OpenAD: Open-World Autonomous Driving Benchmark for 3D Object Detection

Zhongyu Xia[1], Jishuo Li[1], Zhiwei Lin[1], Xinhao Wang[1], Yongtao Wang[1]✉, Ming-Hsuan Yang[2]
[1]Wangxuan Institute of Computer Technology, Peking University [2]University of California, Merced

Paper | Toolkit Code | Online Eval (2D & MLLM) | Online Eval (3D)

> Evaluate Your Open-World 2D/3D Detection Model or MLLM on **OpenAD** Benchmark!

Introduction

Open-World Capabilities to Evaluate:

Different countries, regions, and sensor configuration
Domain generalization

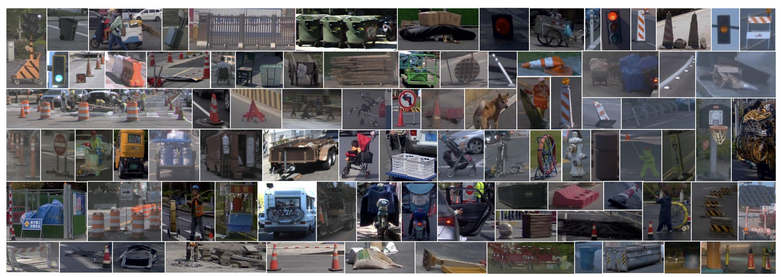
Deal with rare objects, corner cases without user prompt
Open-ended

Motivation & Our Contribution:

- Lack of 3D perception benchmark: We propose the first realworld autonomous driving **benchmark for 3D open-world object detection**, and we design a **labeling pipeline** integrated with MLLM.
- Lack of open-world 3D perception methods: We propose a novel **vision-centric framework for 3D open-world perception**, and further enhance its comprehensive abilities through **General-Specialized Fusion**.

Properties of OpenAD

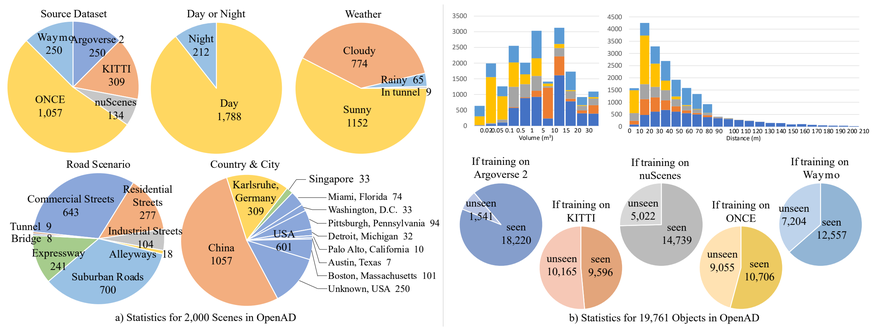
Examples of corner case objects in OpenAD.



OpenAD is the first real-world open-world benchmark for autonomous driving 3D perception. Compared to other real-world datasets, OpenAD boasts greater category diversity and more instances.

Datasets	Sensors	Real	Temporal	Scenes	Classes	Instances	GroundTruth
GTACrash [31]	Cam.	✗	✔	7,720	1	24K*	**Bbox**(2D)
StreetHazards [25]	Cam.	✗	✔	1,500	1	1.5K*	Sem. mask(2D)
Synthetic Fire Hydrants [7]	Cam.	✗	✗	30,000	1	30K*	**Bbox**(2D)
Synthetic Crosswalks [7]	Cam.	✗	✗	20,000	1	20K*	**Bbox**(2D)
CARLA-WildLife [45]	Cam. Depth	✗	✔	26	18	65	**Inst. mask**(2D)
MUAD [19]	Cam. Depth	✗	✗	4,641	9	30K	Sem. mask(2D)
AnoVox [5]	Cam. Lidar	✗	✔	1,368	35	1.4K	**Inst.mask**(2D,**3D**)
YouTubeCrash [31]	Cam.	✔	✔	2,400	1	12K*	**Bbox**(2D)
RoadAnomaly21[12]	Cam.	✔	✗	110	1	0.1K*	Sem. mask(2D)
Street Obstacle Sequences [45]	Cam. Depth	✔	✔	20	13	30*	**Inst. mask**(2D)
Vistas-NP[21]	Cam.	✔	✗	11,167	4	11K*	Sem. mask(2D)
Lost and Found[49]	Cam.	✔	✔	112	42	0.2K*	Sem. mask(2D)
Fishyscapes[4]	Cam.	✔	✗	375	1	0.5K*	Sem. mask(2D)
RoadObstacle21[12]	Cam.	✔	✔	412	1	1.5K*	Sem. mask(2D)
BDD-Anomaly[25]	Cam.	✔	✗	810	3	4.5K	Sem. mask(2D)
CODA[33]	Cam. Lidar	✔	✔	1,500	34	5.9K	**Bbox**(2D)
OpenAD (ours)	Cam. Lidar	✔	✔	2,000	206	19.8K	**Bbox**(2D,**3D**)

Data composition of OpenAD.

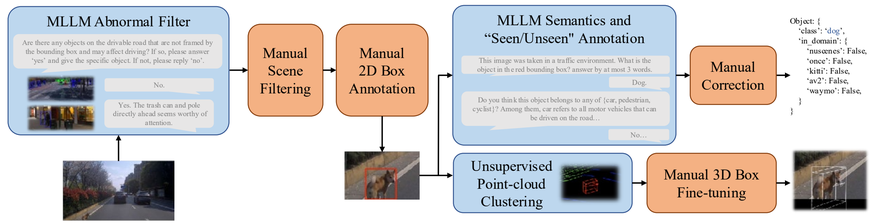


a) Statistics for 2,000 Scenes in OpenAD

b) Statistics for 19,761 Objects in OpenAD

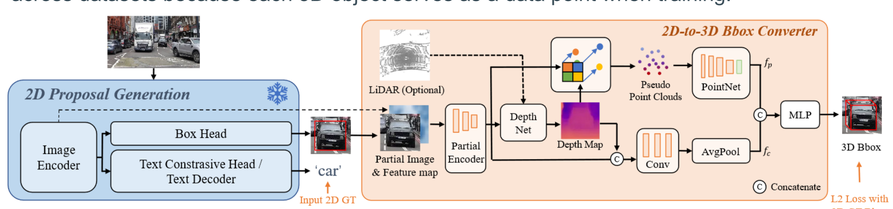
Construction of OpenAD

OpenAD is built upon a semi-automated MLLM corner case discovery and annotation pipeline.



Baseline Method of OpenAD

After obtaining 2D proposals from any frozen open-world 2D object detection model, we train a 2D-to-3D Bbox Converter to predict 3D bounding boxes. It is lightweight and easy to train across datasets because each 3D object serves as a data point when training.

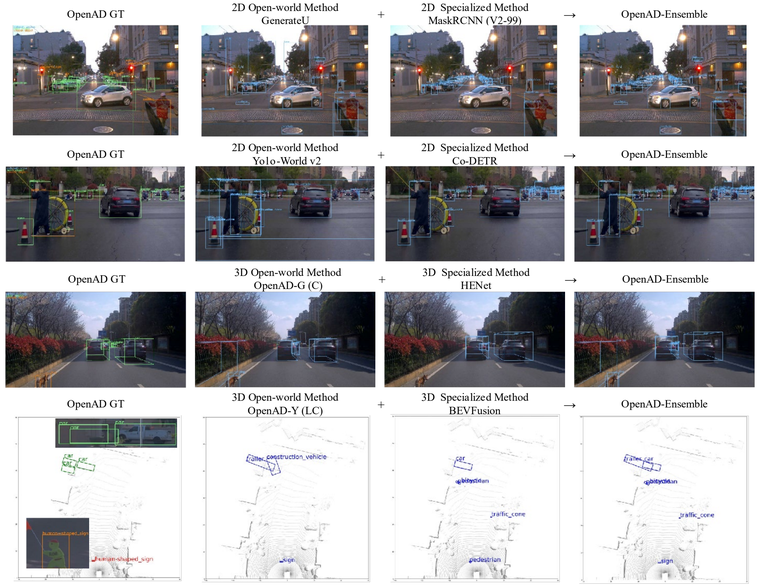


Experiments

Method	Backbone/Base-model	AP↑	AR↑	ATE↓	ASE↓	AR^{nusc}_{seen} ↑	AR^{nusc}_{unseen} ↑	AR^{others}_{seen} ↑	AR^{others}_{unseen} ↑
GLIP [34]	Swin-L	7.14	16.01	6.581	0.1352	1.83	1.28	2.33	1.05
VL-SAM [39]	ViT-H	8.46	17.50	6.630	0.1355	9.66	5.41	9.13	3.43
OWL-ViT v2 [47]	ViT-L	9.70	21.17	6.284	0.1461	21.42	4.66	18.97	**8.01**
GenerateU [15]	Swin-L	9.77	21.75	6.743	0.1360	12.74	7.18	18.79	7.31
YOLO-World v2 [14]	YOLOv8-X	10.20	23.46	7.489	0.1397	18.68	**10.16**	20.61	7.27
GroundingDino [41]	Swin-L	8.52	26.67	6.499	0.1432	20.53	4.21	21.26	7.36
MaskRCNN [24]	ResNet50	12.76	20.07	6.126	0.1359	27.77	0.00	23.41	0.07
MaskRCNN [24]	VovNetv2-99	12.32	21.09	5.746	0.1338	30.21	0.00	21.74	0.09
DETR [10]	ResNet50	12.46	20.35	6.066	0.1346	28.27	0.00	21.35	0.03
DINO [11]	ResNet50	15.24	23.41	5.679	**0.1258**	35.49	0.00	26.39	0.02
Co-DETR [72]	ResNet50	15.65	24.63	5.421	0.1270	38.82	0.00	27.96	0.03
Co-DETR [72]	Swin-L	16.21	27.76	**5.386**	0.1287	45.41	0.00	26.14	0.01
OpenAD-Ens	YOLO-world + MaskRCNN(V2-99)	13.28	29.74	6.726	0.1409	33.30	10.05	26.92	7.20
OpenAD-Ens	YOLO-world + Co-DETR(Swin-L)	**16.94**	**34.38**	6.457	0.1368	**46.65**	10.06	**30.39**	7.20

