Sound3DVDet: 3D Sound Source Detection using Multiview Microphone Array and RGB Images

Yuhang He[1*†] Sangyun Shin[1†] Anoop Cherian[2] Niki Trigoni[1] Andrew Markham[1]
[1]Department of Computer Science, University of Oxford, Oxford, UK.
[2]Mitsubishi Electric Research Labs, Cambridge, MA, US.

Abstract

Spatial localization of 3D sound sources is an important problem in many real world scenarios, especially when the sources may not have any visually distinguishable characteristic; e.g., finding a gas leak, a malfunctioning motor, etc. In this paper, we cast this task in a novel audio-visual setting, by introducing an acoustic-camera rig consisting of a centered pinhole RGB camera and a uniform circular array of four coplanar microphones. Using this setup, we propose Sound3DVDet *– a 3D sound source localization Transformer model that treats this task as a set prediction problem. It first learns a set of initial sound source locations (dubbed* queries*) from a single view of the microphone array signal, then feeds the query set to a sequence of Transformer-like layers for refinement. Each query arising from each layer repeatedly aggregates sound source cues from other views. We deeply supervise the initial sound source queries, intermediate layer queries, and the final output by measuring their respective discrepancy against ground truth queries via bipartite matching. To evaluate our method, we introduce a new dataset: Sound3DVDet Dataset, consisting of nearly 6k scenes produced using the SoundSpaces simulator. We conduct extensive experiments on our dataset and show the efficacy of our approach against closely related methods, demonstrating significant improvements in the localization accuracy. Code is available at* `https://github.com/yuhanghe01/Sound3DVDet`.

1. Introduction

In this work, we propose to accurately detect 3D sound sources by jointly exploiting multiview audio-visual cross-modal information. We assume sound sources lie on object's physical surface, constantly and repetitively emitting sounds independently, our goal is to pinpoint its 3D position and class label by "looking at and listening to" the joint visual-acoustic scene. Unlike previous works that assume that the sound is strongly correlated with a visual cue/object (*e.g.,* the sound comes from particular objects like a church bell, a train, or a clock) [30, 66, 67], we assume that the sound source is only weakly associated with vision. For example, the sound source is either too small to be visually observable or the sound is coming from a novel object. There are a number of real and challenging application scenarios that meet this setting. For example, industrial gas leakage detection requires a robot to pinpoint a leak that shows no visual difference compared with a normal gas pipe - the only clue is the acoustic emission from the defect. Although we may have a rough estimation of 3D sound source position (*e.g.,* we may know the sound comes from a specific area based on

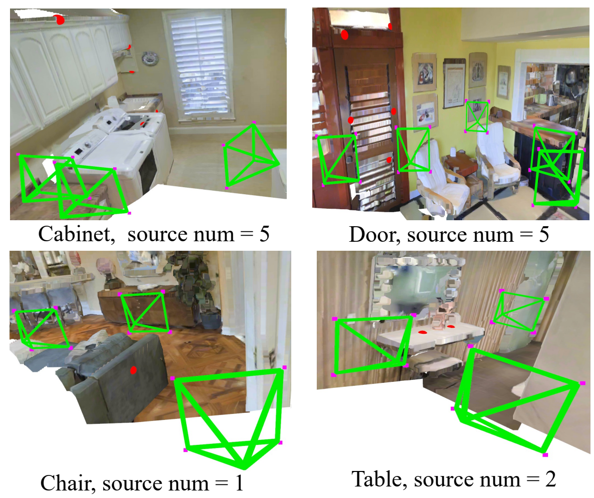

Figure 1. **Sound3DVDet Task Illustration**: Multiple 3D sound sources (red ball) are emitted by visually uninformative objects, we use an acoustic-camera device to record the multi-view, multi-modal visual-acoustic scene. Each recording consists of an RGB image at a known pose (green) and a four-channel microphone array (magenta). The number of sound sources and their classes are arbitrary. The sound sources arbitrarily lie on texture homogeneous (top row) or discriminative regions (bottom row).

*Corresponding author, Email: yuhang.he@cs.ox.ac.uk. Part of the work was done while interning at MERL.
†Equal Contribution

some prior knowledge), how to precisely localize this within a local area remains a challenging task.

In this work, we propose to use an acoustic-camera to record the local area from multiple views. The acoustic-camera is a device equipped with a centered pinhole camera and four microphones in a uniform array. The camera and microphones are coplanar and synchronized so that they record the scene from different viewpoints with known camera poses. At each viewpoint, the RGB image and the multi-channel microphone array signal are recorded simultaneously. The motivations for using multiview audio-visual data are two-fold: first, observing the scene both acoustically and visually from multiple viewpoints enables us to gain a diverse understanding of the sound source; second, multiview RGB images provide useful cues for localizing 3D sound sources. The fundamental idea is to use multiview RGB images to set an "on-the-surface" constraint. A 3D sound source's location when projected onto different RGB image planes are "matching points" when this location lies on the object's surface. Any position shift off the surface (either below or above the surface) leads to the corresponding projections to be "non-matching" points (see Fig. 3).

Based on the multivew acoustic and visual recordings (see Fig. 1 for sample visualization), we propose Sound3DVDet, a novel 3D sound source localization framework that can efficiently handle arbitrary sources. Drawing inspiration from the Transformer architecture design [74] and the current popular set-based object detection methods [10, 44, 76], *Sound3DVDet* treats 3D sound source detection as a set-prediction problem. It directly predicts a set of 3D sound source queries from multiview acoustic-camera recordings, each query corresponds to a potential 3D sound source. To learn discriminative query representations, *Sound3DVDet* first initializes the 3D sound source queries from an individual microphone array sound signal by explicitly using the inter-channel phase difference. Then it refines these queries using a sequence of Transformer layers by improving the cross-modal consistency between acoustic cues and image matching. The final query representations are decoded into 3D sound source positions and class labels through a detection head neural network. During training, the predicted queries are matched with ground truth via bipartite matching [34] and the whole neural network is optimised by minimizing the discrepancy between prediction and ground truth. To further refine 3D sound sources' locations, we deeply supervise [35] the learning of queries arising from all intermediate layers of Transformer, including the initial queries from the microphone array recording (see Fig. 2).

