OpenMask3D: Open-Vocabulary 3D Instance Segmentation - Rebuttal Tables and Figures

	Novel Classes			*Base Classes*			*All Classes*			
Method	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	tail (AP)
OpenScene [46] (2D Fusion)	7.6	10.3	12.3	11.1	**15.0**	17.7	8.5	11.6	13.8	6.1
OpenScene [46] (3D Distill)	1.8	2.3	2.7	10.1	13.4	15.4	4.1	5.3	6.1	0.4
OpenScene [46] (2D/3D Ensemble)	2.4	2.8	3.3	10.4	13.7	16.3	4.6	5.8	6.8	0.9
OpenMask3D (Ours)	**10.4**	**12.9**	**15.3**	**12.1**	**15.0**	**17.9**	**10.9**	**13.5**	**16.0**	**10.0**

Table 1: 3D instance segmentation results using masks from mask module trained on ScanNet20 annotations, evaluated on the ScanNet200 dataset [51]. We identify 53 classes (such as chair, folded chair, table, dining table ...) that are semantically close to the original ScanNet20 classes, and group them as "Base". Remaining 147 classes are grouped as "Novel". We also report results on the full set of labels, titled "All".

Model	AP	AP_{50}	AP_{25}
Open-vocabulary			
OpenScene [46] (2D Fusion)	10.9	15.6	17.3
OpenScene [46] (3D Distill)	8.2	10.5	12.6
OpenScene [46] (2D/3D Ensemble)	8.2	10.4	13.3
OpenMask3D (rendered RGB-D)	11.6	14.9	18.4
OpenMask3D (fast config.)	11.9	17.1	23.3
OpenMask3D (base config.)	**13.1**	**18.4**	**24.2**

Table 2: 3D instance segmentation results on the **Replica dataset.**

Levels	Ratio of Exp.	AP	AP_{50}	AP_{25}
1	0.1	11.3	16.0	20.2
3	0.1	**13.1**	**18.4**	**24.2**
5	0.1	12.8	17.6	22.6
3	0.05	12.9	18.1	23.5
3	0.1	**13.1**	**18.4**	**24.2**
3	0.2	12.8	17.7	22.9

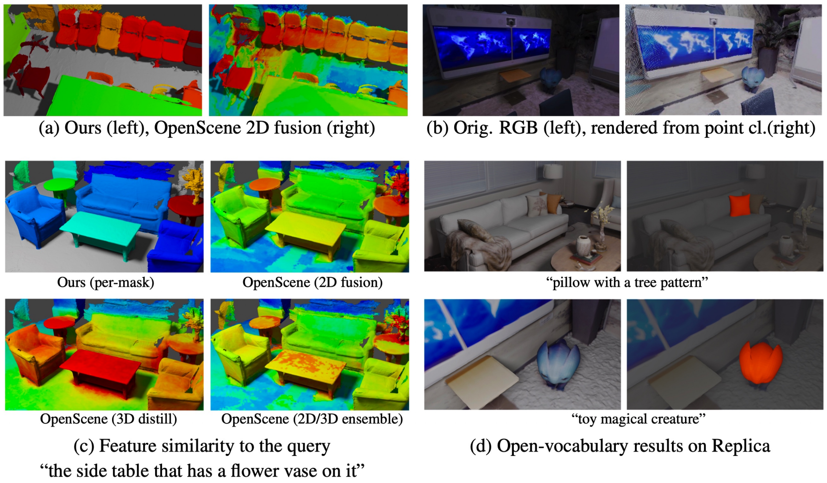
Table 3: Ablation study of the multi-scale cropping hyperparameters on the Replica dataset.

Model	Checkpoints	Memory
OpenMask3D (ours)		
SAM [30]	ViT-H	8 GB
SAM [30]	ViT-B	4 GB
CLIP [49]	ViT-L/14@336px	4 GB
OpenScene [46]		
OpenSeg [15]	from [46] repository	> 30 GB [1]

Table 4: Memory requirements of foundation models used in OpenMask3D and OpenScene [46].

Function	Checkpoints	Time
OpenMask3D (ours)		
SAM.set_image() [30]	ViT-H	0.497
SAM.predict() [30]	ViT-H	0.006
SAM.set_image() [30]	ViT-B	0.109
SAM.predict() [30]	ViT-B	0.005
CLIP.preprocess() [49]	ViT-L/14@336px	0.004
CLIP.encode_image() [49]	ViT-L/14@336px	0.015
OpenScene [46]		
OpenSeg.predict() [15]	from [46] repository	0.917 s

Table 5: Time requirements for atomic operations of foundation models. Values collected as averages during the computation of features for a scene of ScanNet200.


(a) Ours (left), OpenScene 2D fusion (right)

(b) Orig. RGB (left), rendered from point cl.(right)

Ours (per-mask)

OpenScene (2D fusion)

"pillow with a tree pattern"

OpenScene (3D distill)

OpenScene (2D/3D ensemble)

"toy magical creature"

(c) Feature similarity to the query "the side table that has a flower vase on it"

(d) Open-vocabulary results on Replica

Figure 1: Qualitative results (best viewed on a screen)