Supplementary Materials for CMD

Jinhao Deng[1,*], Wei Ye[1,*], Hai Wu[1], Xun Huang[1], Qiming Xia[1], Xin Li[3], Jin Fang[2], Wei Li[2,✉], Chenglu Wen[1,✉], and Cheng Wang[1]

[1] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China
[2] Inceptio, Shanghai, China
[3] Section of Visual Computing and Interactive Media, Texas A&M University, Texas, USA

1 Details for CMD

Sequences. Our CMD covers multiple road and illumination conditions. Due to space constraints, the visualizat·ion in Section 3 selectively features three representative scenes: tunnel, suburban road, and bridge. However, in reality, we carefully selected night distinct scenarios, namely urban road, urban highway, bridge, residential area, tunnel, campus, village, suburban road, and suburban highway. Additionally, our CMD encompasses five different lighting conditions: daylight, cloudy, dusk, and night. Table 1 presents the corresponding weather and lighting conditions for all sequences, while Fig. 1 provides visualizations of these different scenes under varying lighting conditions. The comprehensive coverage of scenes in CMD, along with a well-balanced partitioning of training, validation, and test sets, enhances the dataset's ability to train models with superior generalization performance. This design also facilitates further research into domain adaptation algorithms across diverse scenarios and lighting conditions.

Annotation details. More detailed annotation specifications are listed in this stage. For each bounding box, there are 6 kinds of information labeled, including:

- category,
- tracking ID,
- label confidence,
- 9 DoF bounding box,
- motion state,
- occlusion.

1) For category, tracking ID and 9 DoF bounding box, more details are provided in Section 3.3. 2) Regarding label confidence, this information is derived from the number of points within the bounding box and the point cloud's coverage area, as illustrated in Table 2. The term "coverage" refers to the minimum ratio between the point distribution along the two horizontal axes of the

* Equal contribution
✉ Corresponding authors: W. Li (liweimcc@gmail.com); C. Wen (clwen@xmu.edu.cn)

3D bounding box, namely length and width. 3) Occlusion is manually annotated and categorized into three levels: 0 indicates the object is completely visible with less than 10% occlusion, 1 denotes partial occlusion with 10% to 50% coverage, and 2 signifies substantial occlusion with more than 50% coverage. Additionally, situations where occlusion is not observable through manual inspection are labeled as 3. The annotation of label confidence and occlusion is crucial for assessing the difficulty of object detection and the accuracy of labels, allowing for multi-level analysis of detection results and expanding the potential applications of CMD. 4) For motion state, objects are classified into rigid (e.g., car) and non-rigid (e.g., adult pedestrian) categories. Rigid objects are labeled as stopping, parking, or moving. Non-rigid pedestrians are categorized as ped standing, ped sitting, ped lying, or ped moving. Annotating motion states in this manner is essential for distinguishing between static and dynamic objects in the scene, providing valuable insights for subsequent tracking and related tasks.

Seq	Split	Road	Illumination	Seq	Split	Road	Illumination
1	train	Highway	Bright	26	train	Campus	Normal
2	train	Highway	Bright	27	test	Campus	Dark
3	train	Highway	Bright	28	train	Campus	Normal
4	val	Urban	Bright	29	train	Campus	Normal
5	test	Highway	Bright	30	val	Campus	Normal
6	train	Highway	Bright	31	test	Campus	Normal
7	val	Bridge	Bright	32	train	Campus	Normal
8	val	Bridge	Bright	33	train	Village	Normal
9	train	Bridge	Bright	34	test	Village	Normal
10	train	Bridge	Bright	35	train	Village	Normal
11	test	Bridge	Bright	36	val	Village	Normal
12	train	Bridge	Bright	37	train	Suburban	Dim
13	train	Bridge	Bright	38	val	Suburban	Dim
14	train	Highway	Night	39	train	Suburban	Dim
15	train	Urban	Dim	40	train	Suburban	Dim
16	val	Urban	Dim	41	test	Suburban	Dim
17	test	Urban	Dim	42	train	Suburban	Dim
18	test	Highway	Night	43	val	Suburban	Dim
19	train	Tunnel	Night	44	train	Suburban	Dim
20	train	Tunnel	Night	45	test	Suburban	Dim
21	val	Tunnel	Night	46	train	Suburban	Dim
22	train	Tunnel	Night	47	train	Suburban	Dim
23	train	Campus	Dark	48	train	Suburban	Dim
24	val	Campus	Dark	49	test	Suburban	Dim
25	train	Campus	Normal	50	train	Highway	Night

Table 1: Details of sequence. Seq. refers to sequence.

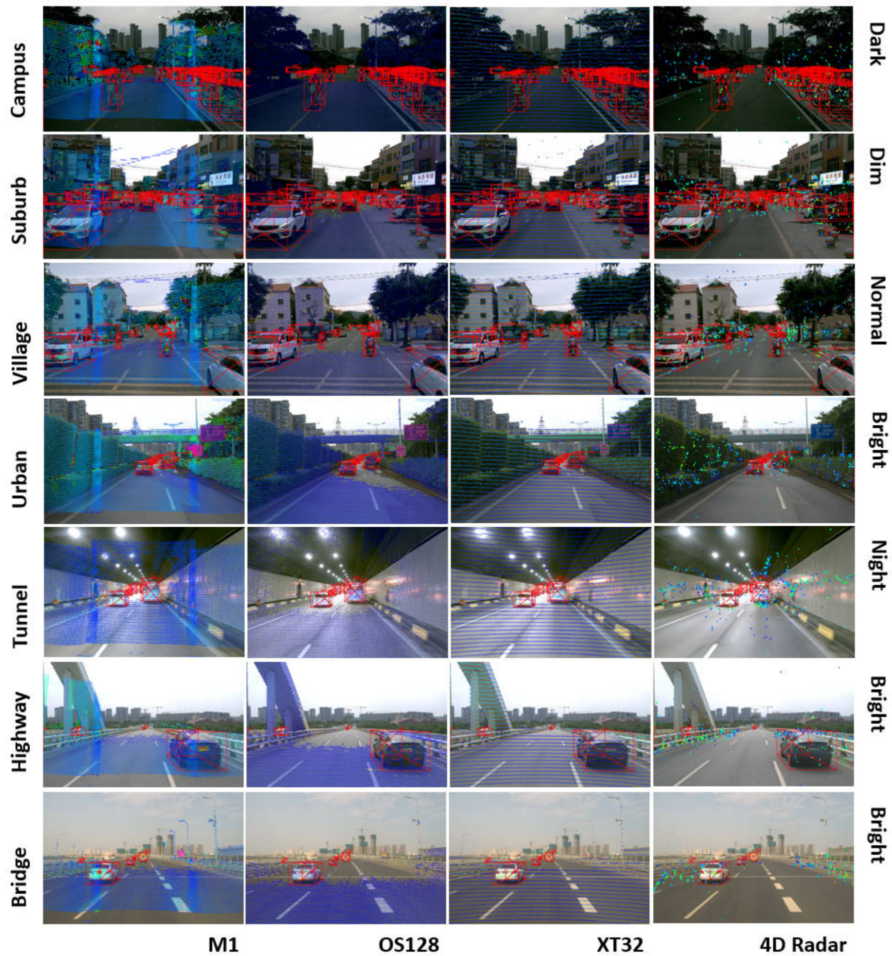

Fig. 1: Visualization for multiple scenarios.

label conf.	points number	coverage
1	<10	-
2	>10	-
3	>30	>30%
4	>50	>50%
5	>100	>80%

Table 2: Label confidence

Source	mAP	Car				Truck				Ped				Cyc			
		all	0-30m	30-50m	50m-inf	all	0-30m	30-50m	50m-inf	all	0-30m	30-50m	50m-inf	all	0-30m	30-50m	50m-inf
Oracle	6.17	7.92	20.06	7.69	1.44	6.14	10.67	15.65	0.11	3.74	8.33	0	0	6.88	20.6	4.42	0
OS128	0.03	0.01	0.03	0	0	0.01	0.02	0.01	0.01	0	0	0	0	0.09	1.03	0.04	0
M1	0.06	0.03	0.05	0	0	0.01	0	0	0	0.15	0.35	0	0	0.06	0.38	0.01	0
XT32	0.04	0	0	0	0	0	0	0	0	0.07	0.32	0	0	0.09	0.82	0.02	0

Table 3: Domain adaptation with 4D radar as target domain.

2 More Experiment Results

Results for 4D Radar. Due to space constraints, experiments involving 4D radar are not included in the main text. Nevertheless, we conducted comprehensive experimental validations on it. As shown in Table 3, the model's Mean Average Precision (mAP) on both training and validation datasets for radar is inherently low, with a specific mAP value of 6.17. This outcome is primarily attributed to: 1) the inherently sparse nature of 4D radar data (nearly 1000 points per second), and 2) the considerably low angular resolution of 4D radar (horizontal angle resolution of 1.9° and vertical angle resolution of 3.9°). Given this foundational model, the domain adaptation effectiveness is also limited. Coupled with the significant original data disparity, domain adaptation for the 4D radar poses even greater challenges. We hope that future work will address this issue with innovative designs to overcome these challenges effectively.

More 3D object detetion on our CMD. In the main paper, considering that Voxel-RCNN is a widely used classical voxel-based detector, we selected it as the primary 3D object detector to construct our benchmark using various domain adaptation methods. In fact, our CMD is also suitable for validating the effectiveness of point cloud-based or multi-modal object detection methods. As shown in Table 4, we selected three different point cloud-based 3D object detectors: CasA, PV-RCNN, and PV-RCNN++, which achieved mAP scores of 31.0, 35.1, and 37.7, respectively. Additionally, we used the multi-modal BEVFusion, achieving an mAP score of 35.0 for the FOV region of the front-facing camera, and an mAP score of 37.0 by combining LiDAR and 4D millimeter-wave radar features using feature concatenation.

Modal	Method	mAP	Car	Truck	Ped.	Cyc.
L	CasA	31.0	43.2/67.2	19.7/30.4	19.2/31.9	42.0/61.1
L	PV-RCNN	35.1	46.0/75.0	21.9/36.9	25.0/40.6	47.4/74.0
L	PV-RCNN++	37.7	47.7/77.8	22.8/40.6	24.8/41.7	55.4/80.8
L+C	BEVFusion	35.0	42.0/72.4	24.2/47.8	19.3/35.3	54.6/80.5
L+R	FeatureConcat	37.0	47.3/78.0	25.4/43.6	26.3/41.9	49.0/72.3

Table 4: Results on CMD with more 3D object detectors.

DIG on Waymo to KITTI. To validate the effectiveness of our proposed Density-Intensity-Geometry (DIG) method in the widely studied cross-dataset

Car AP@40 (IoU=0.5)	Easy	Mod.	Hard
DT	89.25	76.00	77.57
DIG	89.63	85.17	84.58

Table 5: Performance of DIG for domain adaptation from Waymo to KITTI.

domain adaptation setting, we conducted relevant experiments. As shown in Table 5, for the domain adaptation task from Waymo to KITTI, the DIG method demonstrates significant improvements in the Car category, with increases of 9.17 and 7.01 AP in the "Moderate" and "Hard" levels, respectively.