

RoboSense: Large-scale Dataset and Benchmark for Egocentric Robot Perception and Navigation in Crowded and Unstructured Environments

Haisheng Su^{1,2}, Feixiang Song², Cong Ma², Wei Wu^{2,3}, Junchi Yan^{1♡} (♡*corresponding author*)

¹School of Computer Science, Shanghai Jiao Tong University ²SenseAuto Research ³Tsinghua University

How to Improve Robot Perception Capacities in Crowded and Unstructured Environments ?

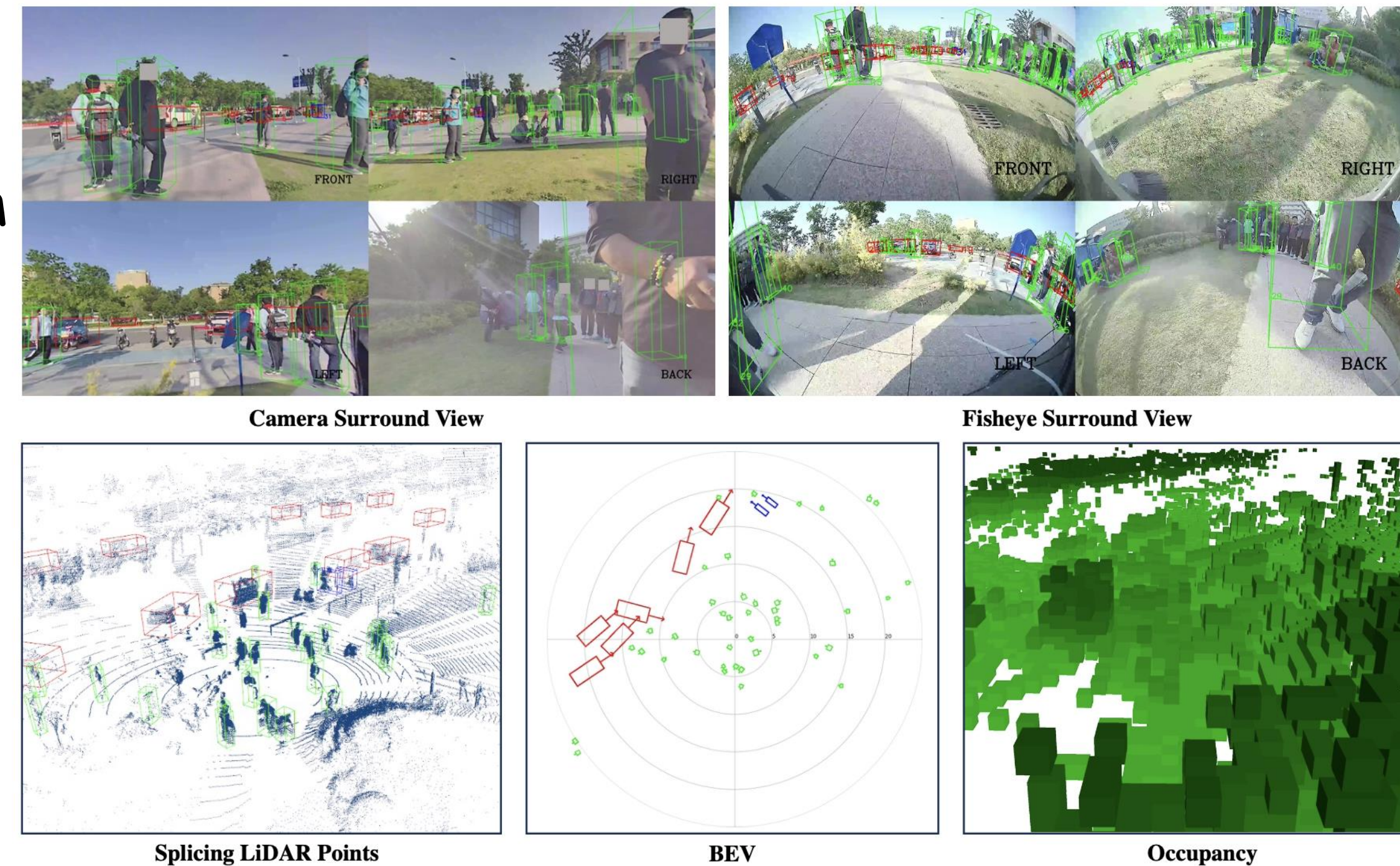
💡 Contribution:

- ✓ **Unique Collected Domain:** Unstructured navigation routes
- ✓ **Diverse Annotation:** 3D box + Trajectory + Occupancy
- ✓ **Social Crowded Scene :** Closer obstacles, with a mode at 5m



