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, UK
[2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, US

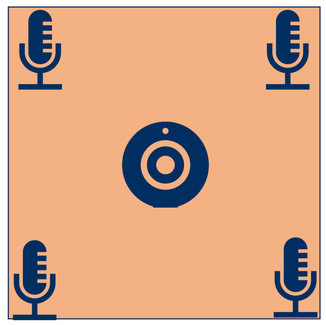
1. Problem Definition

Given a set of 3D sound sources, we aim to
1. localize their spatial [x, y, z] coordinates.
2. classify their semantic label.

where sound sources,
1. arbitrarily lie on the physical surface of objects
2. visually non-observable (too small/no vis-entity).

from multiview Mic-Array and RGB images.

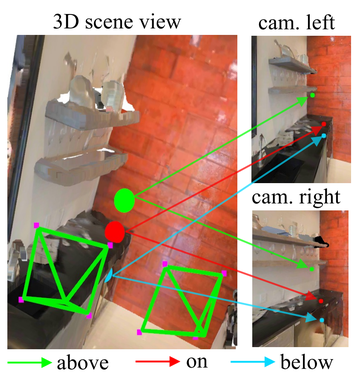
2. Acoustic-Camera Rig



Co-planar Rig, where
1. pinhole RGB camera in the center.
2. four Mics distribute at four corners.

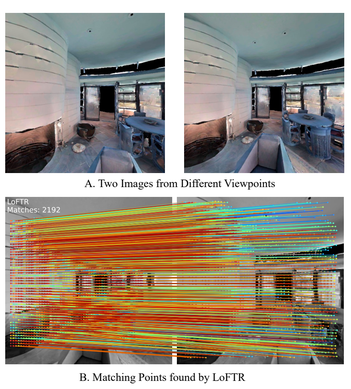
Use the rig to record the sound sources from closeby multiviews with known camera poses.

3. On-the-Surface Constraint



On-the-surface sound sources,
Their projections onto different RGB images are matching points.
Otherwise (either below or above), they are less likely to be matching points (visually dissimilar).

4. On-the-Surface Cues from RGBs



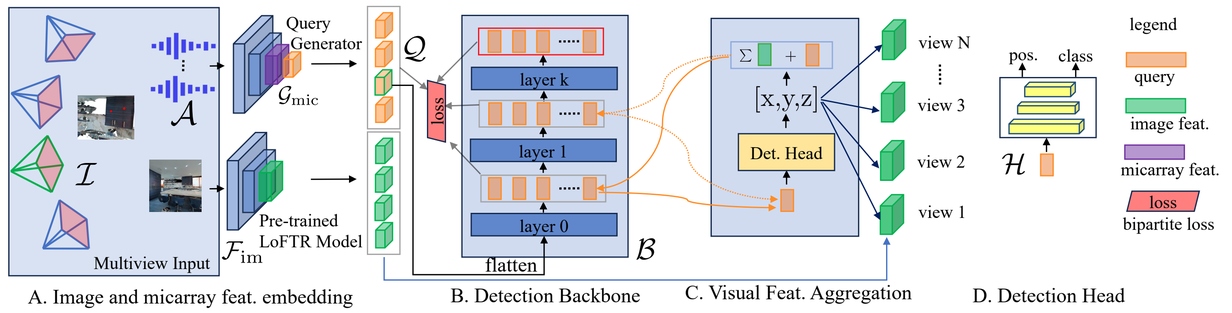
We adopt LoFTR[1] to pre-extract RGB feature. The advantage is that it is capable of generating matching points on texture homogeneous area.

[1] Jiamin Sun et al., LoFTR: Detector-Free Local Feature Matching with Transformers. CVPR 2021.

5. Sound3DVDet Idea Sketch

- Treat it as a *Set Prediction* problem.
- Mic-Array signal gives initial sound sources.
- Initial sound sources are iteratively optimized by aggregating multiview RGBs informed feature.
- Optimized sound sources are matched with ground truth with *Hungarian Algorithm*.

6. Sound3DVDet Pipeline

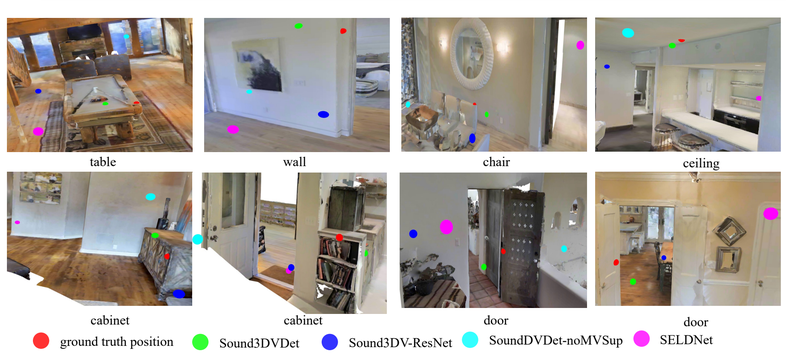


- Four modules: Query Generator; Backbone; Detection Head; Visual Feature Aggregation;

7. Experiment Result

- Simulate with Sound-Spaces 2.0 simulator.
- 6 objects: *wall, chair, table, door, ceiling and cabinet.*
- 5 sources: *telephone-ring, siren, alarm, fireplace, etc*.
- On both texture discriminative and homogeneous area.

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



Conclusion:
1. Novel audio visual research direction.
2. New baseline and evaluation metrics.
3. Hope to motivate more research.