Since there is no publicly available dataset suitable for our task, we use the SoundSpaces 2.0 [12] simulator to create a dataset with 6.2k samples. Experimental results show the our framework outperforms the comparing methods by 20%, 30% and 0.25 in mAP, mAR and mALE metrics, respectively. In summary, we make the three main contributions: **1.** We propose a novel task: 3D sound source detection from a moving acoustic-camera with known camera poses. The acoustic-camera jointly records microphone-array signals and RGB images. The sound source is assumed to lie on an object's physical surface, but may not be visually distinguishable. **2.** We propose **Sound3DVDet**, a novel framework to jointly harness a microphone array and RGB images to accurately detect 3D sound sources. **3.** We introduce a new dataset: *Sound3DVDet dataset*, using which we provide experiments using our model, demonstrating state-of-the-art results on sound source localization and classification.

2. Related Work

Sound Source Detection. There are many works focusing on 3D sound source detection purely from microphone array signals [1, 8, 9, 23, 27, 29]. They either detect 3D sound source direction of arrival (DoA) [1, 8, 27, 29] or spatial physical position $[x, y, z]$ [23, 28]. 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 the static sound sources can vary.

Multiview based Object Detection. Since the seminal work on DETR [10] that learns object proposals in 2D using a Transformer, many works have been proposed that extend the single view used in DETR to multiple views. Extending the core concept of DETR, DETR3D [76] proposed to use Transformer based encoder and decoder for 3D detection with multi-view for learning sparse object queries. Based on DETR3D [76], huge progress has been made in parameterizing 3D detection into polar coordinates [14], focusing on a bottleneck caused by truncated instances with graph structure learning (GSL) [16], incorporating 2D features from the image into 3D domain [44, 45], and using dense queries with at predefined spatial locations for each query [33, 40, 81].

Sound Vision Joint Learning. Exploiting the relationship between audio and visual modalities has gained considerable attention recently in various tasks [85]. Among many tasks, studies that are closely related to ours are audio-visual separation [2, 19, 20, 22, 47, 51, 86], as well as localization and navigation [21, 31, 52, 55, 59, 61, 62, 68, 83, 84]. Most works have made impressive progress in scenarios where audio and vision are tightly correlated (*e.g.*, that is object of interest is always in the camera frustum and it sound is audible and the task is to localize source in 2D space [49, 50, 78]).

Image Feature Matching aims at finding correspondences between images. This line of research could be broadly divided into detector-based and detector-free methods. While detector-based methods first detect salient pixels (keypoints) for comparison [3–5, 15, 24, 36, 60, 63, 72, 75, 79, 80] , detector-free-based methods try to find denser correspondences [32, 39, 43, 58, 65, 69–71].

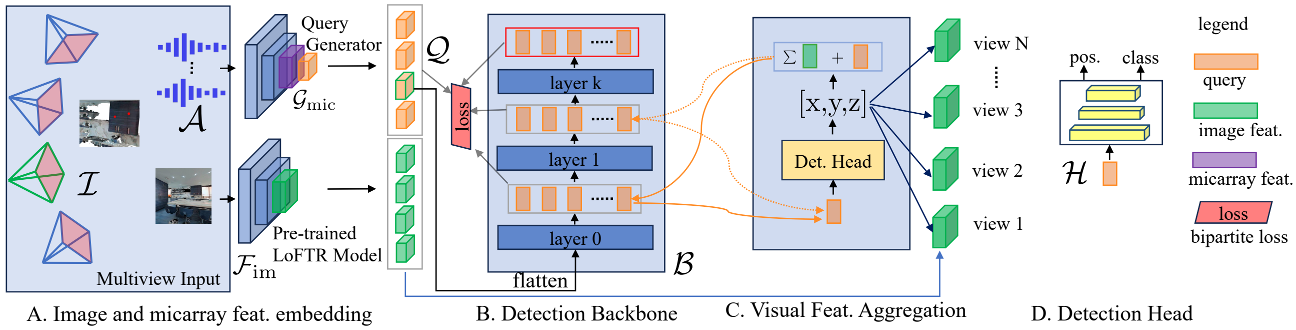

Figure 2. **Sound3DVDet Pipeline Illustration**: A. For each single view, we use a learnable sound source query generator (Sec. 3.3) to jointly obtain the microphone array signal feature embedding and initial sound source queries, pre-trained image model to get RGB image feature embedding (Sec. 3.4), respectively. Then, we randomly choose a reference view (left-most camera in green color) and flatten its initial learned sound source queries. The flattened sound source queries serve as Transformer's tokens and are fed to the detection backbone for further refinement (Sec. 3.5). In each intermediate layer in the backbone, we aggregate multiview visual sound source cues for each query by involving "on-the-surface" constraint. This is achieved by first using detection head $\boldsymbol{\mathcal{H}}$ to decode each query into world positions and then projecting this across the multiple views with the known camera poses (sub-figure B). We deeply supervise all sound source queries during training. During inference, we use the final queries (surrounded by red box) to predict 3D sound sources.

Deeply-Supervised Learning has been extensively explored [26, 35, 37, 38, 64] during the past several years. The main idea is to add extra supervision to various intermediate layers of a deep neural network in order to train deeper neural networks more efficiently. In our work, we adopt a similar idea to deeply supervise the training of feature hallucination and action generation.

3. Multiview based 3D Sound Source Detection

3.1. Problem Formulation

In this paper, we assume a 3D enclosed room environment with an arbitrary number of point sound sources lying on indoor objects' physical surface. These sound sources are constantly and repetitively emitting anechoic sound waveforms. The objects we use are commonly seen indoor objects, including furniture (chair, cabinet, table, *etc.*) and architectural structure (wall, door, ceiling, *etc.*). We also assume we have a rough estimation of the sound sources locations either from prior knowledge or other sound source detection techniques. For example, we may know the sound of gas leak comes from a particular wall in a specific room, because the gas pipes traverse along that wall. Moreover, we assume these sources have no apparent visually distinguishable characteristic, which means that we cannot directly detect them from images alone.

