Find n' Propagate
Open-Vocabulary 3D Object Detection in Urban Environments

Djamahl Etchegaray[1] Zi Huang[1] Tatsuya Harada[2] Yadan Luo [1]

[1]UQMM Lab, University of Queensland, Brisbane, Australia [2]The University of Tokyo, Tokyo, Japan.

Motivation

We explore open-vocabulary 3D object detection using LiDAR data (TOP) and multi-view imagery (BOTTOM), designing four baselines: (1) Top-down Projection, (2) Top-down Self-train, (3) Top-down Clustering, and (4) Bottom-up Weakly-supervised detection, as shown in Fig. 1.

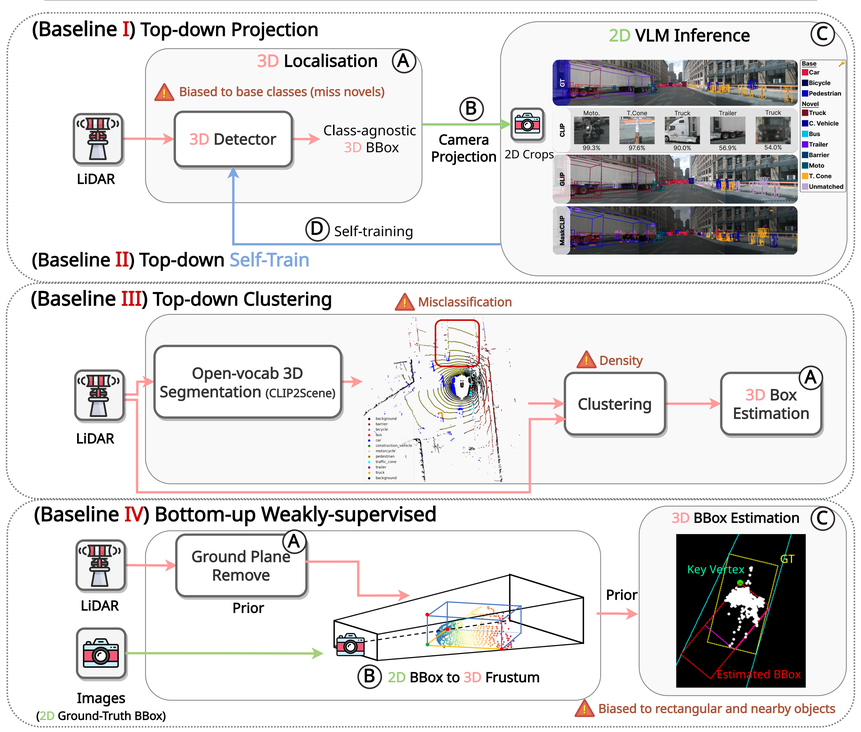
Top-down Approaches:

- Baseline I - Class-agnostic bounding boxes using vision-language models (VLMs):
  - ✓ Leverages 2D open-vocabulary learning.
  - ✗ Overfits to known classes, missing novel objects.
  - ✗ Occlusion leads to misclassification.
- Baseline II - Self-train:
  - ✓ Improves open-vocabulary detection via self-training.
- Baseline III - Clustering:
  - ✓ Uses 3D segmentation for bounding box estimation.

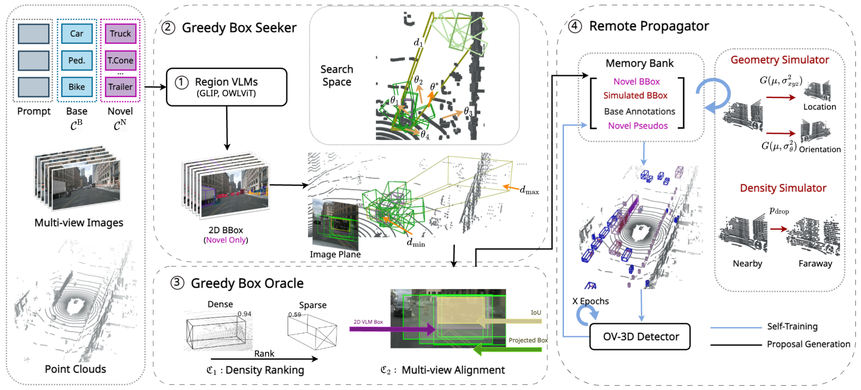
Bottom-up Approach:

- Training-free:
  - ✓ Lifts 2D annotations into 3D.
  - ✓ No base annotations needed, generalizes well to diverse objects.
- Baseline IV, FGR[2]:
  - ✓ Removes background points (e.g., ground, other objects).
  - ✗ Assumes rectangular objects.
  - ✗ Parameter tuning required for each class.
  - ✗ Difficult to segment singular objects from frustum points.

Baselines



Framework



Our Approach

Goal: Utilize pre-trained VLMs to discover novel objects and lift them to 3D.

1. Utilize predictions from region VLMs to discover novel objects near the camera, using our **Greedy Box Seeker**.
2. Feed newly determined boxes into the **Greedy Box Oracle** module to filter out low-quality detections.
3. For distant or occluded novel objects, utilise predicted 3D proposals through a **Remote Propagator**, replicating geometry of remote objects.
4. Incrementally improve detection of novel instances during self-training with a coupled memory bank.

Figure 1. The **Greedy Box Oracle** utilises density and multi-view alignment to rank the incoming proposals and filter the best for each 2D bounding box. Green boxes show the top-5 proposals (by these criteria), yellow is the frustum and pink are the ground-truth annotations.

