NEURAL INFORMATION PROCESSING SYSTEMS

Towards 3D Objectness Learning in an Open World

Taichi Liu[1], Zhenyu Wang[2], Ruofeng Liu[3], Guang Wang[4], Desheng Zhang[1]

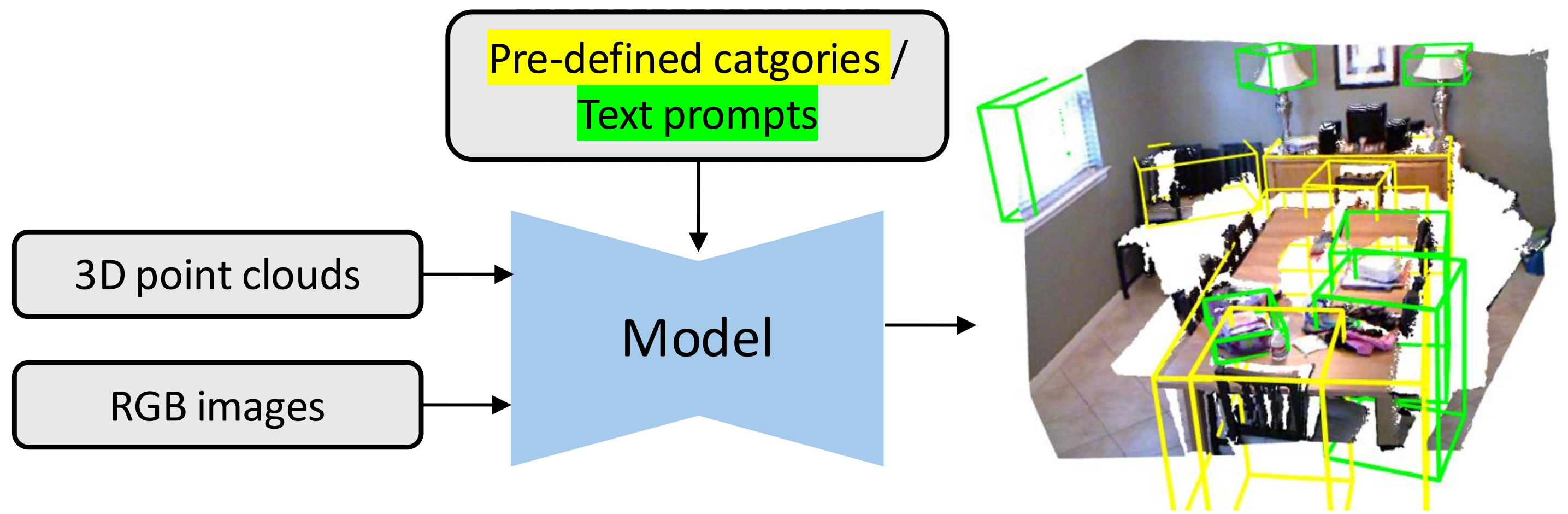
[1]Rutgers University, [2]Tsinghua University, [3]Michigan State University, [4]Florida State University

RUTGERS   MICHIGAN STATE UNIVERSITY   Tsinghua University   FLORIDA STATE UNIVERSITY

Goal

- Detect ***all*** objects in a 3D scene, including both seen objects and novel objects unseen during training.

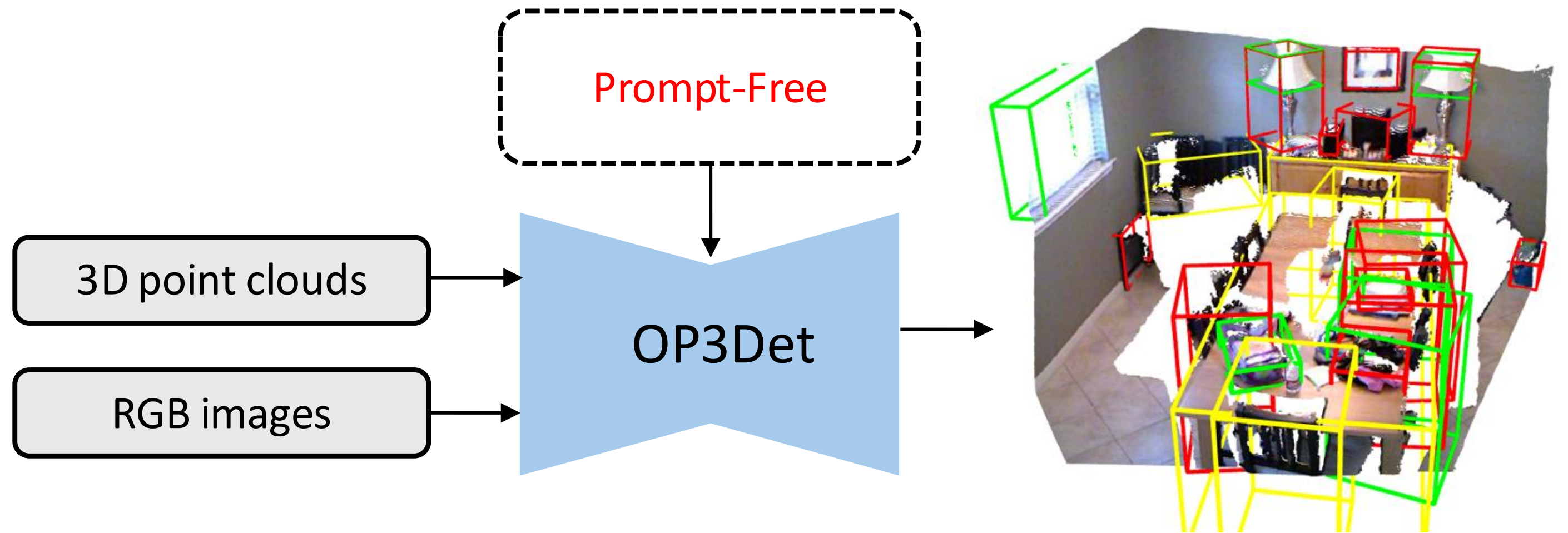
Problem & Motivation



Closed-set / Open-vocab

- Closed-set detection: fails to detect unseen objects.
- Open-vocab detection: unable to define all objects via text prompts.
- **Class-agnostic 3D object detection** - objects are identified and localized based on their intrinsic properties rather than pre-defined semantic labels.