➤ Sensor Setup

➤ Dataset Example



😊 Benchmark: Perception and Prediction

Task	Method	Vehicle@ $p=5\%/10\%$			Cyclist@ $p=5\%/10\%$			Pedestrian@ $p=5\%/10\%$		
		3D AP $\uparrow$	AOS $\uparrow$	ASE $\downarrow$	3D AP $\uparrow$	AOS $\uparrow$	ASE $\downarrow$	3D AP $\uparrow$	AOS $\uparrow$	ASE $\downarrow$
LiDAR 3D Detection	PointPillar [15]	72.5/53.0	73.5/61.1	20.6/16.1	44.2/32.8	45.4/38.3	64.2/54.3	62.7/38.2	45.3/34.1	38.3/27.2
	SECOND [39]	78.8/63.1	80.2/69.4	19.8/15.7	53.8/43.5	57.2/49.9	67.7/55.7	70.8/ 47.2	54.6/43.2	40.1/29.3
	PVRCNN [29]	74.6/57.4	77.4/67.7	16.4 /15.4	53.6/41.4	55.7/50.1	62.5/61.9	66.4/39.1	50.1/37.0	40.4/25.5
	Transfusion-L [1]	83.6/65.1	84.5/73.8	19.7/16.0	59.7/47.0	78.0/70.8	82.1/72.9	72.3/42.8	60.5/48.7	45.1/37.4
Multi-view 3D Detection	BEVDet [12]	76.2/30.2	40.4/25.9	17.3/11.2	42.3/25.7	36.1/30.2	56.5/42.1	47.4/28.5	48.6/36.5	30.2/18.8
	BEVDet4D [11]	77.2/31.1	41.1/26.4	16.8/10.8	42.0/24.8	33.9/27.7	55.3/ 41.2	48.1/29.3	46.6/37.6	27.5 /21.3
	BEVDepth [16]	77.8/31.3	40.9/26.3	16.7/10.7	43.3/27.0	34.9/30.2	52.2 /46.6	50.1/31.3	46.7/37.9	28.0/21.4
	BEVFormer [17]	78.2/32.0	41.6/26.7	16.5/ 10.6	44.1/27.6	34.9/30.5	51.3/44.3	50.2/32.3	46.3/38.0	28.1/ 17.9

Task	Detector	Layouts	Metric	Detection			Tracking			
				Range(m)			sAMOTA $\uparrow$	AMOTP $\uparrow$	MT $\uparrow$	ML $\downarrow$
				[0, 5]	[5, 10]	[10, 30]				
Multi-view 3D Perception	BEVDepth [16]	4C	3D AP	54.9/16.0	60.1/18.3	53.7/33.1	44.03	29.95	20.23	54.01
			AOS	44.8/19.7	37.0/18.8	34.5/26.9				
		4F	3D AP	61.1/16.9	70.6/19.9	50.8/29.0	39.56	27.10	18.02	61.74
			AOS	58.7/27.5	41.3/23.5	36.1/27.4				
		4C + 4F	3D AP	68.9 /20.5	75.2 /22.9	64.2/38.6	51.16	35.68	25.21	48.07
			AOS	53.9/24.4	43.1/22.5	39.6/30.9				
LiDAR 3D Perception	PointPillar [15]	4L	3D AP	59.2/19.3	73.1/42.0	71.0/65.4	44.77	33.65	25.04	54.08
Multi-modal 3D Perception	BEVDepth [16] + Pointpillar [15]	8V + 4L	3D AP	61.3/ 36.9	61.3/ 54.6	54.4/52.6	43.32	43.18	34.74	40.82
			AOS	64.8 /49.6	78.7 /75.0	79.4 /78.4				

Method	minADE (m) $\downarrow$	minFDE (m) $\downarrow$	MR $\downarrow$	EPA $\uparrow$
Constant Pos.*	2.42	3.01	0.319	0.680
Constant Vel.*	1.59	3.54	0.219	0.780
Constant Pos. $\uparrow$	1.52	1.95	0.267	0.243
Random Vel. $\uparrow$	2.56	3.85	0.872	0.029
ViP3D [9]	1.31	1.55	0.196	0.283
PnPNet [18]	0.89	1.12	0.172	0.313

Range(m)	mIoU-3D $\uparrow$	mIoU-BEV $\uparrow$
[0, 12.8]	24.6	29.7
[0, 2]	39.6	48.2
[2, 5]	30.7	36.7
[5, 12.8]	16.1	19.7

----- Supported tasks -----

➤ Perception:

1. LiDAR 3D Object Detection
2. Multi-view/modal 3D Detection
3. Multi-Object Tracking

➤ Prediction:

