OpenM3D: Open Vocabulary Multi-view Indoor 3D Object Detection without Human Annotations

Supplementary Material

1. Visualization

This section showcases visualizations of 3D pseudo boxes generated by our method, along with additional qualitative results from OpenM3D.

Visualize 3D Pseudo Boxes. The localization capability of our pseudo boxes has been validated in Table 1, 2 of the main paper. In Fig. 1, we show some examples of our 3D pseudo boxes and their corresponding 3D segmentations. To clearly present our pseudo boxes, we organize them based on two distinct ranges—small and medium—using the volumes of the boxes. Moreover, our 3D pseudo boxes can accurately locate novel objects, as illustrated in Fig. 2, in addition to those annotated in the ground truth. These results validate the localization capability of our generated class-agnostic pseudo boxes for various potential objects in the scene, paving the way for open-vocabulary 3D object detection.

More Qualitative Results. We present more qualitative results of open-vocabulary 3D object detection obtained by OpenM3D in Fig. 3. Some detection results for tail and novel objects are also shown in Fig. 4. With general prompts used in CLIP, OpenM3D demonstrates consistent 3D detections across multiple classes. This strongly showcases OpenM3D's capability in open-vocabulary 3D object detection.

2. Experiment Details

2.1. Implementation Details

Frame Selection for Generating 2D Segments. To achieve fine-grained SAM [3] results for each frame, we aim to automatically choose frames with distinct outlines as SAM inputs for improving segmentation quality. As a result, we utilize Laplacian calculations to determine sharpness as the basis for selecting frames. For every scene in ScanNet200 and ARKitScenes, we divide all frames into intervals based on chronological order. Within each interval, we select the frame with the highest sharpness value. This process was repeated until 300 frames were chosen, iterating through the remaining frames in each round.

2D Segments Filtering and Refinement. When generating 2D segments from an image, we noticed that SAM may generate excessively small segments. Such problematic segments confuse the CLIP image encoder, resulting in poor embedding quality. Multiple small segments may map to the same voxel and further worsen the open-vocabulary classification of OpenM3D. We thus add preprocessing steps to exclude such patches: We set a minimum bounding-box size of 30 pixels, and a 0.02 ratio threshold of observed 3D points within each segment's bounding box.

Given that CLIP is trained using real-world images, our approach involves incorporating the surrounding regions of the bounding box when calculating the CLIP embedding for each 2D segment, to provide a scenario similar to real-world images. In addition to the image patch tightly cropped by the bounding box around each segment, we include patches from areas surrounding the segment with dimensions of 110% and 120% relative to the size of the bounding box. In OV-3DET, Lu *et al*. use a predefined vocabulary with 364 categories for pseudo box generation, we follow the same and use the vocabulary to improve on CLIP segment embeddings. Specifically, for each segment, we compare the embedding of the segment to the text embedding of all categories, and use the embedding of the closest category. Please refer to Alg. 1 for the pseudo code of 3D pseudo box generation.

Voxel and 3D Volume. The feature volume measures 6.4 $\times$ 6.4 $\times$ 2.56 meters, with a voxel size of 0.16 meters in all three dimensions.

2.2. 3D Pseudo Box

3D Pseudo Box on ScanNetv2. The evaluation result for ScanNetv2 [2] is presented in Table 1. Similar to the performance on ScanNet200, our method consistently outperforms OV-3DET [6] and SAM3D [9] in terms of precision at IoU@0.25 and IoU@0.50, while maintaining a comparable recall with SAM3D. This validates the contribution of the graph embedding-based clustering strategy, which simultaneously considers the 2D segmentation results across all frames. This approach helps mitigate the impact of segmentation errors from individual frames.

3D Pseudo Box in Different Subset on ScanNet200. In Table 2, we showcase the detailed 3D pseudo box evaluation in ScanNet200 for different subsets (head, common, tail). The evaluation computed overall precision without considering classes, given our pseudo boxes lack class infor-

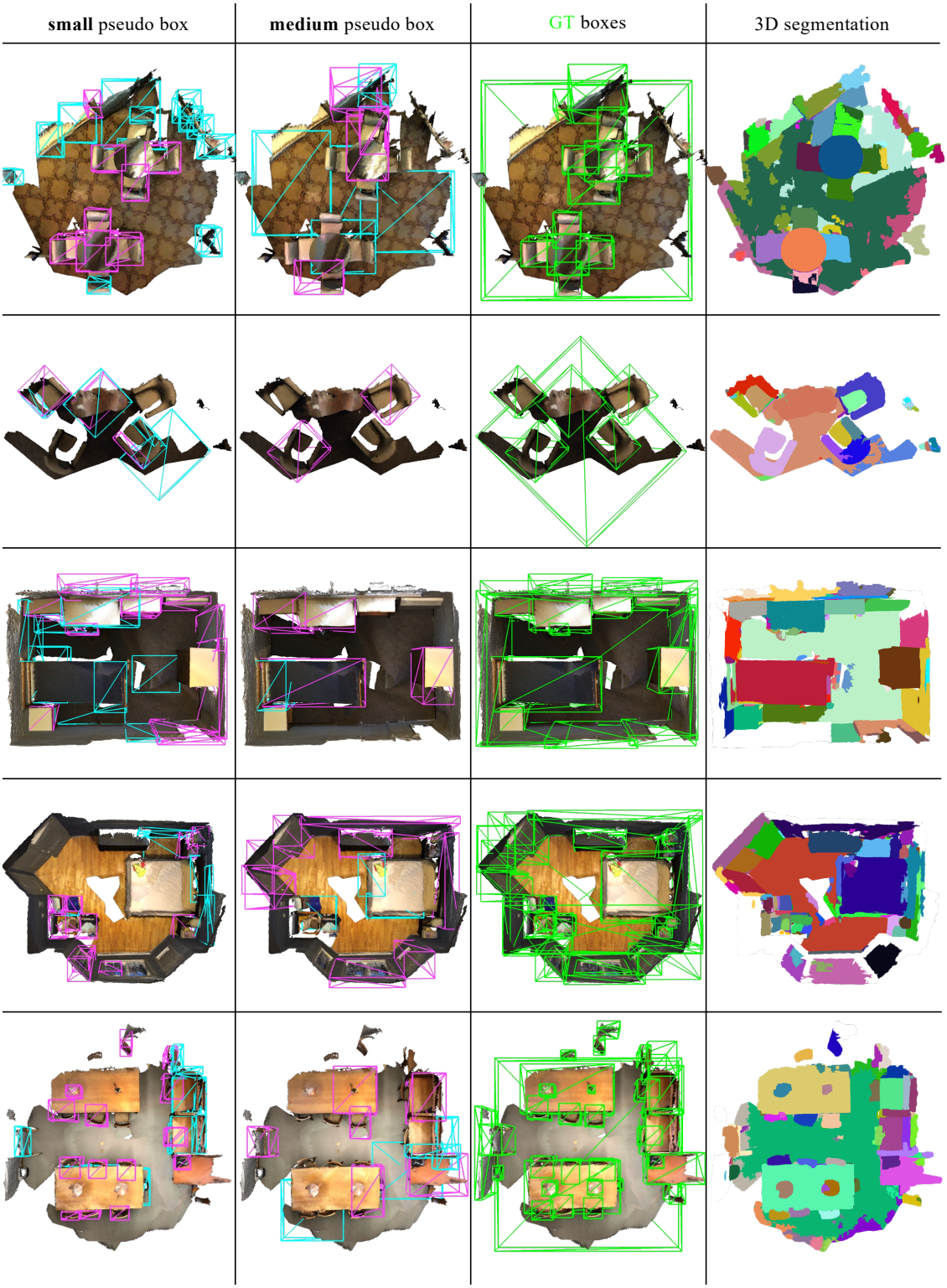

Figure 1. **Visualize Pseudo Boxes of OpenM3D** on ScanNet200. We visualize our 3D pseudo boxes using two different volume sizes (small and medium). In this visualization, cyan represents false positives, while magenta represents true positives matching the GT boxes at IoU@0.25.

