2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

OV-MAP : Open-Vocabulary Zero-Shot 3D Instance Segmentation Map for Robots

Juno Kim[1*] Yesol Park[1*] Hye-Jung Yoon[1*] Byoung-Tak Zhang[1,2,3]

***Abstract*— We introduce OV-MAP, a novel approach to open-world 3D mapping for mobile robots by integrating open-features into 3D maps to enhance object recognition capabilities. A significant challenge arises when overlapping features from adjacent voxels reduce instance-level precision, as features spill over voxel boundaries, blending neighboring regions together. Our method overcomes this by employing a class-agnostic segmentation model to project 2D masks into 3D space, combined with a supplemented depth image created by merging raw and synthetic depth from point clouds. This approach, along with a 3D mask voting mechanism, enables accurate zero-shot 3D instance segmentation without relying on 3D supervised segmentation models. We assess the effectiveness of our method through comprehensive experiments on public datasets such as ScanNet200 and Replica, demonstrating superior zero-shot performance, robustness, and adaptability across diverse environments. Additionally, we conducted real-world experiments to demonstrate our method's adaptability and robustness when applied to diverse real-world environments.**

I. INTRODUCTION

In recent years, mobile robots have increasingly required robust 3D mapping capabilities to operate in complex, unstructured environments. Open-vocabulary 3D mapping, where the system can recognize and segment objects without being explicitly trained on specific object categories, plays a crucial role in enhancing these capabilities. The challenge of creating zero-shot 3D scene mappings from an open-world perspective arises from the complexity of 3D environments and the scarcity of large-scale open-world 3D datasets. Current approaches often rely on 2D open-vocabulary models, such as CLIP [1], to project features from RGB images into 3D space [2]–[5]. By embedding CLIP-space features into the 3D space's voxels, these methods aim to construct 3D maps that generalize to diverse environments.

However, many of these per-voxel mapping methods suffer from the problem where neighboring voxels share overly similar features, causing boundaries between instances to blur and reducing the precision of instance-level segmentation. This problem is especially pronounced when queried with a sentence or complex input, leading to poor object segmentation quality and unreliable mappings. Consequently, these approaches struggle to achieve the necessary instance-level precision required for accurate open-vocabulary 3D mapping.

Fig. 1. **Illustration of Proposed System.** A mobile robot equipped with OV-MAP accurately identifies objects on a per-instance according to input queries.

To address these limitations, we propose OV-MAP (Open-vocabulary zero-shot 3D instance segmentation map), a novel zero-shot open-vocabulary 3D mapping method that operates from a per-instance perspective. This approach is inspired by advancements in zero-shot 3D segmentation [6] and open-vocabulary 3D instance labeling [5]. OV-MAP confines CLIP features to individual 3D instances by developing a zero-shot 3D instance segmented map through a 3D mask voting process. The process begins by projecting 2D masks from class-agnostic segmentation models, such as those in [7], [8], into 3D space using RGB-D images of the entire scene. These candidate 3D masks are then merged and refined through a voting mechanism applied over mesh-segmented areas, as described in [9]. The dominant group of 3D masks is assigned to each area after the voting process.

This process allows OV-MAP to generate 3D instance proposals without relying on 3D supervised learning models. Each instance is labeled using CLIP, based on the highest-scoring view from the 2D RGB images. A key advantage of this approach is its independence from supervised 3D instance segmentation models, making it adaptable to any scene equipped with RGB-D and point cloud data. Fig. 1 illustrates an example of OV-MAP being used by a mobile robot to recognize objects based on input queries.

*Authors have equal contributions
[1]Interdisciplinary Program in AI, Seoul National University
[2]Artificial Intelligence Institute, Seoul National University
[3]Department of Computer Science, Seoul National University

This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) [RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and (RS-2021-II212068-AIHub/10%, RS-2021-II211343-GSAI/15%, 2022-0-00951-LBA/15%, 2022-0-00953-PICA/20%), NRF (RS-2024-00353991/20%, RS-2023-00274280/10%), and KEIT (RS-2024-00423940/10%) grant funded by the Korean government.

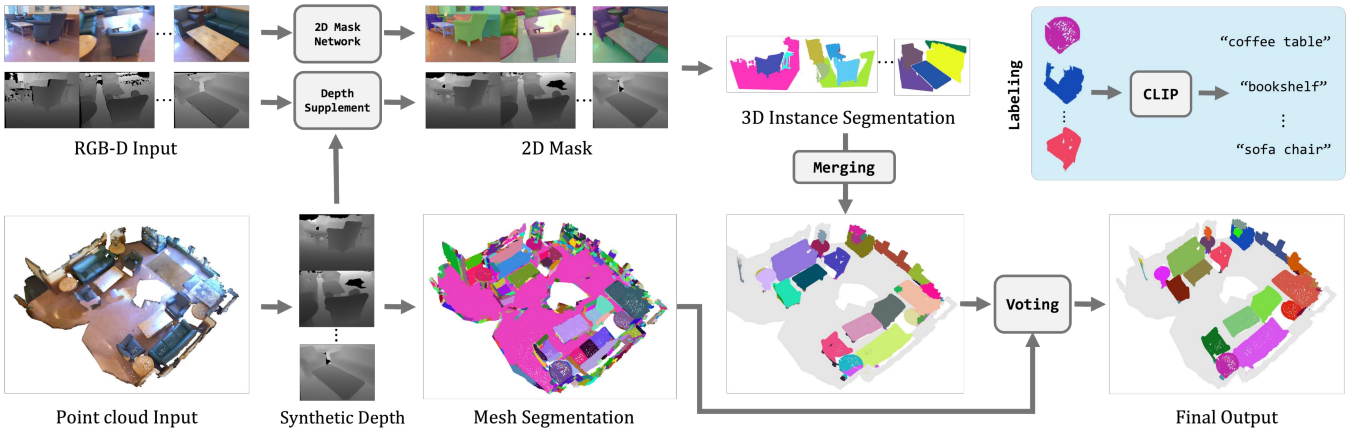


Fig. 2. **Overview of OV-MAP Creation.** The pipeline begins with RGB-D input and point cloud data and progresses to the final 3D instance segmentation maps. Depth images are first refined with synthetic depth from point clouds. Next, RGB images are processed through a 2D class-agnostic segmentation network to generate 2D masks enriched with CLIP features. These masks are then projected into 3D space using supplemented depth data, forming preliminary 3D masks. The 3D masks are integrated with point cloud data to create 3D instance candidates, which are further refined in the final step using a voting mechanism and mesh segmentation to produce the final 3D instance segmentation maps.

We verify OV-MAP through comprehensive experiments on publicly available datasets, including ScanNet200 [10] and Replica [11], where our approach demonstrates robust zero-shot performance. Additionally, we apply OV-MAP in real-world environments, confirming its effectiveness beyond controlled dataset settings.

Our contributions are as follows:

- We propose a mask voting method to merge 2D masks into 3D representations for open-vocabulary instance segmentation.
- We show that our per-instance zero-shot 3D mapping method enhances precision and generalization without relying on supervised 3D instance segmentation models.
- We validate OV-MAP on public datasets and real-world environments, demonstrating its versatility across various applications.

II. Related Works

In the domain of 3D open-vocabulary mapping, research can generally be categorized into two primary approaches: (1) leveraging 2D open-vocabulary models to create per-voxel 3D maps, and (2) employing trained 3D instance prediction models to generate per-instance 3D maps.

A. Per-Voxel 3D Mapping

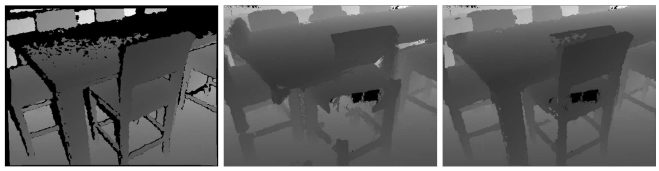
In the context of 3D open-vocabulary mapping, per-voxel approaches rely on extracting features from 2D open-vocabulary models to generate 3D maps at the voxel level. Recent advancements in 2D open-vocabulary segmentation models [12]–[15] have laid the groundwork for exploring 3D open-vocabulary segmentation [2]–[4], [16]–[18]. Projects such as OpenScene [3] and ConceptFusion [4] leverage these 2D open-vocabulary models [12], [13] to extract per-pixel features, which are then projected onto 3D scene voxels to build open-vocabulary representations. Despite these advances, representing 3D scenes at the voxel level presents challenges, particularly the issue where features overflow between voxel boundaries, reducing the accuracy of open-vocabulary queries and making it difficult to distinguish between adjacent instances.

B. Per-Instance 3D Mapping

To overcome the limitations of per-voxel methods, per-instance approaches focus on generating 3D maps at the instance level. OpenMask3D [5], for instance, proposes a method that generates 3D instance proposals using Mask3D [19] and enriches each instance with CLIP features extracted from 2D RGB images. This approach has shown potential for improving mapping accuracy in datasets like ScanNet [20]. However, one limitation of OpenMask3D is that it relies on Mask3D, a 3D supervised model, which struggles to generalize to novel scenes, leading to sub-optimal 3D instance predictions. Our work addresses this issue by introducing a new per-instance mapping approach that enhances generalization across diverse environments. Specifically, we integrate outputs from 2D class-agnostic segmentation model [7] into the 3D space, similar to the method used in SAM3D [6].

III. Method

The process we've developed is outlined in Fig. 2. Starting with RGB-D images from a scene and its reconstructed point cloud, our first step involves generating a list of 3D candidate masks. Each mask is uniquely identified by group IDs generated using class-agnostic 2D pre-trained segmentation models (Sec. III-A). Following this, we initiate a merging process, where similar 3D masks bearing different group IDs are merged under unified group IDs. After this merging phase, we engage in a dominant voting process, allocating the leading 3D mask groups to respective mesh segmented areas (Sec. III-B). This results in final group IDs that are directly associated with individual 3D instances within the scene. In the final step, we enrich each 3D instance with the



(a) Raw Depth (b) Synthetic Depth (c) Supplemented Depth

Fig. 3. **Supplemented Depth Image Construction.** The original depth image (a) is enhanced using the synthetic image (b), producing the supplemented depth image (c).

top-scoring per-mask open-vocabulary CLIP features derived from 2D RGB images (Sec. III-C). This comprehensive approach ensures each 3D instance is accurately defined and enriched within the scene.

A. Candidate 3D Masks Generation

The first step in our approach involves extracting accurate 2D masks for each RGB image using a 2D class-agnostic segmentation model. These masks are then projected onto the 3D surface points of the scene's point cloud, generating 3D masks with unique group IDs.

2D Mask Generation. Given RGB images with a resolution of $H \times W$, denoted as I_t^{RGB}, we obtain 2D masks from the 2D segmentation model $S^{2D}(I_t^{RGB})$, denoted as $m_t^{2D} \in \mathbb{R}^{H \times W \times M}$, where M is the predicted masks, t is the sequence of total frames in the scene. For S^{2D}, we experiment with the class-agnostic segmentation model, CropFormer [7].

Depth Image Construction. To project m_t^{2D} to the 3D point cloud space, depth images are required. However, real-world depth images do not guarantee perfect depth information for every pixel, such as on glass surfaces. To address this, we supplement the missing parts of the depth image, I_t^d, with information from a synthetic depth image, $I_t^{d'}$, derived from a fully constructed point cloud P. The supplementation process can be described as follows:

Given a raw depth image I_t^d and a synthetic depth image $I_t^{d'}$, the supplemented depth image I_t^{sd} is constructed by:

$$I_t^{sd}(i,j) = \begin{cases} I_t^{d'}(i,j) & \text{if } I_t^d(i,j) = 0 \text{ or } I_t^{d'}(i,j) = 0 \\ I_t^d(i,j) & \text{otherwise} \end{cases} \quad (1)$$

for each pixel coordinate (i,j). This approach ensures that the depth image incorporates reliable depth information from the synthetic image where the original depth image is lacking, as shown in Fig. 3.

3D Mask Generation. Given I_t^{sd} and P, and a specific frame index t, we can generate multiple 3D masks, $m_{t,g}^{3D} \in \mathbb{R}^{L \times 3}$, each with L points and a unique group identifier g. This is achieved by projecting a 2D mask, m_t^{2D}, onto P using the corresponding depth images, I_t^{sd}. To facilitate this projection, we first establish a grid of pixel coordinates, $\mathbf{u} = (u,v)$, that matches the resolution of I_t^{sd}. These coordinates are then merged with the depth information to create a set of 3D points in the camera coordinate system, represented as $(u, v, \frac{I^{sd}(u,v)}{s})$, where s is the scaling factor for depth. Subsequently, the inverse of the intrinsic camera matrix, K, is applied to these points to transition them to camera space coordinates. The final step involves transforming these camera space points to world coordinates by applying the pose transformation matrix, T, thereby yielding the 3D mask, $m_{t,g}^{3D}$, that correlates with the original 2D mask, m_t^{2D}.

B. Dominant Group Voting for 3D Instance Segmentation

Upon obtaining candidate 3D masks, denoted by $m_{t,g}^{3D}$, projected onto the point cloud P, we embark on a process of merging these masks and voting to ascertain the dominant area within each mesh segmentation based on [9]. This approach facilitates the identification of dominant groups within the segmentation meshes. As a result, each mesh segmentation is attributed a unique dominant group identifier, g, distinguishing it as a separate entity within the 3D instance segmentations.

Merging. To delve deeper into the merging process, consider $m_{t,g}^{3D}$, where we combine every two frames of a point cloud from $2t$ and $2t+1$ frames to create t' combinations, effectively halving the length of t. Within this setup, each pair of unique group identifiers, g' and g'', in $m_{t',g}^{3D}$, is evaluated for its overlapping rate *OR*. This rate determines whether the groups g' and g'' should be merged into the same group g'. The *OR* is calculated as follows:

$$OR = \frac{\text{overlapping}(m_{t',g'}^{3D}, m_{t',g''}^{3D})}{\max(\text{len}(m_{t',g'}^{3D}), \text{len}(m_{t',g''}^{3D}))} \quad (2)$$

If the *OR* for a smaller group surpasses a predefined threshold, then group g'' is merged into group g', represented by the following:

$$\text{if } OR > \text{threshold, then assign } m_{t',g''}^{3D} \text{ to } m_{t',g'}^{3D}. \quad (3)$$

Our merging criterion is specifically designed to differ from methods used by others, like [6], which merges groups based predominantly on the overlapping rate of smaller groups. Although this approach can result in a higher rate of group consolidation, it bears the drawback of larger 3D masks potentially subsuming smaller ones, consequently eroding the granularity of detail. