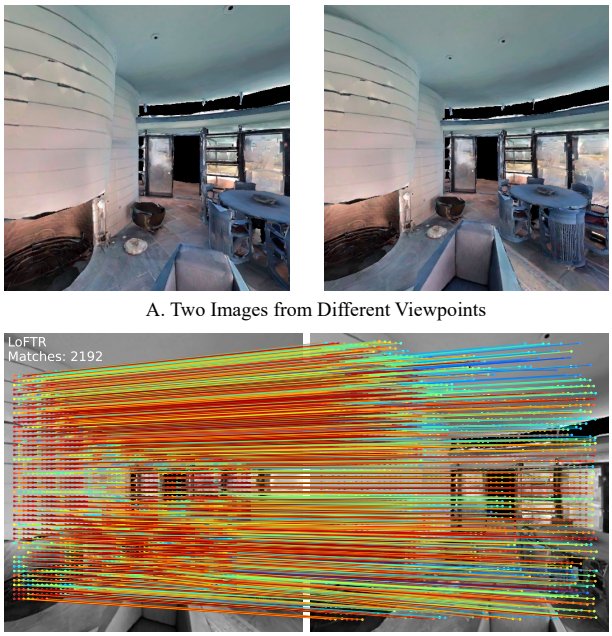
SoundLoc3D Supplementary Materials

1. More Discussion on LoFTR

We adopt LoFTR [9] to extract RGB image feature, which provides RGB informed sound source "on-the-surface" appearance consistency constraint across multiviews. LoFTR [9] is a feature matching model that provides "matching point" across RGB images, it naturally fits for our situation because we depend on such "matching point" to infer 3D sound source's spatial localization. Due to its coarse-to-fine learning strategy, LoFTR is capable of retrieving matching points on both texture homogeneous and texture discriminative area. We show such an example in Fig. 1, from which we can clearly see that dense matching point pairs are generated on the texture homogeneous wall and ceiling area. It thus shows LoFTR [9] can provide useful sound source clues for 3D sound source position, regardless of the position's visual appearance. In Sec. 5.1, we show LoFTR RGB image feature extractor generates better performance than ImageNet pre-trained ResNet50 [4] image feature extractor.



A. Two Images from Different Viewpoints

B. Matching Points found by LoFTR

Figure 1. **LoFTR extracted matching points visualization**. **A.** Two RGB images from different views. They contain large texture homogeneous area, including wall and ceiling. **B.** LoFTR manages to predict dense matching points even on these texture homogeneous areas. We utilize such characteristic to give robust sound source's visual appearance information to constrain the sound source to lie on an object's physical surface.

2. Network Architecture

SoundLoc3D network architecture is given in Table 1. The trainable parameter size of our network is 3.8 M. It is worth noting that our proposed *SoundLoc3D* framework is scalable. Its model complexity can be easily scaled up by adding, for example, more Feature Mixer layers (Transformer encoder layer) or increasing the query embedding size.

3. More Discussion on Dataset Creation

In the supplementary material, we provide the statistics of the created large dataset in Table 2 w.r.t. different physical object class. In this table, we can observe that the "wall" and "ceiling" consist of the largest portion of the dataset, which reflects the real scenario. We further provide some visualizations of the created dataset in Fig. 2, from which we can have an intuitive understanding of how the dataset look like.

We followed the data creation method introduced in Sound3DVDet [6] to create the dataset. We skip the sampled position when no depth map can be collected, so the dataset used by Sound3DVDet [6] and this paper is not exactly the same (the reported Sound3DVDet result in this paper is slightly different from the result reported in the original Sound3DVDet paper). We will release the data creation code and created dataset if this paper is accepted. It is worth noting that, although we placed a 3D sound source on a specific physical object surface in the dataset we have created, the 3D sound sources can freely lie on an arbitrary physical surface. In another word, the sound source placement is independent of physical objects.

Table 1. *SoundLoc3D* network architecture illustration. In the Query Generator $\mathcal{G}$, the 2D convolution kernel size is 3×3 and the stride is 2.

Query Generator $\mathcal{G}$: Input: $[10, 256, 256]$			
Layer Name	In-ch. Num.	Out-ch. Num.	feature size
Conv2D	10	32	$[32, 128, 128]$
Conv2D	32	64	$[64, 64, 64]$
Conv2D	64	128	$[128, 32, 32]$
Conv2D	128	256	$[256, 16, 16]$
Conv2D	256	512	$[512, 8, 8]$
Conv2D	512	256	$[256, 4, 4]$
Query Generator Output: $[16, 256]$			
RGB Informed Feature Aggregation			
LoFT input: $[256, 64, 64]$			
Aggregation output: $[16, 256]$			
Feature Mixer $\mathcal{M}$			
Transformer Layer Num			1
Token Num			16
Head Num			4
FFT Dim			1024
Output			$[16, 256]$
Query Decoder $\mathcal{D}$			
Position Regression Head			
Linear + BN + ReLU	256	128	$[16, 128]$
Linear	128	3	$[16, 3]$
Classification Head			
Linear	256	class num	$[16, \text{classnum}]$

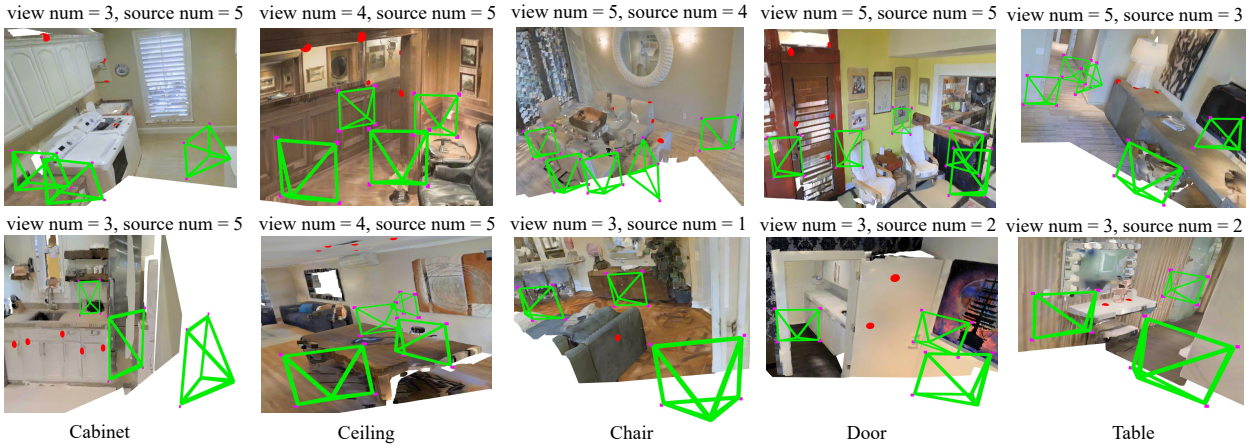


Figure 2. *SoundLoc3D* experiment data visualization: We visualize the sample data we used in our experiment.

Table 2. Created Multiview Mic-Array and RGBD Dataset Summary w.r.t. each Physical Object Category.

Object	Texture-homo.	Texture-disc.	Source Num.	View Num.
wall	975	717	1-5	4
ceiling	727	614	1-5	4
table	464	461	1-5	4
door	712	702	1-5	4
cabinet	286	292	1-5	4
chair	100	222	1-5	4
sum	3264	3008	/	/

4. More Details on Train and Test Configuration

During training, *SoundLoc3D* integrates the predictions (queries) from each single view by introducing global losses so that *SoundLoc3D* gets optimized by both individual view prediction and crossview prediction consistency. During test, since we adopt *set prediction* strategy to predict sound sources, the sound sources predicted from different views may be different from

each other in terms of predicted sound source 3D position class, we treat the *set prediction* from different views separately and compare per-view prediction with ground truth separately to get the final evaluation metric.

5. More Experiment Result

We train all models with the same experimental setting presented in the main paper. We train all models three times independently and report the mean value. We do not report the standard deviation because of the space limit in Table 7 and Table 6. All the standard deviations are within 0.02.

5.1. More Ablation Study

Table 3. More Ablation Study Quantitative Result.

Methods	**mAP** (↑)	**mAR** (↑)	**mALE** (↓)
SL3D_Res50	0.488 ± 0.002	0.932 ± 0.020	0.520 ± 0.005
SL3D_noDeepSup	0.487 ± 0.002	0.960 ± 0.001	0.391 ± 0.002
Ours Sound3DLoc	**0.518** ± 0.010	**0.999** ± 0.001	**0.320** ± 0.001

In the main paper, we reported three ablation studies. We further report another two ablation studies in Table 3 to validate the efficiency of SoundLoc3D.

