ImOV3D: Learning Open Vocabulary Point Clouds 3D Object Detection from Only 2D Images

Timing Yang[1,2*] Yuanliang Ju[1,2*] Li Yi[1,2,3†]

Shanghai Qi Zhi Institute[1], IIIS Tsinghua University[2], Shanghai AI Lab[3]

上海期智研究院 SHANGHAI QI ZHI INSTITUTE

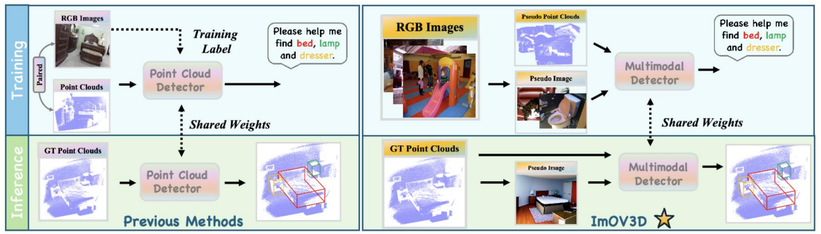
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

清華大學 交叉信息研究院 Institute for Interdisciplinary Information Sciences, Tsinghua University

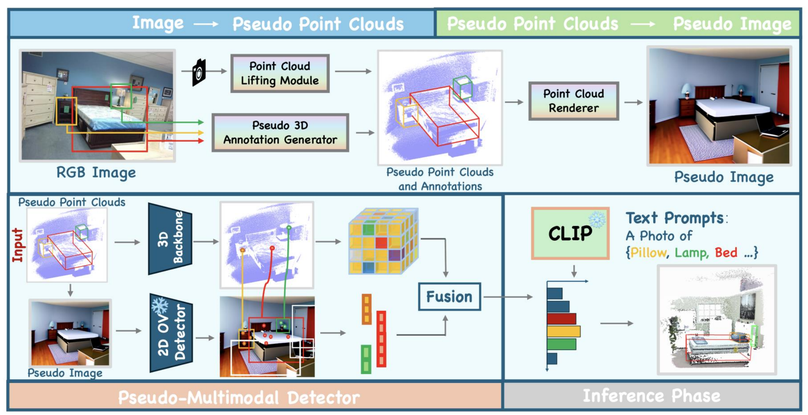
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation:

- How to address the bottleneck of limited 3D data in Open Vocabulary 3D Object Detection?
- How to minimize the domain gap to better transfer 2D knowledge to 3D, thereby enhancing the performance of Open Vocabulary 3D Object Detection?



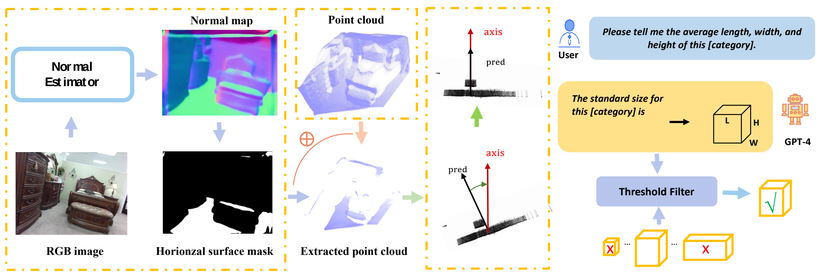
Method:



Loss Function

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{loc}} + \sum_i W_i \times \text{CrossEntropy}(\text{Cls-header}(\mathcal{F}_i) \cdot \mathcal{F}_{\text{text}})$$

Core Modules:



Main Results:

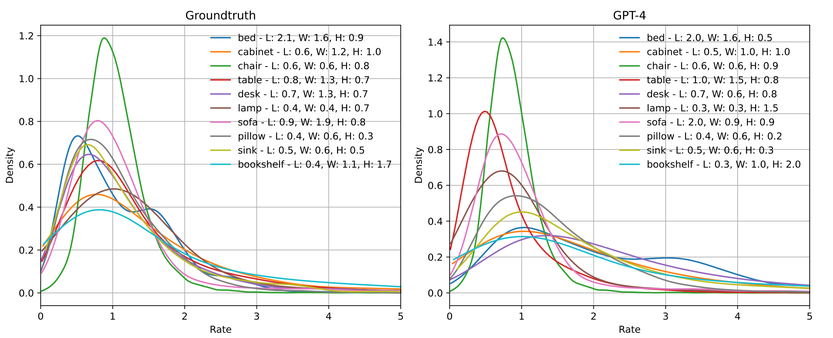
Stage	Data Type	Method	Input	Training Strategy	SUNRGBD mAP@0.25	ScanNet mAP@0.25
Pre-training	Pseudo Data	OV-VoteNet [38]	Point Cloud	One-Stage	5.18	5.86
		OV-3DETR [34]	Point Cloud	One-Stage	5.24	5.30
		OV-3DET [30]	Point Cloud + Image	Two-Stage	5.47	5.69
		Ours	**Point Cloud**	**One-Stage**	**12.61** ↑ **7.14**	**12.64** ↑ **6.78**

Stage	Method	Input	Training Strategy	SUNRGBD mAP@0.25	ScanNet mAP@0.25
Adap-tation	OV-3DET [30]	Point Cloud + Image	Two-Stage	20.46	18.02
	CoDA [5]	Point Cloud	One-Stage	—	19.32
	Ours	**Point Cloud**	**One-Stage**	**22.53**↑ **2.07**	**21.45**↑ **2.13**

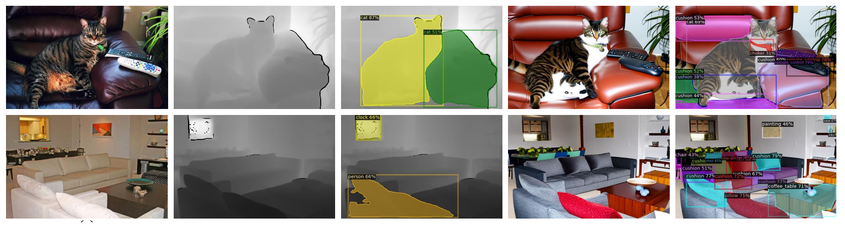
Ablation Study:

3D Data Revision

Stage	Train Phase Prior Size	Rotation Correction	Inference Phase Semantic Size	SUNRGBD mAP@0.25	ScanNet mAP@0.25
Pre-training	✗	✗	✗	8.35	8.33
	✓	✗	✗	10.00	9.60
	✗	✓	✗	9.65	10.29
	✓	✓	✗	11.33	11.64
	✓	✓	✓	**12.61**	**12.64**



Depth Map *VS* Pseudo Images

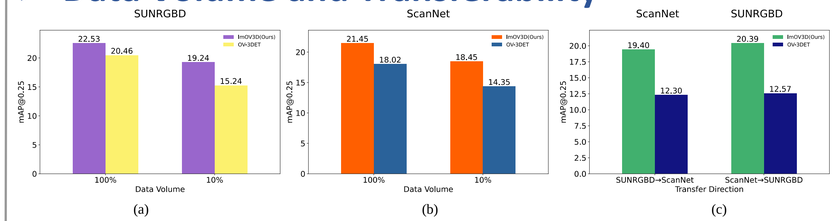


Stage	Rendered Images Data Types	SUNRGBD mAP@0.25	ScanNet mAP@0.25
Pre-training	Depth Map	4.38	4.47
	Pseudo Images	**12.61**	**12.64**

ImOV3D Website

ImOV3D Code

Data Volume and Transferability



Fine-tuned 2D Detector

Pretraining	Adaptation	SUNRGBD mAP@0.25	ScanNet mAP@0.25
-	2D Off-the-shelf + 3D Adaptation	18.8	18.96
Off-the-shelf + 3D Pretraining	2D Off-the-shelf + 3D Adaptation	19.67	19.25
2D Pretraining + 3D Pretraining	2D Adaptation + 3D Adaptation	**22.53**	**21.45**

Visualization: