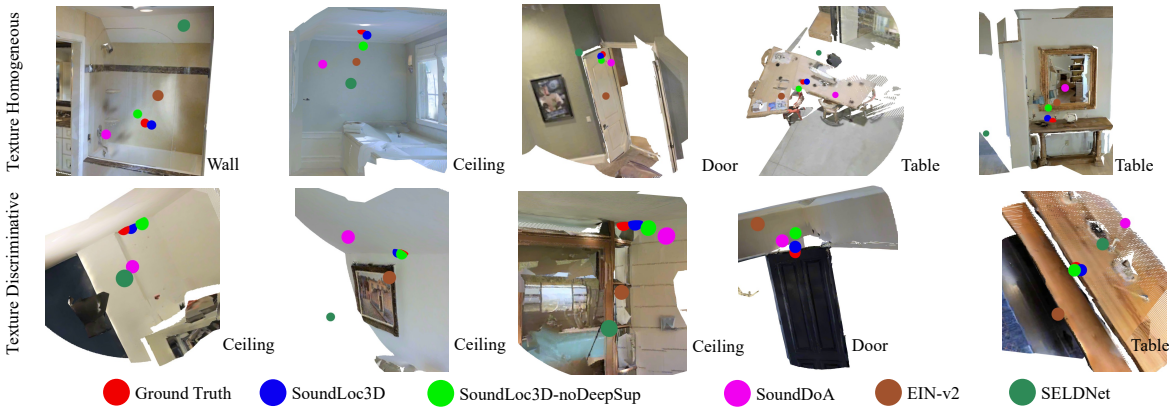


Figure 4. **Localization Result Visualization**: We visualize the sound source localization result in the 3D visual space by different methods as well as its ground truth position. Zoom in for better visualization. We provide data and visualization code and in Supplementary material.

Table 5. Ablation study on microphone number. CNum indicates the GCC-Phat feature channel number. Standard deviation ≤ 0.02.

Method	Mic Num. = 4 (default)			Mic Num. = 6			Mic Num. = 8		
	mAP	mALE	CNum	mAP	mALE	CNum	mAP	mALE	CNum
SoundLoc3D	0.520	0.313	6	0.527	0.308	15	0.530	0.297	28

Table 6. mAP w.r.t. RGB (Sound3DVDet) and RGB-D (SoundLoc3D) measurement inaccuracy.

Method	$\delta = 0$	0.1	0.2	0.3
Sound3DVDet	0.309	0.278	0.243	0.200
SoundLoc3D	**0.516**	**0.507**	**0.498**	**0.480**

Num	Sound3DVDet	Ours
5	0.267	0.497
7	0.255	0.490
9	0.254	0.490

Table 7. mAP (↑) w.r.t. number of source classes. we run the experiment on 400-train, 100-test dataset by increasing sound class number to 7 and 9 (added bird/engine/turbine/fan sound). We can conclude that increasing source class number leads to negligible performance reduction, showing *SoundLoc3D* is capable of handling multiple sound source classes situation.

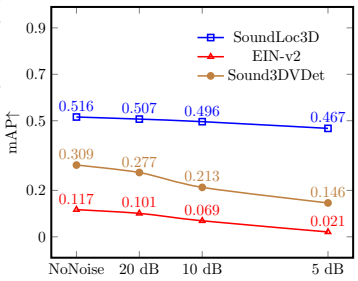


Figure 5. **Ambient noise test**: we add white Gaussian ambient noise measured by SNR in dB.

performance gain comes with extra computations.

4. View Number Impact. Note that the above setup uses four views. To assess the influence of the number of views on performance, we further curated a new dataset where we fixed the number of sound sources and classes but changed the number of views in each acoustic scene. Specifically, we involve three sound sources: telephone, siren and alarm. The sources are placed on the "wall" and the number of views for each acoustic scene varies from 1 to 8. In total, 1,000 acoustic scenes are generated, with 800/100 split for training and test. Five *Sound3DLoc* models are trained on this dataset using views in $\{1, 2, 4, 6, 8\}$. The results are given in Table 3, and it shows two key points: 1) a conspicuous enhancement in performance as the views increase from 1 to 4, and 2) nearly no performance gain as views continue to increase. This indicates that incorporating multiview recordings is beneficial for the task, but the extent of improvement soon diminishes as more views are incorporated.

4.3. Robustness of the Framework

While it is preferable to test our framework on a real-world dataset, collecting such data is both technically and practically challenging. For example, it is difficult to place the sound source on objects' physical surface and it is "visually invisible". To this end, we study the robustness in two settings to best imitate the real-world. First, by including additive white Gaussian ambient acoustic noise to the "wall" data subset, where the amount of acoustic noise is measured by signal-to-noise ratio (SNR, the lower of it, more noise is added). The mAP variations are shown in Fig. 5, where we see both EIN-v2 [7] and Sound3DVDet [25] see a significant drop but *SoundLoc3D* maintains its performance. Second, we add camera pose noise to imitate real-world RGB-D or RGB measurements. Specifically, we add a Gaussian noise $N(0, \delta)$ (mean 0, std δ) to camera rotation parameter (pitch, roll, yaw) to generate RGB-D or RGB measurement inaccuracy. The result in Table 6 shows the robustness of SoundLoc3D. Third, we add more sound classes: from 5 to 9. The results in Table 7 show *SoundLoc3D* can handle detection under more sound source classes. In summary, *SoundLoc3D* is capable of localizing and classifying invisible 3D sound sources from multiview RGB-D Mic-Array recordings. It is efficient, scalable and robust to measurement noise and ambient noise interference that are common in real world, demonstrating its potential to be employed in real world.

Conclusions and Limitations We show multiview RGB-D and Mic-Array recordings can be used to estimate invisible sound sources' spatial position and semantic class. Building and experimenting with a real RGB-D acoustic-camera rig is an important future direction.

References

[1] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen. Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 1, 2, 3, 4, 6, 7

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The Conversation: Deep Audio-Visual Speech Enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 2

[3] Piyush Bagad, Floor Eijkelboom, Mark Fokkema, Danilo de Goede, Paul Hilders, and Miltiadis Kofinas. C-3PO: Towards Rotation Equivariant Feature Detection and Description. In *European Conference on Computer Vision Workshops (ECCVW)*, 2022. 2

[4] Fabio Bellavia. SIFT Matching by Context Exposed. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2

[5] M. S. Brandstein and H. F. Silverman. A Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997. 4

[6] C. Busso, S. Hernanz, Chi-Wei Chu, Soon il Kwon, Sung Lee, P.G. Georgiou, I. Cohen, and S. Narayanan. Smart Room: Participant and Speaker Localization and Identification. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005. 1

[7] Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D Plumbley. An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 1, 2, 3, 4, 6, 7, 8

[8] Yin Cao, Turab Iqbal, Qiuqiang Kong, Yue Zhong, Wenwu Wang, and Mark D Plumbley. Event-Independent Network for Polyphonic Sound Event Localization and Detection. In *DCASE Workshop*, 2020. 2, 4

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*, 2017. 6

[11] Changan Chen and Wei Sun and David Harwath and Kristen Grauman. Learning audio-visual dereverberation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023. 2

[12] 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. In *Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022. 6

[13] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar Parametrization for Vision-Based Surround-View 3D Detection. *arXiv preprint arXiv:2206.10965*, 2022. 2

[14] Ying Chen, Dihe Huang, Shang Xu, Jianlin Liu, and Yong Liu. Guide Local Feature Matching by Overlap Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2

[15] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Graph-DETR3D: rethinking overlapping regions for multi-view 3D object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 4

[16] Lauren M. Chronister, Tessa A. Rhinehart, Aidan Place, and Justin Kitzes. An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information, 2021. 1

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural ntworks on sequence modelling. In *Advances Neural Information Processing System (NeurIPS)*, 2014. 7

[18] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-visual Model for Speech Separation. *arXiv preprint arXiv:1804.03619*, 2018. 2

[19] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through Noise: Visually driven Speaker Separation and Enhancement. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018. 2

[20] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to Separate Object Sounds by Watching Unlabeled Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[21] Francois Grondin, James Glass, Iwona Sobieraj, and Plumbley Mark D. A study of the complexity and accuracy of direction of arrival estimation methods based on gcc-phat for a pair of close microphones. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2019. 1, 3

[22] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 2, 3

[23] Yuhang He, Irving Fang, Yiming Li, Rushi Bhavesh Shah, and Chen Feng. Metric-Free Exploration for Topological Mapping by Task and Motion Imitation in Feature Space. In *Robotics: Science and Systems (RSS)*, 2023. 6

[24] Yuhang He and Andrew Markham. SoundDoA: Learn Sound Source Direction of Arrival and Semantics from Sound Raw Waveforms. In *Interspeech*, 2022. 1, 2, 6, 7

[25] Yuhang He, Sangyun Shin, Anoop Cherian, Niki Trigoni, and Andrew Markham. Sound3DVDet: 3D Sound Source Detection Using Multiview Microphone Array and RGB Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5496–5507, January 2024. 2, 3, 6, 7, 8

[26] Yuhang He, Niki Trigoni, and Andrew Markham. SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform. In *International Conference on Machine Learning (ICML)*, 2021. 1, 2, 6, 7

[27] Yasuhide Hyodo, Chihiro Sugai, Junya Suzuki, Masafumi Takahashi, Masahiko Koizumi, Asako Tomura, Yuki Mitsufuji, and Yota Komoriya. Psychophysiological effect of immersive spatial audio experience enhanced using sound field synthesis. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021. 1

[28] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 5, 6

[29] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015. 6

[30] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution Correspondence Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[31] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense Correspondence across Scenes and its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010. 2

[32] Xingtong Liu, Yiping Zheng, Benjamin Killeen, Masaru Ishii, Gregory D Hager, Russell H Taylor, and Mathias Unberath. Extremely Dense Point Correspondences Using a Learned Feature Descriptor. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[33] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position Embedding Transformation for Multi-View 3D Object Detection. *European Conference on Computer Vision (ECCV)*, 2022. 2, 4

[34] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A Unified Framework for 3D Perception from Multi-Camera Images. *arXiv preprint arXiv:2206.01256*, 2022. 2, 4

[35] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representation (ICLR)*, 2019. 7

[36] Rui Lu, Zhiyao Duan, and Changshui Zhang. Listen and Look: Audio-visual Matching assisted Speech Source Separation. In *IEEE Signal Processing Letters*, 2018. 2

[37] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *International Conference on Machine Learning (ICML)*, 2023. 1

[38] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[39] Giovanni Morrone, Sonia Bergamaschi, Luca Pasa, Luciano Fadiga, Vadim Tikhanoff, and Leonardo Badino. Face Landmark-based Speaker-independent Audio-visual Speech Enhancement in Multi-talker Environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 2

[40] Thi Ngoc Tho Nguyen, Karn N. Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon-Seng Gan. SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022. 6, 7

[41] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond Mono to Binaural: Generating Binaural Audio from Mono Audio with Depth and Cross Modal Attention. In *Proceedings of the IEEE/CVF Conference on Winter Conference on Application of Computer Vision (WACV)*, 2022. 2

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019. 7

[43] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual Object Localization and Separation using Low-rank and Sparsity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 2

[44] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient Neighbourhood Consensus Networks via Submanifold Sparse Convolutions. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning Feature Matching with Graph Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[46] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[47] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to Localize Sound Source in Visual Scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[48] Roneel V Sharan and Tom J Moir. Robust Audio Surveillance Using Spectrogram Image Texture Feature. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015. 1

[49] Xuelun Shen, Qian Hu, Xin Li, and Cheng Wang. A Detector-oblivious Multi-arm Network for Keypoint Matching. *IEEE Transactions on Image Processing*, 2023. 2

[50] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4, 6

[51] Thi Ngoc Tho Nguyen, Douglas L. Jones, Karn N. Watcharasupat, Huy Phan, and Woon-Seng Gan. SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays. In *IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022. 6, 7

[52] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual Event Localization in Unconstrained Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[53] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local Universal Network for Dense Flow and Correspondences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[54] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning Local Features with Policy Gradient. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4

[55] Bert Van Den Broeck, Alexander Bertrand, Peter Karsmakers, Bart Vanrumste, Hugo Van hamme, and Marc Moonen. Time-domain Generalized Cross Correlation Phase Transform Sound Source Localization for Small Microphone Arrays. In *The 5th European DSP in Education and Research Conference*, 2012. 4

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jacob Uszkoreit, Llion Jones Jone, Aidan N. Gomez, and Lukasz Kaiser. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5, 7

[57] Charles Verron, Mitsuko Aramaki, Richard Kronland-Martinet, and Grégory Pallone. A 3-d immersive synthesizer for environmental sounds. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 2010. 1

[58] H. Wang and P. Chu. Voice Source Localization for Automatic Camera Pointing System in Videoconferencing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997. 1

[59] Qing Wang, Huaxin Wu, Zijun Jing, Feng Ma, Yi Fang, Yuxuan Wang, Tairan Chen, Jia Pan, Jun Du, and Chin-hui Lee. The ustc-iflytek system for sound event localization and detection of dcase2020 challenge. In *DCASE workshop*, 2020. 3

[60] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. In *The Conference on Robot Learning*, 2021. 2, 3, 4

[61] Peter W. Wessels, Jeroen v. Sande, and Frits Van der Eerden. Detection and localization of impulsive sound events for environmental noise assessment. *The Journal of the Acoustical Society of America*, 2017. 1

[62] Tao Xie, Kun Dai, Ke Wang, Ruifeng Li, and Lijun Zhao. DeepMatcher: A Deep Transformer-based Network for Robust and Accurate Local Feature Matching. *arXiv preprint arXiv:2301.02993*, 2023. 4

[63] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A Proposal-based Paradigm for Self-supervised Sound Source Localization in Videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[64] Pei Yan, Yihua Tan, Shengzhou Xiong, Yuan Tai, and Yansheng Li. Learning Soft Estimator of Keypoint Scale and Orientation with Probabilistic Covariant Loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[65] Dong Yu, Morten Kolbæ k, Zheng-Hua Tan, and Jesper Jensen. Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 7

[66] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1327–1336, June 2023. 2

[67] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 1

[68] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4