In this paper, we introduce an **acoustic-camera** device to record the local acoustic-visual scene from different viewpoints with known camera poses, each single view recording consists of an RGB image and a microphone array acoustic signal. An acoustic-camera is a device consisting of a pinhole camera and a microphone array that records raw waveforms from each microphone. A microphone array consists of a spatial arrangement of microphones. As sound propagates at roughly 330 m/s at room temperature, the received sound waveforms by any pair of microphones have a time-delay (or phase difference) due to their different distances to the sound source. Using the recorded multi-channel sound waveforms, the sound sources' spatial location and semantic class can be estimated. We use a small array (four microphones with a 10 cm spacing) in this work which is inexpensive and easy to use. This gives an azimuthal far-field angular uncertainty of approximately $10 - 15°$ for frequencies in the range of 500 Hz to 2000 Hz with a sampling frequency of 22050 Hz - see e.g., [6, 13] for more details. Our aim is to use the movement of the acoustic-camera to precisely locate the positions of multiple sound-sources in 3D and their class labels.

Formally, let a multiview acoustic-camera recording be denoted $\boldsymbol{\mathcal{R}}_{\text{av}} = \{(\mathcal{A}_i, I_i, T_i)\}_{i=1}^n$, where $\mathcal{A}_i \in \mathbb{R}^{4\times w}$ is the i-th view of four-channel microphone array sound waveforms $\mathcal{A}_i = [a_{i1}, a_{i2}, a_{i3}, a_{i4}]$, $I_i \in \mathbb{R}^{C\times H\times W}$ is the i-th RGB image (of size $3 \times 512 \times 512$), $T_i \in \mathbb{R}^{3\times 4}$ is the i-th view camera pose (including both the intrinsic and extrinsic parameters), and n is the number of views. Further, let M be the number of static sound sources, expressed as $\mathcal{S} = \{(p_k, c_k)\}_{k=1}^M$, where $p_k \in \mathbb{R}^3$ indicates the 3D position: $p_k = [x_k, y_k, z_k]$ and $c_k \in \mathbb{Z}$ indicates the class label. Our goal is to design a model $\boldsymbol{\Theta}$ to detect 3D sound sources from multiview acoustic-camera recordings, that is:

$$\boldsymbol{\Theta}(\{(\mathcal{A}_i, I_i)|T_i\}_{i=1}^N) \rightarrow \mathcal{S}. \quad (1)$$

3.2. Sound3DVDet Framework Overview

Motivated by [10, 44, 76], we treat 3D sound source detection/localization as a set prediction problem. Given multiview acoustic-recordings $\{(\mathcal{A}_i, I_i)\}_{i=1}^n$, our *Sound3DVDet* model $\boldsymbol{\Theta}$ learns a set of sound source *queries*[1] for a reference

[1]Here and in the subsequent sections, we go by the nomenclature in [10] and call the target variables as *queries*, which correspond to neural

view (*e.g.,* the i-th view) $\mathcal{Q}_i = \{Q_{i1}, Q_{i2}, \cdots, Q_{iK}\}$. Each query $Q_{ik} \in \mathbb{R}^d$ is a potential 3D sound source embedding that can be fed to a detection head network $\boldsymbol{\mathcal{H}}$ to be decoded into its corresponding 3D position and class label. During training, we adopt bipartite matching (a.k.a Hungarian algorithm) [34] to find the best assignment between queries and ground truth sound sources, and optimize the whole neural network Θ with the loss incurred by this bipartite matching. During inference, the predicted sound source queries are directly used to output 3D sound sources; we do not assume to use any post-processing (*e.g.* non-maximum suppression (NMS) for detection redundancy removal [42,56,57]).

Our *Sound3DVDet* fully embraces the sound source cues arising from a single view microphone array signal and multiview RGB images to detect 3D sound sources. From our empirical observations, we find that usually the microphone array signals from a single view can provide coarse estimations to the sound source locations. Leveraging this observation, we propose to learn initial sound source queries from such a single view of the microphone array signals by a query generator network $\boldsymbol{\mathcal{G}}_{\text{mic}}$ (see Sec. 3.3), and subsequently optimize these initial queries through a backbone network $\boldsymbol{\mathcal{B}}$ (see Sec. 3.5). The backbone neural network is a stack of L Transformer encoder layers into which the sound source queries are input as tokens, which then sequentially pass through these L layers. The sound source queries output by a preceding encoder layer is refined by the subsequent encoder layer via: 1) inter-query interaction through Transformer multihead self-attention (MHSA) and feed-forward networks (FFN) and 2) the visual source position cues aggregated from the multiview recordings. Since the same sound source queries are passed through the entire neural network to be refined gradually, we propose to deeply supervise [26,35] the queries arising from the different intermediate layers. We experimentally find such deep supervision enables the neural network to learn better sound source query representations.

In summary, *Sound3DVDet* $\boldsymbol{\Theta}$ consists of a source query generator $\boldsymbol{\mathcal{G}}_{\text{mic}}$, a detection head $\boldsymbol{\mathcal{H}}$, a backbone $\boldsymbol{\mathcal{B}}$ and an RGB image feature extractor $\boldsymbol{\mathcal{F}}_{\text{im}}$, $\boldsymbol{\Theta} = (\boldsymbol{\mathcal{G}}_{\text{mic}}, \boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{F}}_{\text{im}})$. While $\boldsymbol{\mathcal{G}}_{\text{mic}}$, $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{B}}$ are learnable neural networks, $\boldsymbol{\mathcal{F}}_{\text{im}}$ is pre-trained RGB image feature extraction model. Figure 2 shows the pipeline, which works as:

1. At each iteration, *Sound3DVDet* takes as input a multiview acoustic-camera recording $\{(\mathcal{A}_i, I_i)\}_{i=1}^n$. The multiview images I are fed to $\boldsymbol{\mathcal{F}}_{\text{im}}$ to get the image feature maps. The multiview microphone array signals $\mathcal{A}$ are fed to $\boldsymbol{\mathcal{G}}_{\text{mic}}$ to obtain initial sound source queries $\mathcal{Q}_{\text{init}}$.

2. Go through all initial queries, each time select one reference view $\mathcal{Q}_{\text{init},r}$ (*e.g.* the r-th view, $r = 1, \cdots, N$) and pass it to $\boldsymbol{\mathcal{B}}$ for refinement. For each intermediate output in $\boldsymbol{\mathcal{B}}$, we aggregate source cues from multiview RGB images.

3. During training, we deeply supervise all source queries: 1) from query generator $\boldsymbol{\mathcal{G}}_{\text{mic}}$. 2) from intermediate queries in $\boldsymbol{\mathcal{B}}$. 3) from the final output queries in $\boldsymbol{\mathcal{B}}$. During inference, we use the final output queries in $\boldsymbol{\mathcal{B}}$ to predict 3D sound source locations and their labels.

representations of the sound sources.

3.3. Source Queries from Microphone Array Signal

A single-view microphone array signal (four-channel sound waveforms) contains enough information for estimating a 3D sound source's spatial position and class label. Specifically, the class label is encoded in each sound-channel waveform's time-frequency (TF) representation and the spatial position is encoded in the inter-channel phase difference (a.k.a time-delay). Following the common practice [1, 8, 23], for each single-channel one dimensional sound waveform, we first apply the short time Fourier transform (STFT) to transform it into a 2D TF representation and then convert it to log-mel scale. To extract the inter-channel phase difference, we compute the generalized cross-correlation phase transform (GCC-Phat [7], represented as a 2D map) feature between any microphone pair. GCC-Phat has been widely used for microphone array signals [1,8,9,73]. In our case, we create 6 GCC-Phat maps as we compute it for all potential microphone pairs from the four microphones ($\binom{4}{2} = 6$). By concatenating the 6 GCC-Phat maps with the four TF representation maps, we obtain a 10-channel 2D feature map, $F_{\text{mic}} \in \mathbb{R}^{10 \times H_1 \times W_1}$ (in our case, $H_1 = W_1 = 256$).

The source query generator $\boldsymbol{\mathcal{G}}_{\text{mic}}$ takes as input the 10-channel feature map F_{mic}, and applies a sequence of 2D convolutions to consecutively reduce the feature spatial resolution while increasing their channel size. The resolution reduction is achieved by setting the 2D convolution *stride=2*. In our case, we treat the last layer output as the initial source queries $\mathcal{Q}_{\text{init}}$. At the same time, we take the penultimate layer output as the microphone array signal feature embedding $f_{\mathcal{A}_i}$. We will use such microphone array signal embedding in one of our ablation studies to test if further aggregating multiview acoustic signal improves the performance.

$$\mathcal{Q}_{\text{init},i}, f_{\mathcal{A}_i} = \boldsymbol{\mathcal{G}}_{\text{mic}}(F_{\text{mic},i}),\ F_{\text{mic},i} \leftarrow \mathcal{A}_i,\ i = 1, \cdots, N \tag{2}$$

where $\mathcal{Q}_{\text{init},i}$ is the i-th frame sound source queries. During training, we iterate over all views, each time treat the investigating view initial queries as the reference queries $\mathcal{Q}_{\text{init},r}$ and flatten into tokens before feeding to backbone $\boldsymbol{\mathcal{B}}$ for further refinement.

3.4. Visual On-the-Surface Constraint

Since we do not assume the 3D sound source has any obvious visual entity in each single view image, we cannot

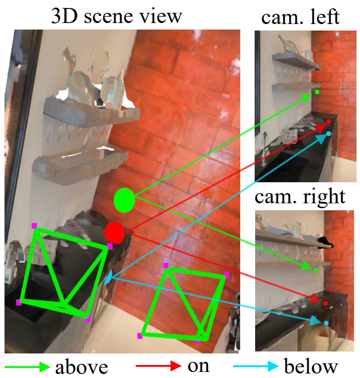

Figure 3. **Visual On-the-Surface Constraint:** While a 3D sound source's projections onto images are visually close matching points if the source lies on the surface (red ball), the projections becomes non-matching points once the source is shifted to either above (green) or below (blue) the surface.

directly detect the sound source in each image. Thus, to make audio-visual learning feasible and use the multiview visual information, we propose to impose an "on-the-surface" constraint on the sound sources – that is, the sound sources are assumed to lie on the physical surface of some object seen in RGB images from all views. Such an assumption allows for an elegant formulation of an audio-visual location consistency. Specifically, if a 3D sound source lies on an object's surface, its projections onto the multiview RGB images are "matching points" [16,44,76]. Any position shift away from the surface (either below of above the surface) makes the projections less likely to be "matching points" (see Fig. 3 for illustration). The task then becomes finding a way that is capable of accurately measuring the "matchness" for projections from multiview RGB images.

Unlike traditional image matching methods [60,72,77,80, 87] that focus on finding corresponding points in discriminative image regions, we proceed in the opposite way to decide the "matchness" for multiview 2D pixels from the projections of the predicted locations of the sources. Furthermore, these 2D pixels can lie in regions that may be textured, discriminative, or homogeneous in the 2D images. Therefore, the resulting RGB image embedding needs to be representative enough in providing matching information across multiple views, regardless of the positions of the matching points. To this end, we depend on the pre-trained feature matching model LoFTR [65] to obtain feature embedding for each RGB image. LoFTR [65] is trained for feature matching in a coarse to fine manner, it is capable of finding matching points even in texture homogeneous regions. Benefiting from this advantage, we are able to reasonably measure the matchness for projections on texture homogeneous area (like walls). We extract its coarse-level representation as the initial embedding (of size $256 \times 64 \times 64$), and further introduce an extra Fully-connected layer (FC) to further adjust the embedding to fit our scene dataset (also increase the feature size from 256 to 512),

$$f_{\mathcal{I}} = \mathrm{FC}(\mathrm{LoFTR}(\mathcal{I})) \tag{3}$$

where $f_I \in \mathbb{R}^{512\times 64\times 64}$ ($I \in \mathcal{I}$). $\boldsymbol{\mathcal{F}}_{\mathrm{im}} = \mathrm{FC}(\mathrm{LoFTR}(\cdot))$. We find that adopting the pretrained model for feature matching gives better performances than using the ImageNet [18] pretrained model (*e.g.*, ResNet50 [25], see Experiment). We provide more discussion on how LoFTR helps set "on-the-surface" constraint in supplementary material.