Contribution

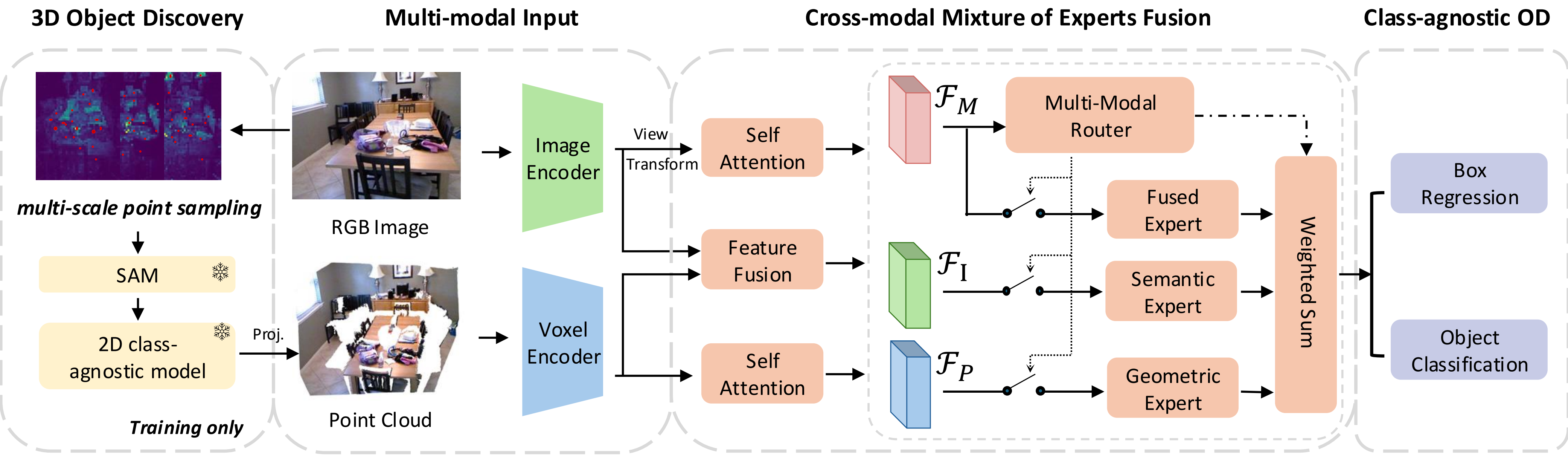


- We introduce the task of **class-agnostic** open-world 3D object detection.
- We propose OP3Det, **a multi-modal 3D detector** for learning **open-world 3D objectness.**
- We provide insights into the benefits of a class-agnostic approach, highlighting its **strong generalization ability** across various downstream tasks.

Problem

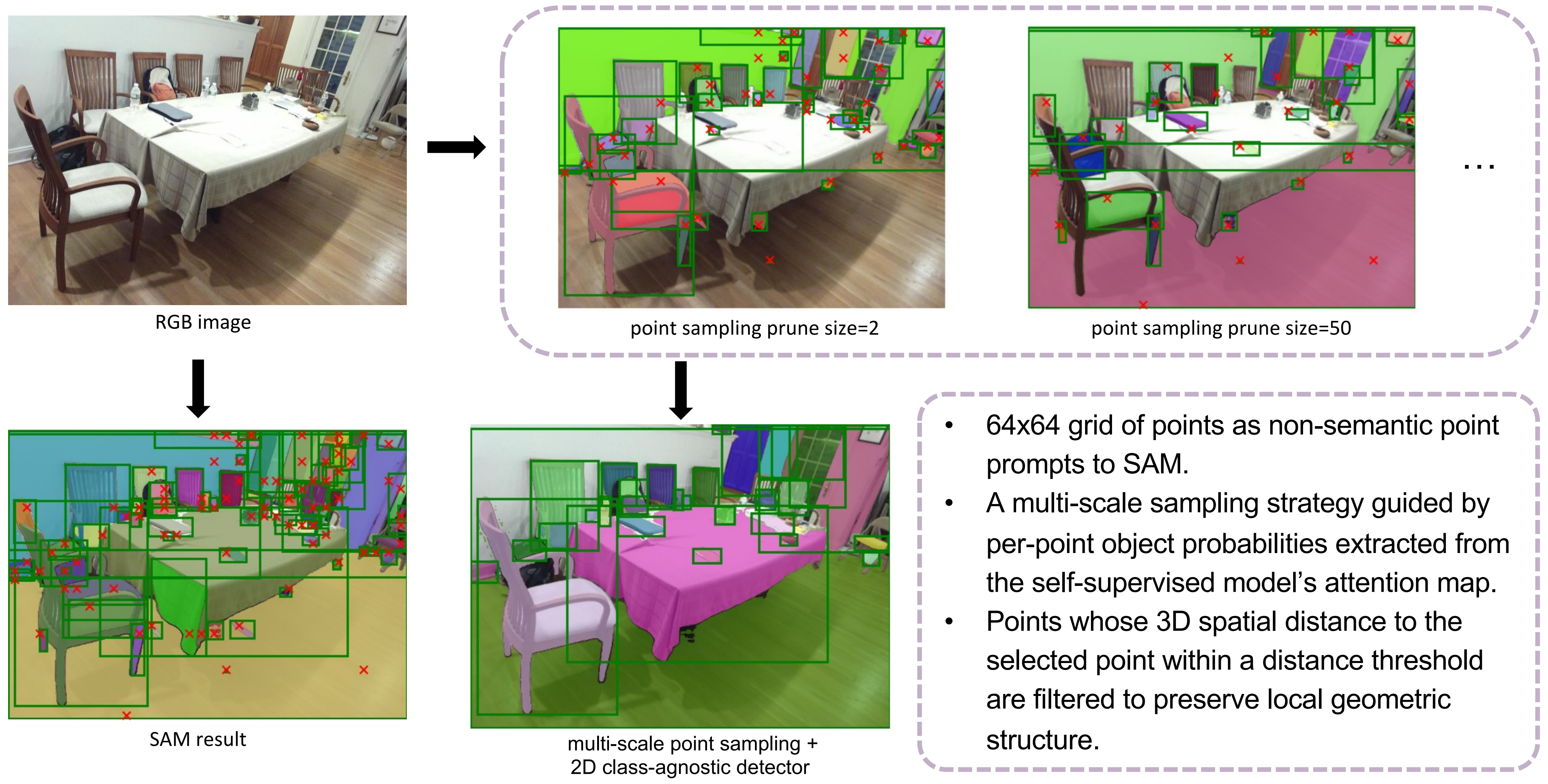
- 3D point cloud data are extremely **limited in both the scale of data and annotated categories**, simply shifting from class-specific to class-agnostic 3D classification is ineffective.
- Open-vocabulary 3D models for class-agnostic detection faces significant challenges due to **vocabulary expansion** and **semantic overlap** in hand-crafted text prompts.

Methodology



Overview: OP3Det mainly consists of two modules: **3D Object Discovery** and **Cross-modal Mixture of Experts Fusion**.

- **3D Object Discovery**: Leverage 2D and 3D cross-priors enables the discovery of novel objects prior to training.
- SAM offers strong zero-shot segmentation for generating 2D masks. However, its outputs are often fragmented, introducing annotation noise that degrades 3D objectness learning.


RGB image
point sampling prune size=2
point sampling prune size=50
…
SAM result
multi-scale point sampling + 2D class-agnostic detector

- 64x64 grid of points as non-semantic point prompts to SAM.
- A multi-scale sampling strategy guided by per-point object probabilities extracted from the self-supervised model's attention map.
- Points whose 3D spatial distance to the selected point within a distance threshold are filtered to preserve local geometric structure.

- **Multi-Modal Fusion Study**: Investigate simple fusion strategies (addition / concatenation), and both lead to performance drops.

PC	Img	method	AR_{novel}	AR_{all}	AR_{base}
✓		-	69.2	87.9	92.5
	✓	-	38.4	64.4	72.5
✓	✓	addition	65.4	85.6	91.4
✓	✓	concatenation	66.0	85.8	92.1
✓	✓	CM-MoE	**78.8**	**89.7**	**93.1**

- **Cross-modal Mixture of Experts Fusion**: selectively route 2D semantic, 3D geometric, and fused multi-modal features to enable dynamic uni-modal and multi-modal fusion, enhancing 3D objectness learning in an open world.

Experiments

- **Cross-category (Class-agnostic 3D object detection)**

Indoor - SUN RGB-D and ScanNet dataset

Method	SUN RGB-D				ScanNet			
	AR_{novel}	AR_{all}	AR_{base}	AP_{all}	AR_{novel}	AR_{all}	AR_{base}	AP_{all}
closed-world 3D object detection methods								
VoteNet [16]	33.7	68.3	79.1	55.1	35.3	44.6	56.1	13.8
GroupFree [60]	41.8	69.9	78.7	49.2	32.1	40.9	51.8	9.4
FCAF3D [17]	65.3	86.5	92.7	62.0	61.7	71.3	83.2	24.7
Uni3DETR [15]	51.8	82.1	91.6	61.3	54.6	67.6	80.1	16.9
Tr3D [56]	62.1	84.8	91.9	53.4	47.1	58.1	71.6	17.2
open-vocabulary 3D object detection methods								
Det-PointCLIPv2 [8]	22.4	31.1	64.5	10.2	33.1	38.7	55.9	3.1
3D-CLIP [26]	23.6	32.3	66.8	25.7	32.9	36.2	55.5	5.6
CoDA [7]	33.9	60.2	71.5	48.2	44.3	53.4	68.3	23.9
OV-Uni3DETR [9]	62.8	82.5	88.8	57.4	67.6	71.6	76.5	25.9
ImOV3D [61]	46.9	63.1	74.1	28.3	56.9	70.6	77.9	25.0
class-agnostic open-world 3D object detection method								
OP3Det (ours)	**78.8**	**89.7**	**93.1**	**65.4**	**79.9**	**83.2**	**87.3**	**28.6**

Outdoor - KITTI dataset

Method	AP_{3D}			AP_{BEV}		
	easy	medium*	hard	easy	medium	hard
closed-world 3D object detection methods						
SECOND [69]	61.05	62.36	61.36	63.15	69.00	68.46
PointPillar [70]	59.54	62.13	60.04	63.04	68.87	66.75
Part-A^2 [71]	61.28	63.43	63.57	62.93	69.04	69.88
3DSDD [72]	61.42	62.34	62.06	62.93	68.50	68.29
PV-RCNN [73]	59.88	65.18	65.67	63.01	69.36	70.42
Uni3DETR [15]	63.54	65.74	65.43	62.74	69.01	69.87
open-vocabulary 3D object detection method						
OV-Uni3DETR [9]	62.66	63.20	62.82	64.33	69.15	68.98
class-agnostic open-world 3D object detection method						
OP3Det (ours)	**63.56**	**66.75**	**66.42**	**65.13**	**71.37**	**70.34**

- **Cross-dataset (Class-agnostic 3D object detection)**

Method	ScanNet → SUN RGB-D				SUN RGB-D → ScanNet			
	AR_{25}	AR_{50}	AP_{25}	AP_{50}	AR_{25}	AR_{50}	AP_{25}	AP_{50}
closed-world 3D object detection methods								
VoteNet [16]	34.8	2.0	10.8	0.1	30.4	6.3	9.6	1.2
GroupFree3D [60]	41.4	0.4	1.9	0.1	39.4	5.2	8.7	0.1
FCAF3D [17]	59.3	8.1	17.9	0.6	47.7	14.6	12.9	1.9
Uni3DETR [15]	51.3	6.4	11.9	0.2	45.7	10.9	11.3	1.3
Tr3D [56]	54.6	4.5	11.4	0.2	45.2	10.7	9.4	1.6
open-vocabulary 3D object detection methods								
CoDA [7]	21.4	2.8	6.2	0.1	32.7	5.2	8.9	0.4
OV-Uni3DETR [9]	49.5	3.2	8.1	0.3	52.0	15.4	9.5	0.8
class-agnostic open-world 3D object detection method								
OP3Det (ours)	**73.1**	**10.7**	**22.3**	**1.1**	**77.9**	**37.3**	**21.2**	**5.1**

- **Cross-category (Class-sepcific 3D object detection)**

Method	SUN RGB-D			ScanNet		
	AP_{novel}	AP_{base}	AP_{all}	AP_{novel}	AP_{base}	AP_{all}
CoDA [7]	6.71	38.72	13.66	6.54	21.57	9.04
INHA [74]	8.91	42.17	16.18	7.79	25.1	10.68
CoDAv2 [75]	9.17	42.04	16.31	9.12	23.35	11.49
OV-Uni3DETR [9]	12.96	49.25	20.85	15.21	31.86	17.99
GLRD [76]	12.96	49.40	20.88	17.29	26.78	18.87
OP3Det (ours)	**14.31**	**49.63**	**21.99**	**17.77**	**32.12**	**20.16**

- **Visulazation**