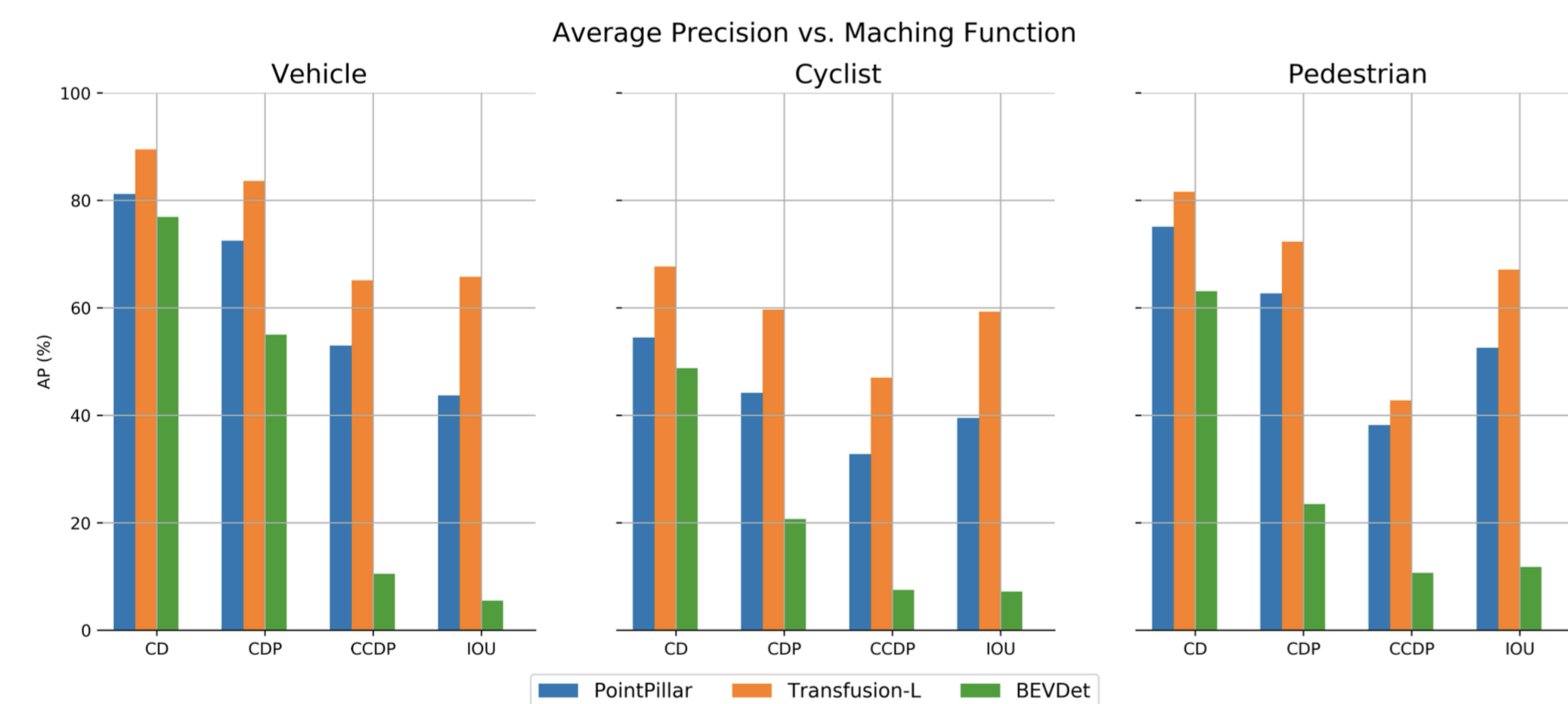
1. Motion Forecasting
2. Occupancy Prediction

Dataset Comparison and Constitution

Dataset	Year	Size (hr)	Ann. Scenes	Ann. Frames	With Trajectory	Multi-view Overlapping	Sensor Layouts	3D Boxes (Total)	3D Boxes $\uparrow$ ($\leq 5m$)
KITTI [7]	2012	1.5	22	15K	✗	✗	4C+1L	80K	638
Cityscapes [5]	2016	-	-	25K	✗	✗	1C	0	0
ApolloScope [13]	2016	2	-	144K	✗	✗	1L	70K	4.7K
H3D [23]	2019	0.77	160	27K	✗	✓	3C+1L	1.1M	-
Lyft L5 [14]	2019	2.6	366	55K	✓	✓	7C+3L	1.3M	-
nuScenes [3]	2019	5.5	1K	40K	✓	✓	6C+1L	1.4M	9.8K
Argoverse [4]	2019	0.6	113	22K	✓	✓	9C+2L	993K	15K
Waymo Open [32]	2019	6.4	1K	200K $\ddagger$	✓	✓	5C+5L	12M $\ddagger$	123K $\ddagger$
BDD100k [40]	2020	1K	100k	100k	✗	✗	1C	0	0
RoboSense (Ours)	2024	42	7.6K	133K	✓	✓	4C+4F+4L	1.4M	173K

Scene-ID	Distribution			Ratio of Dataset			Num of Sequences	Num of Frames	Num of 3D Boxes	Num of Trajectories
	Day	Night	Scene	Train	Test	Val				
S-1	56%	44%	20%	50%	40%	10%	1.5K	26K	310K	36K
S-2	69%	31%	30%				2.3K	42K	293K	37K
S-3	71%	29%	17%				1.2K	22K	284K	64K
S-4	83%	17%	7%				0.5K	9K	144K	22K
S-5	70%	30%	20%	0%	100%	0%	1.6K	26K	297K	44K
S-6	22%	78%	6%				0.5K	8K	88K	13K
Total	65%	35%	100%	46%	44%	10%	7.6K	133K	1.4M	216K

Metric Comparison: Average Precision vs. Matching Function



IOU: Intersection Over Union, CD: Center Distance
CDP: CD Proportion, CCDP: Closest Collision Distance Proportion



SCAN ME