Evaluation of 2D object detection (middle column, bottom) and 3D object detection methods (right column, top). Each table shows the evaluation results of open-world methods (top), specialized methods (middle), and ensemble methods (bottom) on the OpenAD benchmark. $AR^{nusc}$and AR^{others}demonstrate the in-domain and out-domain capabilities, respectively. AR_{seen} and AR_{unseen} showcase the detection abilities for common objects and rare objects, respectively. General-Specialized Fusion is implemented using NMS.

Method	Modality	Backbone/Base-model	AP↑	AR↑	ATE↓	ASE↓	AR^{nusc}_{seen} ↑	AR^{nusc}_{unseen} ↑	AR^{others}_{seen} ↑	AR^{others}_{unseen} ↑
OpenAD-G	C	GenerateU	6.01	12.90	1.342	0.504	11.35	3.64	15.18	3.71
OpenAD-Y	C	YOLOWorld	6.26	13.89	1.338	0.487	14.64	7.18	18.79	3.53
FnP [18]	L	SECOND	8.85	18.97	0.848	0.493	18.49	10.82	23.42	7.47
OpenAD-G	LC	GenerateU	15.14	34.46	1.056	0.649	14.54	11.15	26.48	**16.95**
OpenAD-Y	LC	YOLOWorld	15.54	36.07	1.063	0.646	29.99	**12.73**	25.88	14.17
BEVDet [27]	C	ResNet50	9.42	15.63	1.183	0.438	36.46	0.00	14.11	0.00
BEVFormer [36]	C	ResNet50	10.08	19.36	1.125	0.440	39.38	0.00	15.85	0.00
BEVFormer [36]	C	ResNet101-DCN	14.43	22.73	0.978	0.444	51.86	0.00	16.59	0.03
BEVDepth4D [26]	C	ResNet50	12.33	20.70	1.118	0.480	39.75	0.00	17.94	0.02
BEVStereo [35]	C	ResNet50	11.12	18.27	1.133	0.431	36.73	0.00	16.21	0.00
BEVStereo [35]	C	VovNetv2-99	10.58	16.03	1.118	0.388	51.69	0.00	13.05	0.01
HENet [60]	C	Vov2-99 + R50	11.58	17.48	1.070	**0.386**	52.02	0.00	14.65	0.01
SparseBEV [40]	C	ResNet50	7.61	16.97	1.142	0.435	60.04	0.00	7.48	0.02
SparseBEV [40]	C	VovNetv2-99	7.64	16.93	1.103	0.431	61.36	0.00	7.09	0.01
BEVFormer v2 [62]	C	ResNet50	14.64	33.13	1.064	0.554	56.63	0.00	27.16	0.08
Centerpoint [67]	L	SECOND	13.79	26.79	0.667	0.499	44.23	0.00	11.42	0.04
TransFusion-L [3]	L	SECOND	14.64	34.02	**0.653**	0.655	52.18	0.00	24.02	0.00
BEVFusion [37]	LC	SECOND + Dual-Swin-T	15.57	33.50	0.730	0.449	59.93	0.00	20.64	0.00
OpenAD-Ens	C	OpenAD-Y + HENet	12.36	24.32	1.176	0.420	54.16	7.18	23.37	3.53
OpenAD-Ens	LC	FnP + BEVFusion	16.19	42.08	0.776	0.458	61.74	10.82	28.40	7.47
OpenAD-Ens	LC	OpenAD-Y + BEVFusion	16.22	47.12	0.851	0.511	62.69	12.05	35.62	13.60
OpenAD-Ens	LC	OpenAD-G + BEVFusion	**16.30**	**48.25**	0.858	0.520	**64.84**	10.59	**39.11**	16.85



Acknowledgements

This work was supported by National Key R&D Program of China (Grant No. 2022ZD0160305).