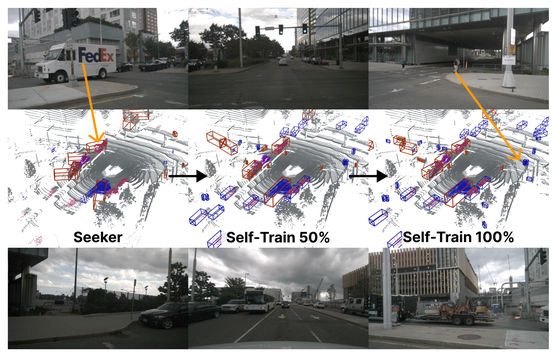


Figure 2. Visualization of our method's iterative ability to recover missed unknowns during self-training. Base and Novel bounding boxes are highlighted for comparison.

Experimental Results

Table 1. **Open-vocabulary evaluation on the nuScenes dataset with four novel classes (§ Setting 1).** -f denotes the variants with logit fusion. All scores are given in percentage. "C2S" denotes CLIP2Scene, "Trans" as Transfusion, and "Center" as Centerpoint.

Method	Box_{3D}	VLM	Arch.	Overall mAP	NDS	AP_B	AP_N
Top-Projection	Base	-	Trans.	34.97	40.02	**58.28**	0.00
		-	Center.	32.84	38.75	54.74	0.00
		CLIP-f	Trans.	31.70	43.24	50.05	4.17
		CLIP-f	Center.	23.78	42.98	39.56	0.13
		Mask-f	Trans.	23.96	42.77	39.07	1.29
		Mask-f	Center.	23.78	42.98	39.55	0.12
		GLIP-f	Trans.	30.64	46.40	46.56	6.77
		GLIP-f	Center.	28.95	46.07	46.55	2.55
(Upper Bound)	GT_{3D}	CLIP	Trans.	45.59	23.48	37.61	-
Top-Clustering	HDBScan	C2S[1]	-	0.70	6.10	0.67	0.76
	DBScan	C2S[1]	-	0.57	3.33	0.51	0.65
Find n' Propagate	Seeker+ST	OWL	Center.	35.06	40.85	48.94	14.25
	Seeker+ST	GLIP	Center.	37.38	39.82	49.54	19.15
	Seeker+ST	OWL	Trans.	42.52	45.13	53.09	26.66
	Seeker+ST	GLIP	Trans.	**44.95**	**47.87**	52.48	**33.65**

Table 2. **Comparisons under the challenging § Setting 2.**

Method	VLM	Arch.	mAP	NDS	AP_B	AP_N
Base	-	Trans.	21.85	24.42	72.85	0.00
Base	-	Center.	21.67	24.40	72.23	0.00
Top-proj.	CLIP-f	Trans.	21.86	24.42	**72.86**	0.00
Top-proj.	CLIP-f	Center.	18.13	33.07	57.37	1.31
Top-proj.	GLIP-f	Trans.	21.66	30.49	71.36	0.36
Top-proj.	GLIP-f	Center.	22.71	33.67	69.78	2.54
Top-proj.	Mask-f	Trans.	23.78	**42.98**	62.46	2.23
Top-proj.	Mask-f	Center.	19.28	33.17	60.00	1.83
Find n' Propagate	GLIP	Trans.	31.44	34.53	67.41	**16.03**
Find n' Propagate	GLIP	Center.	**37.38**	40.28	49.99	18.46

Table 3. **Evaluation with 10 novel classes (§ Setting 3).**

Method	Box_{3D}	VLM	Arch.	Overall AP_N	NDS
Top-Clustering	DBScan	GLIP	-	0.68	4.63
	DBScan	GT_{2D}	-	1.52	7.09
OpenSight [3]		OWL Detic	-	5.40 5.80	12.40 12.70
Ours	Seeker	OWL	-	10.65	18.30
	Seeker	GLIP	-	**16.72**	**22.40**
(Upper Bound)	Seeker	GT_{2D}	-	23.10	22.83

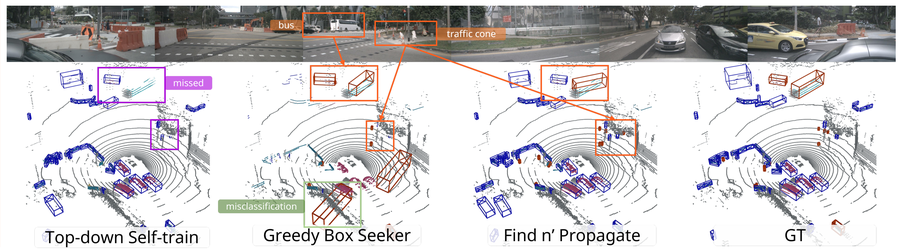


Figure 3. Visualisation of open-vocabulary 3D detection results (§ Setting 1).

References

[1] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7020–7030, 2023.

[2] Yi Wei, Shang Su, Jiwen Lu, and Jie Zhou. FGR: frustum-aware geometric reasoning for weakly supervised 3d vehicle detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4348–4354. IEEE, 2021.

[3] Hu Zhang, Jianhua Xu, Tao Tang, Haiyang Sun, Xin Yu, Zi Huang, and Kaicheng Yu. Opensight: A simple open-vocabulary framework for lidar-based object detection. *CoRR*, abs/2312.08876, 2023.