3.5. Transformer-based Detection Backbone

The initial source queries from the r-th reference view are fed to the detection backbone $\mathcal{B}$ for further learning. The backbone network $\mathcal{B}$ consists of L standard Transformer encoder layers, each of which contains a multi-head self-attention (MHSA) and feed forward network (FFN). The queries, working as Transformer tokens, are optimized in two ways: (i) for a single view source, the multihead attention allows the queries to interact among each other allowing implicitly modeling of the dependency and audio dynamics of sound sources within one view, and (ii) the cross-view consistency, allowing all queries arising from Transformer intermediate layers to actively aggregate source cues from crossmodal multiview RGB images,

$$\mathcal{Q}_{l+1,r} = \boldsymbol{\mathcal{B}}_l(\mathcal{Q}_{l,r}|f_{\mathcal{I}}, \boldsymbol{\mathcal{H}}, T), l = 1, \cdots, L-1 \tag{4}$$

3.6. 3D Sound Source Detection Head

The source detection head $\boldsymbol{\mathcal{H}}$ decodes any query feature (*e.g.*, $q_l \in \mathcal{Q}_l$) into its designated sound source 3D position p and class label c,

$$[p_{i,k}, c_{i,k}] = \boldsymbol{\mathcal{H}}(\mathcal{Q}_{i,k}), \;\; k = 1, \cdots, m \tag{5}$$

where $p_{i,k}$ and $c_{i,k}$ indicates the k-th predicted sound source 3D position expressed in the i-th camera coordinate system, $c_{i,k}$ is the class label. In our implementation, $\boldsymbol{\mathcal{H}}$ consists of two parallel fully-connected layers to regress 3D position and predict class label separately.

3.7. Source Multiview Visual Cue Aggregation

We aggregate multiview RGB images informed sound source cues to improve the sound queries learning. Such aggregation encourages the queries to predict accurate sound source 3D positions because it directly uses the decoded 3D position (via the detection head $\boldsymbol{\mathcal{H}}$ in Eqn. 5) to aggregate source cues. Specifically, given one query $\mathcal{Q}_{l,k}$ arising from the k-th query feature in the l-th detection backbone layer in Eqn. 4, we first apply the detection head $\boldsymbol{\mathcal{H}}$ to decode $\mathcal{Q}_{l,k}$ into its corresponding 3D position $p_{l,k}$ expressed in the reference camera coordinate system (the r-th camera

Input: Multiview data $\{(\mathcal{A}_i, \mathcal{I}_i)|T_i\}_{i=1}^N$, Network $\boldsymbol{\Theta} = (\boldsymbol{\mathcal{G}}_{\text{mic}}, \boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{F}}_{\text{im}})$
$\mathcal{Q}_{\text{init}} \leftarrow \boldsymbol{\mathcal{G}}_{\text{mic}}(\mathcal{A})$, Eqn. 2; $f_{\mathcal{I}} \leftarrow \boldsymbol{\mathcal{F}}_{\text{im}}(\mathcal{I})$, Eqn. 3;
for $r = 1, \cdots, N$ **do**
  for $l = 1, \cdots, L-1$ **do**
    $\mathcal{Q}_{l+1,r} \leftarrow \boldsymbol{\mathcal{B}}_l(\mathcal{Q}_{l,r}|f_{\mathcal{I}}, \boldsymbol{\mathcal{H}}, T)$;
  end
end
Output: $\mathcal{S} \leftarrow \boldsymbol{\mathcal{H}}(\mathcal{Q}_L)$

Algorithm 1: Sound3DV Algorithm Pipeline.

system), and then project it to j-th novel view RGB image plane to get its 2D position $[u_{x,j}, u_{y,j}]$ through the camera poses. Afterwards, we adopt bilinear interpolation ϕ to index the cross-view sound source visual clue $f_{I,r\leftarrow j}$ based on $[u_{x,j}, u_{y,j}]$.

$$f_{I,r\leftarrow j} = \phi_{\text{bilinear}}(f_{I_j})_{[u_{x,j},u_{y,j}]}, \quad j = 1, \cdots, N. \tag{6}$$

If $[u_{x,j}, u_{y,j}]$ is within the j-th RGB plane, we adopt bilinear interpolation in Eqn. 6 to get the feature, otherwise the feature is set 0. Moreover, since the spatial resolution of RGB feature embedding map is much smaller than the original RGB image (RGB image size is 512×512), we follow DETR3D [76] to normalize the valid $[u_{x,j}, u_{y,j}]$ (those lie within the RGB image plane) into $[-1, 1]$ before performing bilinear interpolation. Given all the aggregated multiview RGB image informed source clue features, we merge them into the query through elementwise-add before feeding to next Transformer layer,

$$\mathcal{Q}_{l,k} \leftarrow \mathcal{Q}_{l,k} + \sum_{j=1}^{N} f_{I,r\leftarrow j} \tag{7}$$

Specifically, given one query $\mathcal{Q}_{l,k}$ arising from the k-th query feature in the l-th detection backbone layer in Eqn. 4, we first apply the detection head $\boldsymbol{\mathcal{H}}$ to decode $\mathcal{Q}_{l,k}$ into its corresponding 3D position $p_{l,k,i}$ in the i-th reference camera coordinate system, which is then projected into j-th ($j \neq i$) novel view camera coordinate system $p_{l,k,j} = T_i p_{l,k,i}$. We finally acquire 2D position in image plane $[u_x, v_y]$ by performing perspective projection on $p_{l,k,j}$ with known intrinsic parameters of i-th camera.

3.8. Deeply Supervise All Intermediate Queries

In *Sound3DVDet*, the source queries repetitively appear at different intermediate layers (see Fig. 2). We propose to deeply supervise all intermediate sound source queries learning by directly feeding all of them to detection head $\boldsymbol{\mathcal{H}}$ to predict 3D sound source's position and class label, respectively. We then use bipartite matching [34] loss to supervise all predictions learning. Specifically, we deeply supervise three main sound source queries: the initial queries given by query generator $\mathcal{G}_{\text{mic}}$; intermediate queries from each of the L layers in the backbone network $\boldsymbol{\mathcal{B}}$ and the final queries from the last layer of $\boldsymbol{\mathcal{B}}$.

For bipartite matching, since the number of sound source queries is usually larger than the ground truth sound source number ($M < K$), we explicitly pad `no-source` category $\varnothing$ to the ground truth sound sources to reach the number K. Bipartite matching is then applied to find a one-one correspondence σ^* between prediction and ground truth by taking sound source position closeness and label classification score into account, $\sigma^* = \arg\min_{\sigma\in\mathcal{P}} \sum_{k=1}^{K} -1_{\{C_k\neq\varnothing\}}\hat{p}_{\sigma(k)}(C_k) + 1_{\{C_k=\varnothing\}}\mathcal{L}_{\text{pos}}(P_k, \hat{P}_{\sigma(k)})$, where $\hat{p}_{\sigma(k)}$ and $\hat{P}_{\sigma(k)}$ indicate the predicted label classification probability and 3D position, respectively. $\mathcal{P}$ denotes the permutation set. $\mathcal{L}_{\text{pos}}$ is the L_1 loss for position regression. After finding the best correspondence σ^*, we can then compute the final set prediction loss by combining the classification cross-entropy loss and L_1 position regression loss $\mathcal{L} = \sum_{k=1}^{K} -\log \hat{p}_{\sigma^*(k)}(C_k) + 1_{\{C_k=\varnothing\}}\mathcal{L}_{\text{pos}}(P_k, \hat{P}_{\sigma^*(k)})$. The whole algorithmic visualization is shown in Algorithm 1.

$$\mathcal{L} = \overbrace{\mathcal{L}(\mathcal{Q}_{\text{init}})}^{\text{initial queries}} + \overbrace{\sum_{l=1}^{L-1} \mathcal{L}(\mathcal{Q}_{\boldsymbol{\mathcal{B}}_l})}^{\text{interm. queries}} + \overbrace{\mathcal{L}(\mathcal{Q}_{\boldsymbol{\mathcal{B}}_L})}^{\text{final queries}}. \tag{8}$$

4. Experiments

Dataset Creation: Given the novelty of our problem setup, currently we do not have any publicly available datasets that fit our experimental setup. To this end, we use the SoundSpaces 2.0 [12] simulator to synthesize a new dataset. We load Matterport3D dataset [11] in SoundSpace 2.0. Matterport3D contains large-scale (with average room area $>$100 m^2) and complex indoor room scenes, with which we are able to synthesize data with large visual and acoustic diversity. Specifically, we place multiple point sound sources (source emits sound waveform isotropically) on the surface of 6 commonly seen objects: *wall*, *chair*, *table*, *door*, *ceiling*, *cabinet*. Each sound source emits sound independently. Around the object, we let an agent holding an acoustic-camera to record the object from multiple viewpoints. In our implementation, the multiview acoustic-cameras are recorded roughly at the same height because the agent holds the acoustic-camera at a fixed height position (in our case, at a height of 1.5 m).

Specifically, given an object, we randomly place n ($1 \leq n \leq 10$) sound sources on its surface and ensure any two sources are at least 0.3 m apart (no overlap). Each sound source randomly emits one sound class out of five sound class corpus: *telephone-ring*, *siren*, *alarm*, *fireplace* and *horn-beeps*. The sampling frequency is 21k Hz. By varying the number of sound sources, views and sound classes, we

Table 1. Overall quantitative result across all object categories and sound classes.

Methods	mAP (↑)	mAR (↑)	mALE (↓)
SELDNet [1]	0.101 ± 0.003	0.531 ± 0.000	0.912 ± 0.001
EIN-v2 [8]	0.111 ± 0.003	0.612 ± 0.001	0.877 ± 0.001
SoundDoA [27]	0.123 ± 0.001	0.701 ± 0.001	0.820 ± 0.003
Sound3DVDet	**0.308** ± 0.011	**0.998** ± 0.000	**0.588** ± 0.001

Table 2. Quantitative result comparison between texture homogeneous and texture discrinativative projections of sound sources.

Methods	Texture Homogeneous			Texture Discriminative		
	mAP	mAR	mALE	mAP	mAR	mALE
SELDNet [1]	0.107	0.532	0.910	0.100	0.528	0.934
EIN-v2 [8]	0.115	0.620	0.882	0.117	0.600	0.862
SoundDoA [27]	0.125	0.703	0.821	0.122	0.698	0.820
Sound3DVDet	**0.308**	**0.996**	**0.585**	**0.293**	**0.993**	**0.591**

can flexibly test their individual impact on sound source detection performance. To further test the impact of visual discriminativeness of the RGB image on detection performance, we divide the sound sources into two main categories according to their position in the images: texture-homogeneous area in which the sound source lies around a textured homogeneous area like wall and table surface, texture-discriminative regions in which the sound source lies around regions like corners. More discussion on the creation of the data set is provided in the Supplementary Material. In summary, we have created 5,000/1,250 for train/test, respectively.

Evaluation Metrics: Motivated by existing works on sound event detection [23, 29, 48, 54] and 2D/3D object detection [10, 41, 76], we propose three main evaluation metrics: mean average precision (mAP) and mean average recall (mAR) and mean localization error (ER), to evaluate the performance from various perspectives. It is worth noting that our *Sound3DVDet* directly outputs all sound sources without any post-processing involved.

We first evaluate within each class separately. Given the detected sound source set and ground truth set for a particular class, we first apply bipartite matching algorithm [34] to assign each detected sound source to one ground truth sound source (in some cases, some detections remain unassigned if the detections outnumber the ground truth, and vice versa). After assignment, a detection is a true positive iff it is within a distance threshold with its assigned ground truth, otherwise a false positive. Given a particular threshold, we can accordingly compute the *precision* and *recall*. In our case, rather than fixing one distance threshold, we instead compute across a set of discrete thresholds and further get the average precision (*AP*) and average recall (*AR*) by averaging across all distance thresholds. Finally, we average across all classes to get the mean average precision (*mAP*) and mean average recall (*mAR*). mAP and mAR are two widely adopted evaluation metrics in object detection [10, 29, 41, 76]. In our case, we find that mAP and mAR are relatively dependent on the distance threshold we choose, they do not directly give an understanding how close the predicted sound sources

Table 3. Ablation Study on overall quantitative result across all object categories and sound classes.The top1/top2/top3 performing methods are labelled in red, green and blue color respectively

Methods	mAP (↑)	mAR (↑)	mALE (↓)
S3DVDet_ResNet50	0.236 ± 0.002	0.977 ± 0.006	0.580 ± 0.011
S3DVDet_noDeepS	0.167 ± 0.007	0.994 ± 0.001	0.616 ± 0.004
S3DVDet_noMVSup	0.253 ± 0.018	0.981 ± 0.000	0.603 ± 0.002
SDVDet_mvSound	0.264 ± 0.032	0.994 ± 0.002	0.592 ± 0.008
S3DVDet_wMVIS	0.289 ± 0.006	0.997 ± 0.000	0.595 ± 0.002
Sound3DVDet	0.308 ± 0.011	0.998 ± 0.000	0.588 ± 0.001

Table 4. Ablation Study on quantitative result comparison between texture homogeneous and texture discrinativative projections of sound sources.

Methods	Texture Homogeneous			Texture Discriminative		
	mAP	mAR	mALE	mAP	mAR	mALE
S3DVDet_ResNet50	0.235	0.953	0.583	0.240	0.943	0.579
S3DVDet_noDeepS	0.171	0.988	0.617	0.164	0.977	0.613
S3DVDet_noMVSup	0.254	0.952	0.608	0.168	0.980	0.607
S3DVDet_mvSound	0.274	0.993	0.590	0.253	0.984	0.593
S3DVDet_wMVIS	0.297	0.994	0.593	0.280	0.989	0.597
Sound3DVDet	0.308	0.996	0.585	0.293	0.993	0.591

are to the ground truth. To this end, we further embrace the localization error (LE) metric that are initially used in sound event detection [23, 48, 54]. LE builds on true positive detections, but it goes further to consider the exact the distance between prediction and ground truth. Following mAP and mAR, we first compute average LE across all distance thresholds and finally compute mean average LE (mALE) across all classes. In this work, we adopt three distance thresholds: $[0.5\ m, 0.8\ m, 1.2\ m]$.

Comparison Methods: There are no existing methods that directly work on our proposed problem. We thus propose to compare with three typical microphone array signals based sound source detection baselines: SELDNet [1], EIN-v2 [8] and SoundDoA [27]. SELDNet serves as the baseline for various microphone array based sound source detection, it combines CNN and GRU [17] to detect sound sources; EIN-v2 [8] and SoundDoA [27] are two more recent work, they further adopt Transformer [74] and permutation invariant training [82] to detect sound source.

Implementation Details We implement *Sound3DVDet* with PyTorch [53] and train it on NVIDIA A40. The model parameter size is 19.9 M. We adopt the AdamW optimizer [46], with an initial learning rate 0.0001 and decays every 100 epochs with a decaying rate 0.5. We train each model variants three times independently, and report the mean and variance for each metric separately. We train all models 100 epochs. The source code is given in the supplementary material. We compare with them to test the necessity of involving RGB image and further multiple view recording for 3D sound source detection.

4.1. Experiment Results

Our quantitative results are given in Table 1, from which we can clearly observe that *Sound3DVDet* outperforms all the three comparing methods by a large margin. On aver-

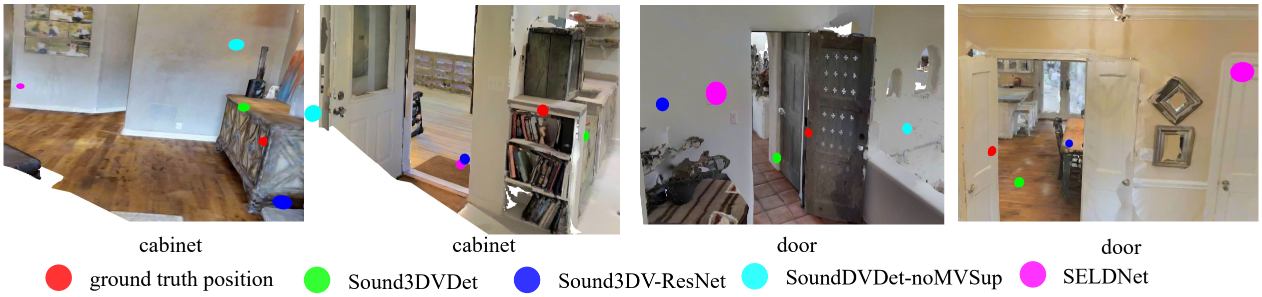

Figure 4. **Qualitative Detection Result Visualization**: We visualize the position of one detected sound source position by different methods as well as its ground truth position. We recommend to zoom in for better visualization.

age, *Sound3DVDet* outperforms the three comparing methods by 20% on mAP, 30% on mAR and 0.25 on mALE. It thus shows our proposed framework works well on 3D sound source detection. We also note that all methods have achieved a much higher mAR than mAP, which means set-based prediction strategy is capable of predicting enough sound sources in each camera view.

The performance in terms of texture difference is shown in Table 2. We can observe from this table that 1) the three comparing methods show inconsistency w.r.t. the texture difference, which is reasonable because they do not explicitly depend on vision information to detect 3D sound sources; 2) *Sound3DVDet* can still achieve reasonably good performance in texture homogeneous area with small performance drop.

4.2. Ablation Studies

We present five ablation studies. The quantitative results are provided in Table 3 and Table 4.

1. Pre-trained Image Matching Feature VS. Classification Feature. As an alternative, we adopt ImageNet [18] pre-trained ResNet50 [25] (S3DVDet_ResNet50) to replace LoFTR [65]. This replacement helps to test what RGB image feature is better for providing "on-the-surface" constraint. From Table 3 and 4, we can see that such replacement leads to obvious performance drop in mAP (≈ 0.6) and mAR (≈ 0.2). It thus shows pre-trained image matching model is better at setting "on-the-surface" constraint, especially in texture homogeneous area. This is also echoed in Table 4, in which we have observed performance drop in detection in texture homogeneous area.

2. Without Deep Supervision When removing the deep supervision from the initial sound source queries and detection backbone intermediate layers (S3DVDet_noDeepS), we have observed significant performance drop (mAP ≈ 1.4, mAR ≈ 0.02, mALE ≈ 0.3), which shows deep supervision strategy is vital to enforce the whole framework to learn more representative sound source queries representation.

3. No Multiview Supervision in which we just rely on single view (microphone array and RGB image) to predict 3D sound sources with cross-view visual feature aggregation (3DVDet_noMVSup). However, the deep supervision module is still kept. We have observed significant overall performance drop. The performance drop becomes significant when the sound sources lie around texture discriminative area. It thus shows multiview supervision is an essential component of *Sound3DVDet*.

4. With Multiview Sound, in which we replace the image feature embedding by the learned microphone array signal embedding (Eqn. 2). It helps to test if it is a better choice to use cross-view image supervision than microphone-array signal. We call this variant S3DVDet_mvSound. From these two tables 3,4, we can clearly see that replacing image with microphone array signal supervision leads to significant performance drop.

5. With Multiview both Image and Sound. In the above test, we show aggregating cross-view acoustic feature leads to inferior performance, but what if we combine image and sound? To this end, we propose a *Sound3DVDet* variant (S3DVDet_wMVIS) that jointly aggregates cross-view image feature and acoustic feature. We have observed performance drop, but the performance drop is not that obvious than other *Sound3DVDet* variants, which in turns shows the importance of involving multiview image feature for 3D sound source prediction.

The ablation studies show the necessity of each component of *Sound3DVDet*. More ablation studies are provided in Supplementary material. We further qualitative visualization in Fig. 4, from which we can see the two *Sound3DVDet* variants and the comparing SELDNet [1] predict sound source incorrectly that either lies in the air or on different object surface. Our proposed framework *Sound3DVDet* can predict the 3D sound source that is closest to the ground truth.

5. Conclusions and Limitations

In this work, we show how to use multiview acoustic-camera recordings to assist localize invisible 3D sound sources. A limitation is that we assume the space between the sources and acoustic-camera is unoccluded, which may not reflect the real settings. Another limitation is that we do not consider situation where the sound sources are moving and dynamic. Using real robotic acoustic-camera is also planned for the future.

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. 2, 4, 7, 8

[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] Axel Barroso-Laguna and Krystian Mikolajczyk. Key. net: Keypoint Detection by Handcrafted and Learned CNN Filters Revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2

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

[6] Oded Bialer, Noa Garnett, and Tom Tirer. Performance advantages of deep neural networks for angle of arrival estimation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3907–3911. IEEE, 2019. 3

[7] 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

[8] 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. 2, 4, 7

[9] 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

[10] 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, 7

[11] 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

[12] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-Visual Navigation in 3D Environments. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[13] Joe C Chen, Kung Yao, and Ralph E Hudson. Acoustic source localization and beamforming: theory and practice. *EURASIP journal on advances in signal processing*, 2003:1–12, 2003. 3

[14] 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

[15] 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

[16] 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. 2, 5

[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] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5, 8

[19] 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

[20] 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

[21] 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

[22] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu. Multi-modal Multi-channel Target Speech Separation. In *IEEE Journal of Selected Topics in Signal Processing*, 2020. 2

[23] 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, 4, 7

[24] Jianfeng He, Yuan Gao, Tianzhu Zhang, Zhe Zhang, and Feng Wu. D2Former: Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-Based Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016. 5, 8

[26] 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. 3, 4

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

[28] Yuhang He and Andrew Markham. SoundSynp: Sound Source Detection from Raw Waveforms with Multi-Scale Synperiodic Filterbanks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. 2

[29] 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. 2, 7

[30] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[31] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torraba. ConceptFusion: Open-set Multimodal 3D Mapping. *Robotics: Science and Systems (RSS)*, 2023. 2

[32] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching across Images. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[33] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3D Object Detection with Polar Transformers. *arXiv preprint arXiv:2206.15398*, 2022. 2

[34] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 2, 4, 6, 7

[35] 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. 2, 3, 4

[36] Jongmin Lee, Byungjin Kim, Seungwook Kim, and Minsu Cho. Learning Rotation-Equivariant Features for Visual Correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[37] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep Supervision with Shape Concepts for Occlusion-Aware 3D Object Parsing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[38] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep Supervision with Intermediate Concepts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 3

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

[40] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BevFormer: Learning Bird's-Eye-View Representation From Multi-Camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, 2014. 7

[42] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot Multibox Detector. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[43] 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

[44] 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, 3, 5

[45] 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

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

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

[48] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for Polyphonic Sound Event Detection. *Applied Sciences*, 2016. 7

[49] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *International Conference on Machine Learning (ICML)*, 2023. 2

[50] 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. 2

[51] 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

[52] Sanjeel Parekh, Alexey Ozerov, Slim Essid, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. Identify, Locate and Separate: Audio-visual Object Extraction in Large Video Collections using Weak Supervision. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019. 2

[53] 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

[54] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020. 7

[55] 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

[56] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical Depth DistributionNetwork for Monocular 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[57] Shaoqing Ren, Kaiming He, Ross Girshich, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 4

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

[59] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised Audio-visual Co-segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 2

[60] 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, 5

[61] 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. 2

[62] Rahul Sharma, Krishna Somandepalli, and Shrikanth Narayanan. Crossmodal Learning for Audio-visual Speech Event Localization. *arXiv preprint arXiv:2003.04358*, 2020. 2

[63] 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

[64] Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao. Deeply-Supervised Knowledge Synergy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[65] 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, 5, 8

[66] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[67] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-Visual Event Localization in Unconstrained Videos. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[68] 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

[69] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. GOCor: Bringing Globally Optimized Correspondence Volumes into Your Neural Network. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[70] 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

[71] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning Accurate Dense Correspondences and When to Trust Them. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[72] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning Local Features with Policy Gradient. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5

[73] 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

[74] 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. 2, 7

[75] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Asian Conference on Computer Vision (ACCV)*, 2022. 2

[76] 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, 5, 6, 7

[77] 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. 5

[78] 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. 2

[79] Fei Xue, Ignas Budvytis, and Roberto Cipolla. SFD2: Semantic-guided Feature Detection and Description. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2

[80] 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, 5

[81] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[82] 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

[83] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The Sound of Motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[84] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The Sound of Pixels. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2

[85] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep Audio-visual Learning: A Survey. *International Journal of Automation and Computing*, 2021. 2

[86] Lingyu Zhu and Esa Rahtu. Separating Sounds from a Single Image. In *ArXiv, Arxiv 2007.07984*, 2020. 2

[87] 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. 5