Towards Robust 3D Robot Perception in Urban Environments: The UT Campus Object Dataset

Arthur Zhang, Chaitanya Eranki, Christina Zhang, Raymond Hong, Pranav Kalyani,
Lochana Kalyanaraman, Arsh Gamare, Maria Esteva, Joydeep Biswas

I. INTRODUCTION

We contribute the UT Campus Object Dataset (CODa), a mobile robot egocentric perception dataset collected at the University of Texas Austin Campus, to train and evaluate state-of-the-art perception models in urban environments. CODa provides significantly more object classes, diverse 3D bounding box annotations, and lighting and weather variations than any other robot perception dataset to date. It is comparable in scope and magnitude to existing AV datasets, including Waymo and nuScenes. We release CODa on the Texas Data Repository [1], pretrained models, and dataset development kit.

CODa contains synchronized 3D point clouds and stereo RGB video from a 128-channel 3D LiDAR and two RGB cameras at 10 fps, RGB-D videos at 7 fps, inertial data from a 9-DOF IMU at 40 Hz, 1.3 million 3D bounding box annotations for 53 semantic classes, 5000 frames of 3D semantic annotations for terrain, and pseudo-ground truth poses. To collect the data our team repeatedly drove the robot in indoor and outdoor areas under rainy, cloudy, sunny, and low-light conditions. Deepen annotators manually labeled 58 minutes of frames, followed by manual quality assurance checks to ensure the accuracy of 3D point cloud labels. We use these labels to show that object detection performance in urban environments is significantly higher when trained on CODa compared to other datasets. We provide benchmarks for 3D object detection and semantic segmentation, with more planned in the future. CODa and its benchmarks will support research toward egocentric 3D perception and navigation in urban environments for the computer vision and robotics communities.

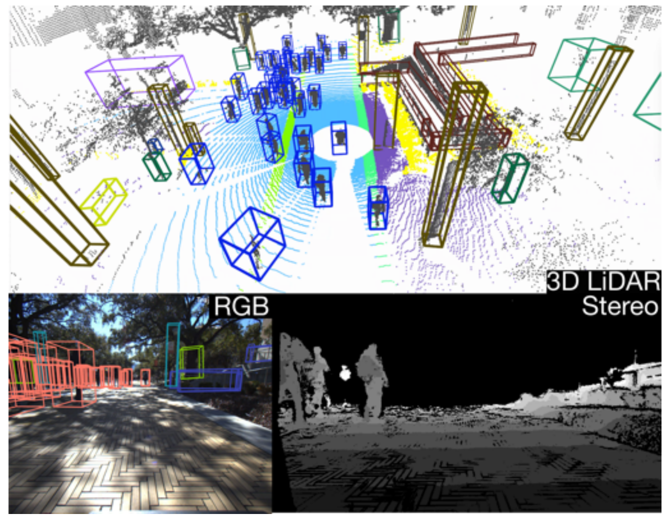

Fig. 1: Three of the five modalities available in CODa. **3D point cloud** (top), **RGB** image with 3D object labels (bottom left), **stereo depth** image (bottom right).

II. RESULTS

We selected the PV-RCNN object detector from OpenPCDet and average the birds' eye view and 3D bounding box metrics from the KITTI Vision Benchmark Suite over pedestrian, vehicle, and cyclist classes. We compared with AV datasets because of similarity in annotations and sensors.

Table I lists object detection performance on CODa when trained on AV datasets, showing that accuracy is greatly improved by training on CODa. Table II presents object detection performance on CODa when trained and tested on various LiDAR resolutions, illustrating that performance improves when LiDAR resolution matches closely between training and testing.

Training		Evaluation					
Pretrain	**Finetune**	**nuScenes**		**Waymo**		**CODa**	
nuScenes	-	**33.85**	**25.41**	x		21.30	15.53
Waymo	-	x		**62.73**	**56.40**	46.20	43.11
CODa	-	x		x		92.08	91.11
nuScenes	CODa	x		x		91.39	90.16
Waymo	CODa	x		x		**93.12**	**92.07**

TABLE I: Average precision (AP) for BEV and 3D object detection. Best result for AP_{BEV} and AP_{3D} in blue and red.

Train \ Test	**CODa-16**		**CODa-32**		**CODa-64**		**CODa-128**	
CODa-16	**75.15**	**73.29**	64.99	63.24	49.17	47.36	21.93	18.94
CODa-32	50.79	47.95	**78.30**	**76.90**	70.49	69.37	59.95	56.59
CODa-64	21.10	22.05	67.27	64.77	**86.20**	**84.48**	77.63	77.53
CODa-128	12.58	12.16	48.05	45.76	76.51	75.38	**92.61**	**91.34**

TABLE II: Effect of point cloud resolution differences on object detection performance.

*This project is partially supported by NSF Awards CAREER-2046955 and IIS-1954778. The authors acknowledge the Texas Advanced Computing Center (TACC) for providing computational resources for the project. All authors are with the University of Texas at Austin, TX 78712, USA.

REFERENCES

[1] A. Zhang, C. Eranki, C. Zhang, R. Hong, P. Kalyani, L. Kalyanaraman, A. Gamare, M. Esteva, and J. Biswas, "UT Campus Object Dataset (CODa)," 2023. [Online]. Available: https://doi.org/10.18738/T8/BBOQMV