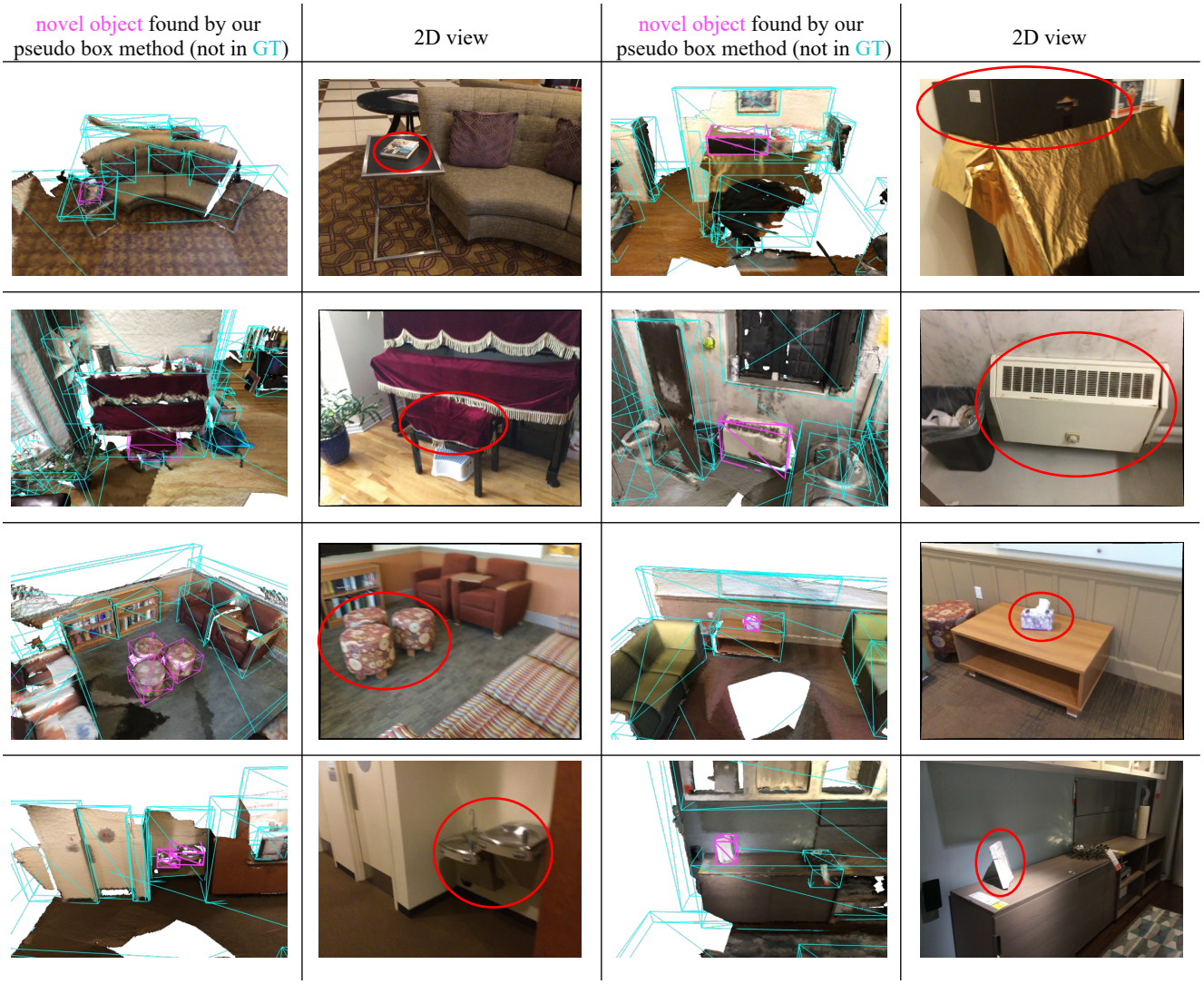

Figure 2. **Localizing Novel Object with Pseudo Box** on ScanNet200.

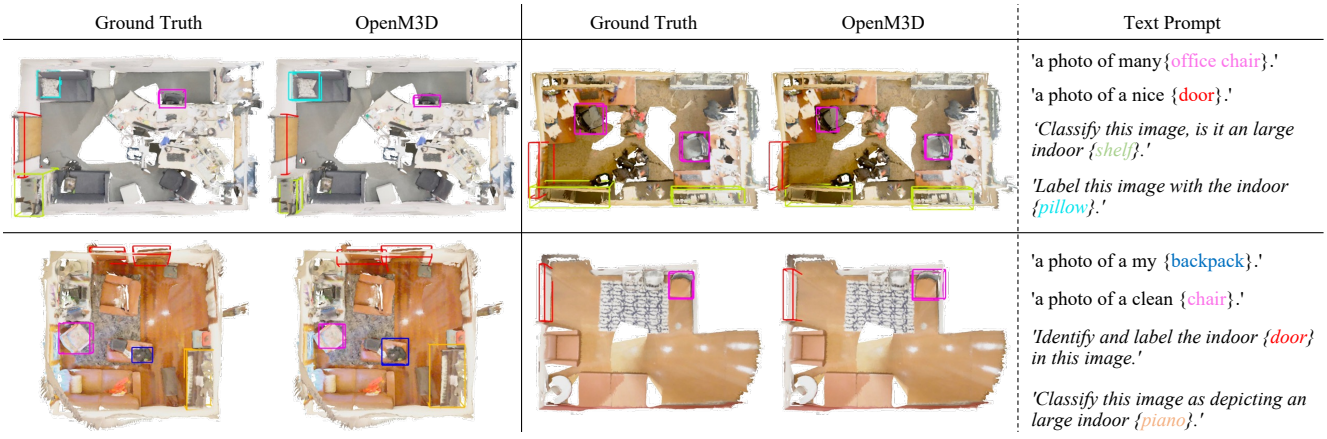

Figure 3. **More Qualitative Results of OpenM3D** on ScanNet200. We show general text prompts used in the ImageNet dataset, as well as prompts from *specific text*.

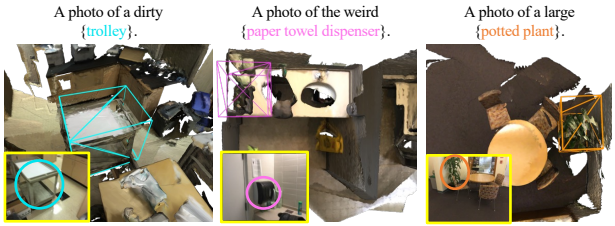

Figure 4. **Novel and tail predictions in OpenM3D**.

Algorithm 1: 3D Pseudo Box Generation

Input : RGB images, corresponding pose $(\mathbf{R}, \mathbf{t})$, intrinsic $\mathbf{K}$, and depth map $\mathbf{D}$
Output : 3D Pseudo Boxes b^{3D}

for *each RGB image* I **do**
  $n_j^{2D} \leftarrow$ Segment2D(I)
  $n_j^{3D} \leftarrow$ Backproject(n_j^{2D}, $(\mathbf{R}, \mathbf{t})$, $\mathbf{K}$, $\mathbf{D}$) ; // Eq.1
end
$Nodes := \{n_j^{3D}\}$
$V \leftarrow$ Voxelize($Nodes$) ; // Voxelize based on 3D coordinates of each node
for *voxel in* V **do**
  for *any pair* (n_j^{3D}, n_k^{3D}) *in voxel* **do**
    $e_{jk} := \text{edge}(n_j^{3D}, n_k^{3D})$; // Eq.2
  end
end
$Edges := \{e_{jk}\}$
$Embedding \leftarrow$ GraphEmbed(GenGraph($Nodes$, $Edges$))
$C \leftarrow$ Clustering($Embedding$) ; // Give each node a clustered group
for $\mathcal{C}_q$ *in* C **do**
  $\hat{n}_q^{3D} := \{n^{3D} \in \mathcal{C}_q\}$; // Collect partial segments in the same cluster q
  $b_q^{3D} :=$ AxisAlignedBox($\hat{n}_q^{3D}$)
end
$b^{3D} := \{b_q^{3D}\}$

mation. While calculating precision in a certain subset, such as "head," only ground truth boxes in head classes are considered. This may result in a lower head precision than the overall precision, as pseudo boxes overlapping with ground truth common/tail classes contribute to false positives in the head precision calculation.

Moreover, we utilize an advanced image segmentation method, CropFormer [4], to acquire more accurate 3D pseudo boxes. CropFormer's improved object-wise understanding reduces the risk of over-segmentation, enhancing the consistency in 2D views. This improvement benefits our 3D pseudo box generation method, resulting in less noisy 3D segments and more precise refinements. Our method prioritizes pseudo box precision over recall for detector training, resulting in higher precision at IoU@0.25 and IoU@0.5 compared to OV-3DET and SAM3D in each subset. This superior quality is evident in our boxes generated based on both SAM and CropFormer. They also achieve significantly better recall than OV-3DET and remain comparable to SAM3D in most settings.

Table 1. **3D Pseudo Box Evaluation** on ScanNetv2. Our 3D pseudo boxes demonstrate higher quality compared to OV-3DET and SAM3D in terms of precision at IoU@0.25 and IoU@0.50.

Method	Precision (%)		Recall (%)	
	@0.25	@0.50	@0.25	@0.50
OV-3DET [6]	4.28	0.20	53.14	25.90
SAM3D [9]	7.39	4.94	70.02	**46.76**
Ours w/o MSR	15.81	7.52	72.62	34.56
Ours	**17.11**	**9.91**	**73.84**	42.80

Table 2. **Detailed 3D Pseudo Box Evaluation with different 2D segmentation** on ScanNet200. We perform a comprehensive evaluation across different subsets of ScanNet200. Additionally, we leverage various 2D segmentation sources to generate pseudo boxes. The use of different 2D segmentation sources in our method results in 3D pseudo boxes of varying quality. For example, when CropFormer is applied, these boxes outperform all other methods in terms of precision and recall at IoU@0.25 and IoU@0.50.

Method	2DSeg	Classes	Precision (%)		Recall (%)	
			@0.25	@0.50	@0.25	@0.50
OV-3DET [6]	Detic [10]	overall	11.62	4.40	21.13	7.99
		head	9.59	3.68	20.39	7.82
		common	1.95	0.74	26.12	9.95
		tail	1.06	0.29	24.22	6.77
SAM3D [9]	SAM [3]	overall	14.48	9.05	57.70	36.07
		head	12.54	7.68	58.44	35.81
		common	1.86	1.33	**56.97**	**40.67**
		tail	0.87	0.59	**44.72**	**30.27**
Ours w/o MSR	SAM [3]	overall	27.09	11.98	52.43	23.18
		head	24.26	10.67	54.90	24.14
		common	2.99	1.40	45.15	21.23
		tail	0.86	0.36	20.65	8.68
Ours	SAM [3]	overall	32.07	18.14	58.30	32.99
		head	28.55	16.00	60.68	34.01
		common	3.66	2.20	51.88	31.68
		tail	1.15	0.69	26.30	22.68
Ours	CropFormer [4]	overall	**35.58**	**22.72**	**62.60**	**39.97**
		head	**31.67**	**19.97**	**65.14**	**41.08**
		common	**3.94**	**2.75**	55.07	38.53
		tail	**1.31**	**0.94**	29.77	21.33

2.3. Baseline Using *3R methods

Recent 3R methods such as MV-Dust3R [7] and VGGT [8] enable 3D scene reconstruction from RGB images and camera poses without requiring depth, aligning well with OpenM3D's inference setting. To establish a baseline, we implemented a multi-stage pipeline that combines VGGT for 3D reconstruction and OVIR-3D [5] for open-vocabulary instance segmentation on ScanNet200. All components were executed using official implementations and default settings, with 3D boxes computed from axis-aligned segment bounds.

This pipeline incurs substantial computational overhead—particularly during 2D-3D fusion—resulting in an inference time of 300 seconds per scene, compared to 0.3 seconds for OpenM3D. In terms of accuracy, it achieved only 5.97% AP@25 (class-agnostic), significantly lower than OpenM3D's 26.92%. We also observed that VGGT often fails to reconstruct fine-grained indoor geometry (see Fig. 5), which is crucial for accurate 2D-3D matching in instance segmentation—a limitation also noted in the OVIR-3D paper.

Overall, this reconstruction-based pipeline is substantially less effective than OpenM3D in both accuracy and efficiency for open-vocabulary 3D object detection.

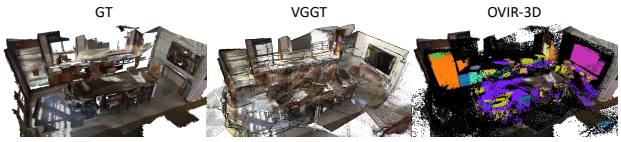

Figure 5. **3R baseline qualitative result.** Comparison between (left) ground-truth ScanNet scene, (middle) VGGT 3D reconstruction using only RGB images and poses, and (right) OVIR-3D segmentation result on the VGGT output. The reconstruction lacks fine-grained indoor geometry, resulting in inaccurate 2D–3D matching and degraded segmentation quality.

2.4. Inference Efficiency

As shown in Table 3. OpenM3D achieves the fastest inference time of 0.3 seconds per scene, using only multi-view RGB images, and significantly outperforms baselines such as OV-3DET (5 s), S2D (2.1 s), and S2D with depth estimation (81 s). Unlike others, it avoids costly CLIP inference and depth prediction, making it highly suitable for real-time 3D detection.

Table 3. **Inference time comparison** on ScanNet200 on a V100 GPU. OpenM3D is over 16× faster than OV-3DET and 270× faster than the depth-estimated S2D baseline.

Method	**OV-3DET**	**S2D**	**S2D Depth Est.**	**Ours**
Inference time (s)	5	2.1	81	**0.3**

Table 4. **Results of OpenM3D trained with different CLIP encoders** on ScanNet200.

CLIP Encoder	mAP@25 (%)	mAR@25 (%)
ViT-L/14	4.23	15.12
ViT-B/16	4.16	15.50
ViT-B/32	4.02	14.74

2.5. Transferability of Pretrained Model

OpenM3D does not rely on predefined 'seen' categories or 3D annotations during training, making it naturally OV - all categories are essentially novel. OpenM3D demonstrates strong performance across head, common, and tail classes in ScanNet200 (see Fig. 4), highlighting its ability to handle rare or unseen classes.

2.6. Ablation Study

CLIP Visual Encoders. We aligned our voxel feature to the pre-trained CLIP feature extracted by the ViT-L/14 image encoder during training. Furthermore, we showcase alternative results employing various other CLIP image encoders in this section. As outlined in Table 4, the use of different CLIP image encoders exhibited negligible impact on both evaluation metrics, namely mAP@25 and mAR@25. This observation emphasizes the robust open-vocabulary classification capability of our method, OpenM3D.

3D Detection with Pseudo Box using CropFormer. When deploying better segmentation models, e.g., CropFormer [4], we can generate more accurate pseudo boxes as detailed in Table 2. The improvement on 2D segmentation benefits our 3D pseudo box generation method on 3D segments refinements. Trained with these boxes, OpenM3D demonstrates a notable improvement of 12.5% in mAP@25, rising from 4.23% to 4.76% on ScanNet200, as shown in Table 5. This highlights the potential of our 3D pseudo boxes on better 2D segmentation. Note that in ARKitScenes, given the sparse point cloud, improving 2D segmentation using CropFormer alone has not significantly improved 3D box metric performance.

3D Detection on ScanNetv2. We reported the results of our model evaluated on the common 18 classes in ScanNetv2 [2] in Table 6. OpenM3D trained with our pseudo boxes consistently outperforms the models trained with SAM3D and OV-3DET on all metrics, including AP@25, AP@50, AR@25, and AR@50. Notably, OpenM3D achieved over 12% and 20% improvements in AP@25 and AR@50, respectively, compared to OV-3DET. Larger gaps were observed, with 7.34% vs. 2.87% in AP@50 and 20.94% vs. 9.27% in AR@50. The substantial improvements brought by our method on AP@50 and AR@50 underscore the limitations associated with solely relying on single-view depth maps and images for bounding box generation. The notable improvements of our pseudo boxes over SAM3D in AP@25

Table 5. **3D Object Detection** on ScanNet200. Our pseudo boxes with CropFormer improve upon SAM.

Method	Trained Box 2DSeg	mAP@25(%)	mAR@25(%)
OV-3DET [6]	Detic [10]	3.13	10.83
SAM3D [9]	SAM [3]	3.92	13.33
Ours	SAM [3]	4.23	**15.12**
	CropFormer [4]	**4.76**	14.62

Table 6. **3D Object Detection** on ScanNetv2. OpenM3D outperforms our method trained on the boxes from OV-3DET and SAM3D.

Method	Trained Box	AP@25 (%)	AR@25 (%)	AP@50 (%)	AR@50 (%)
OpenM3D	OV-3DET [6]	17.65	40.37	2.87	9.27
	SAM3D [9]	16.69	49.39	5.18	19.33
	Ours	**19.76**	**50.40**	**7.34**	**20.94**

and AP@50 metrics showcase the efficacy of our graph embedding-based pseudo boxes. Note that the train/evaluate split applied in [1, 6] differs from the official split by ScanNetv2 [2], making direct comparisons with their reported results challenging.

3. Limitation

The gap between class-agnostic and OV 3D detection implies that the pre-trained CLIP feature can be improved in classifying many semantically similar household objects. We leave this as a future direction.

References

[1] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 6

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 5, 6

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 4, 6

[4] Qi Lu, Jason Kuen, Shen Tiancheng, Gu Jiuxiang, Guo Weidong, Jia Jiaya, Lin Zhe, and Yang Ming-Hsuan. High-quality entity segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 4, 5, 6

[5] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *7th Annual Conference on Robot Learning*, 2023. 5

[6] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 4, 6

[7] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *arXiv preprint arXiv:2412.06974*, 2024. 5

[8] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 5

[9] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 1, 4, 6

[10] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision (ECCV)*, 2022. 4, 6