1. LoFTR vs ResNet LoFTR [9] is better suited to our problem setup as it uses the projections of sound source locations with visual consistency. We test the performance of replacing LoFTR with widely used ImageNet [3] pre-trained ResNet50 [4] as image feature extractor. This variant, we call, *SL3D_Res50* leads to performance drop as well, which indirectly shows visual consistency is a vital cue for sound source localization.

2. Without Deep Supervision. In *Sound3DLoc*, we jointly train both the initial queries and updated queries. We ablate the performance without deep supervision. To this end, we remove the loss (Eqn. (14) in the main paper) added to the initial queries (*SL3D_noDeepSup*). From Table 3, we can see that removing deep supervision strategy leads to performance drop.

In summary, from all ablation studies, we can validate the necessity and importance of each component of our *SoundLoc3D* framework design.

5.2. Quantitative Result w.r.t Sound Source Class

The quantitative result w.r.t. sound source class is given in Table 6. From this table, we can observe that 1) *SoundLoc3D* stays as the best-performing method among all comparing methods and all *SoundLoc3D* variants used in ablation studies. 2) As the training dataset size increases (so the acoustic scenes' visual variation increases accordingly), the three comparing methods have observed performance drop while our proposed *SoundLoc3D* maintains nearly the same performance. It thus shows our proposed *SoundLoc3D* can better handle visual variation challenge.

5.3. Quantitative Result w.r.t Physical Object Class

The detailed quantitative result of our method w.r.t. physical object class is given in Table 7. We can observe from this table that 1) our proposed *SoundLoc3D* outperforms all other *SoundLoc3D* variants across all physical object classes, in terms of both mAP, mAR and mALE metrics. 2) *SoundLoc3D* and its variants achieve better performance on surface flat objects (such as Table, Ceiling and Wall) than on surface uneven objects (Chair and Cabinet and Door). It thus shows that localizing and classifying 3D sound sources on cluttered and uneven surface is a challenging task that requires more future work.

5.4. Comparing Methods with RGBD Image Input

All the 6 comparing methods SELDNet [1], EIN-v2 [2], SoundDoA [5], SoundDet [7], SALSA [8], SALSA-Lite [10] are just based on Mic-Array signal input. One question that naturally arises is that what if feeding the RGBD images to these Mic-Array based methods. To this end, we further run two kinds of extra experiments:

First, we simply combine the Mic-Array signal feature map ($10 \times 256 \times 256$) with its corresponding RGBD image ($4 \times 256 \times 256$) for each single view. We then obtain a 13-channel 2D Mic-Array and RGBD feature map and feed its neural network to localize and classify sound sources. It helps to test if directly concatenating RGBD can improve Mic-Array based methods' performance. We run such test on SELDNet [1], EIN-v2 [2], SALSA [8], SALSA-Lite [10] and Sound3DVDet [6], we do not include SoundDoA [5] and SoundDet [7] because SoundDoA [5] and SoundDet [7] do not directly generate fixed size Mic-Array based 2D feature (they propose learnable filter bank to directly learn from sound raw waveform). The

Table 4. Quantitative results of comparing methods w/o single view RGBD input.

Methods	**mAP** (↑)	**mAR** (↑)	**mALE** (↓)
SELDNet [1]	0.103 ± 0.002	0.501 ± 0.001	0.923 ± 0.001
SELDNet [1] + RGBD	0.093 ± 0.001	0.489 ± 0.002	0.943 ± 0.001
EIN-v2 [2]	0.113 ± 0.002	0.607 ± 0.001	0.878 ± 0.001
EIN-v2 [2] + RGBD	0.101 ± 0.001	0.591 ± 0.001	0.899 ± 0.001
SALSA [8]	0.147 ± 0.002	0.722 ± 0.002	0.793 ± 0.003
SALSA [8] + RGBD	0.133 ± 0.002	0.701 ± 0.001	0.813 ± 0.002
SALSA-Lite [10]	0.130 ± 0.010	0.712 ± 0.003	0.810 ± 0.002
SALSA-Lite [10] + RGBD	0.107 ± 0.006	0.697 ± 0.002	0.831 ± 0.001
Sound3DVDet [6]	0.309 ± 0.010	0.998 ± 0.007	0.586 ± 0.009
Sound3DVDet [6] + RGBD	0.278 ± 0.009	0.892 ± 0.002	0.687 ± 0.007
SoundLoc3D	**0.518** ± 0.010	**0.999** ± 0.001	**0.320** ± 0.001

quantitative result is given in Table 4, from which we can see that simply concatenating RGBD images to Mic-Array feature leads to reduced performance for all comparing methods. It thus shows single view RGBD image does not present explicit 3D sound source localization clue.

Second, we further follow *SoundLoc3D* pipeline to add multiview RGB-informed sound source position visual appearance constraint to the comparing methods. Specifically, we replace *SoundLoc3D*'s Query Generator $\mathcal{G}$ with the comparing Mic-Array based methods, and keep the remaining *SoundLoc3D* component the same (including the Feature Mixer $\mathcal{M}$ and Query Decoder $\mathcal{D}$). Such setting helps test the feasibility of our multiview RGBD feature aggregation scheme. The quantitative result is given in Table 5, from which we can see that involving multiview RGB-D informed 3D sound source clue significantly improves their corresponding performance (in terms of both mAP, mAR and mALE evaluation metric). It thus shows aggregating cross-modal vision-informed clue for sound source localization and classification can dramatically improve the performance, even though the sound source exhibits no visual entity.

Table 5. Quantitative results of comparing methods w/o multiview RGBD-Informed Sound Source Clue Aggregation.

Methods	**mAP** (↑)	**mAR** (↑)	**mALE** (↓)
SELDNet [1]	0.103 ± 0.002	0.501 ± 0.001	0.923 ± 0.001
SELDNet + mvRGBD	0.208 ± 0.001	0.635 ± 0.001	0.852 ± 0.001
EIN-v2 [2]	0.113 ± 0.002	0.607 ± 0.001	0.878 ± 0.001
EIN-v2 + mvRGBD	0.145 ± 0.001	0.687 ± 0.001	0.822 ± 0.001
SALSA [8]	0.147 ± 0.002	0.722 ± 0.002	0.793 ± 0.003
SALSA + mvRGBD	0.289 ± 0.001	0.810 ± 0.001	0.700 ± 0.002
SALSA-Lite [10]	0.130 ± 0.010	0.712 ± 0.003	0.810 ± 0.002
SALSA-Lite + mvRGBD	0.269 ± 0.003	0.792 ± 0.001	0.732 ± 0.001
Sound3DVDet [6]	0.309 ± 0.010	0.998 ± 0.007	0.586 ± 0.009
Sound3DVDet [6] + mvRGBD	0.378 ± 0.006	0.999 ± 0.006	0.501 ± 0.005

5.5. More Qualitative Result

We provide more qualitative result visualization in Fig. 3. From this figure, we can clearly see that *SoundLoc3D* is capable of accurately detect 3D sound sources under various room scenarios. It is better at handling both texture-homogeneous and texture-discriminative situations.

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. 3, 4, 5, 6

[2] 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. 3, 4, 5, 6

[3] 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. 3

Table 6. Quantitative Result w.r.t. Each Sound Source Classes.

Methods	Telephone			Siren			Alarm		
	AP	AR	ALE	AP	AR	ALE	AP	AR	ALE
SELDNet [1]	0.104	0.503	0.918	0.102	0.500	0.924	0.103	0.500	0.924
EIN-v2 [2]	0.112	0.606	0.881	0.114	0.608	0.881	0.112	0.607	0.879
SoundDoA [5]	0.210	0.764	0.801	0.213	0.760	0.800	0.210	0.761	0.793
SALSA [8]	0.144	0.723	0.791	0.146	0.722	0.794	0.149	0.719	0.794
SALSA-Lite [10]	0.126	0.710	0.812	0.131	0.712	0.808	0.128	0.715	0.811
SoundDet [7]	0.119	0.670	0.820	0.120	0.675	0.823	0.117	0.672	0.825
Sound3DVDet [6]	0.308	0.999	0.600	0.320	0.997	0.577	0.320	0.998	0.579
SL3D_noRGB	0.499	0.944	0.513	0.497	0.945	0.510	0.499	0.947	0.509
SL3D_noDepth	0.471	0.911	0.459	0.473	0.908	0.453	0.473	0.908	0.458
SL3D_noCVC	0.500	0.948	0.391	0.504	0.949	0.387	0.497	0.945	0.391
SL3D_noRGBD	0.330	0.730	0.811	0.327	0.736	0.807	0.327	0.729	0.814
SL3D_Res50	0.494	0.930	0.522	0.490	0.937	0.527	0.487	0.933	0.518
SL3D_noDeepSup	0.488	0.962	0.392	0.490	0.961	0.391	0.485	0.965	0.388
Ours *SoundLoc3D*	**0.519**	**0.999**	**0.317**	**0.517**	**0.998**	**0.323**	**0.523**	**0.998**	**0.324**

Methods	Fireplace			Horn-beeps			Overall		
	AP	AR	ALE	AP	AR	ALE	**mAP**	**mAR**	**mALE**
SELDNet [1]	0.101	0.499	0.923	0.104	0.500	0.926	0.103	0.501	0.923
EIN-v2 [2]	0.115	0.608	0.873	0.113	0.605	0.870	0.113	0.607	0.878
SoundDoA [5]	0.214	0.761	0.800	0.212	0.761	0.796	0.212	0.762	0.800
SALSA [8]	0.146	0.723	0.792	0.147	0.720	0.791	0.147	0.722	0.793
SALSA-Lite [10]	0.131	0.714	0.810	0.130	0.710	0.809	0.130	0.712	0.810
SoundDet [7]	0.118	0.676	0.824	0.121	0.671	0.820	0.120	0.674	0.823
Sound3DVDet [6]	0.322	0.999	0.586	0.220	0.996	0.588	0.301	0.998	0.584
SL3D_noRGB	0.499	0.940	0.512	0.498	0.945	0.512	0.498	0.944	0.510
SL3D_noDepth	0.471	0.910	0.457	0.473	0.911	0.459	0.472	0.910	0.457
SL3D_noCVC	0.501	0.950	0.391	0.503	0.946	0.387	0.501	0.948	0.389
SL3D_noRGBD	0.330	0.734	0.809	0.331	0.732	0.813	0.328	0.732	0.810
SL3D_Res50	0.488	0.929	0.522	0.489	0.933	0.519	0.488	0.932	0.520
SL3D_noDeepSup	0.492	0.964	0.394	0.488	0.962	0.387	0.487	0.960	0.391
Ours SoundLoc3D	**0.518**	**0.999**	**0.315**	**0.520**	**0.997**	**0.317**	**0.518**	**0.999**	**0.320**

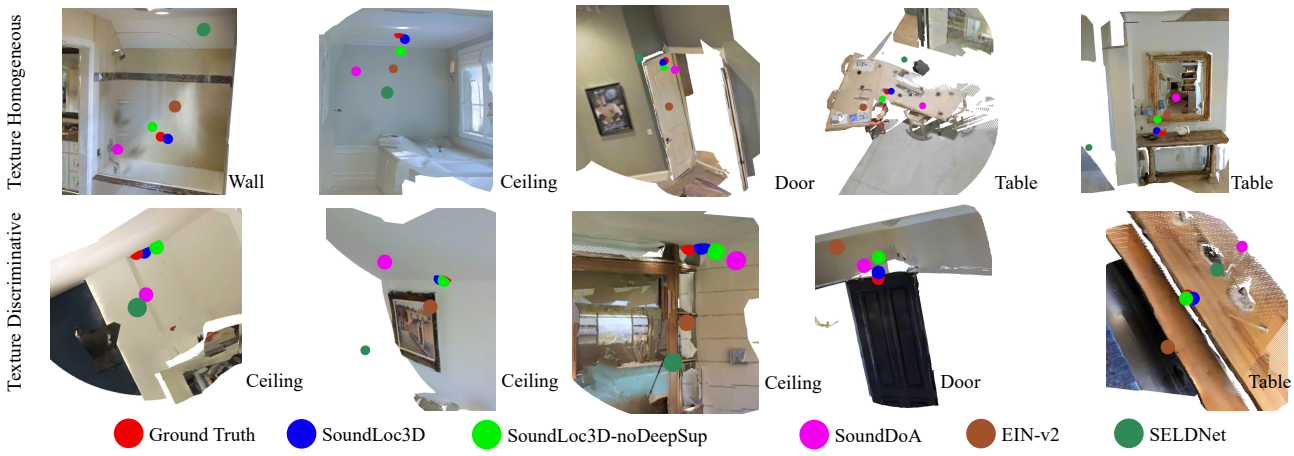


Figure 3. **More qualitative result**: We visualize the localization result for one sound source in different visual scenes. We also provide the visualization source code and data for more directive visualization.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 3

[5] Yuhang He and Andrew Markham. SoundDoA: Learn Sound Source Direction of Arrival and Semantics from Sound Raw Waveforms.

Table 7. Quantitative Result w.r.t. Each Physical Object Class.

Methods	Table			Ceiling			Door		
	mAP	mAR	mALE	mAP	mAR	mALE	mAP	mAR	mALE
SELDNet [1]	0.105	0.505	0.916	0.110	0.507	0.914	0.090	0.490	0.945
EIN-v2 [2]	0.114	0.608	0.870	0.117	0.610	0.865	0.109	0.598	0.870
SoundDoA [5]	0.211	0.765	0.800	0.215	0.766	0.793	0.200	0.750	0.800
SALSA [8]	0.145	0.725	0.790	0.147	0.725	0.787	0.140	0.705	0.803
SALSA-Lite [10]	0.127	0.718	0.822	0.135	0.718	0.800	0.122	0.705	0.821
SoundDet [7]	0.677	0.815	0.122	0.682	0.813	0.110	0.660	0.833	0.108
Sound3DVDet [6]	0.266	0.970	0.611	0.300	0.980	0.602	0.348	0.990	0.581
SL3D_noRGB	0.501	0.946	0.507	0.507	0.951	0.501	0.477	0.926	0.520
SL3D_noDepth	0.473	0.917	0.450	0.476	0.918	0.443	0.460	0.900	0.467
SL3D_noCVC	0.501	0.949	0.389	0.505	0.951	0.382	0.492	0.942	0.395
SL3D_noRGBD	0.332	0.734	0.807	0.335	0.737	0.801	0.321	0.717	0.835
SL3D_Res50	0.495	0.932	0.520	0.501	0.947	0.517	0.472	0.921	0.534
SL3D_noDeepSup	0.489	0.964	0.390	0.488	0.966	0.384	0.484	0.958	0.399
SoundLoc3D	**0.520**	**0.998**	**0.318**	**0.519**	**0.999**	**0.318**	**0.514**	**0.997**	**0.327**

Methods	Chair			Wall			Cabinet		
	mAP	mAR	mALE	mAP	mAR	mALE	mAP	mAR	mALE
SELDNet [1]	0.089	0.481	0.950	0.108	0.505	0.919	0.091	0.497	0.926
EIN-v2 [2]	0.100	0.578	0.894	0.117	0.610	0.868	0.103	0.589	0.898
SoundDoA [5]	0.200	0.750	0.812	0.215	0.767	0.780	0.210	0.757	0.802
SALSA [8]	0.137	0.709	0.810	0.150	0.729	0.771	0.138	0.708	0.813
SALSA-Lite [10]	0.120	0.701	0.841	0.139	0.722	0.798	0.121	0.707	0.818
SoundDet [7]	0.108	0.657	0.849	0.125	0.678	0.809	0.107	0.662	0.820
Sound3DVDet [6]	0.220	0.923	0.613	0.294	0.990	0.588	0.300	0.975	0.579
SL3D_noRGB	0.470	0.920	0.544	0.504	0.950	0.500	0.480	0.927	0.543
SL3D_noDepth	0.461	0.900	0.469	0.483	0.923	0.430	0.462	0.901	0.469
SL3D_noCVC	0.489	0.932	0.410	0.505	0.952	0.367	0.492	0.940	0.396
SL3D_noRGBD	0.312	0.712	0.819	0.339	0.741	0.802	0.317	0.729	0.828
SL3D_Res50	0.477	0.912	0.539	0.498	0.948	0.530	0.480	0.928	0.528
SL3D_noDeepSup	0.478	0.957	0.410	0.498	0.973	0.362	0.469	0.950	0.412
SoundLoc3D	**0.513**	**0.998**	**0.330**	**0.520**	**0.999**	**0.313**	**0.510**	**0.997**	**0.329**

In *Interspeech*, 2022. 3, 5, 6

[6] 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. 1, 3, 4, 5, 6

[7] 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. 3, 5, 6

[8] 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. 3, 4, 5, 6

[9] 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. 1, 3

[10] 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. 3, 4, 5, 6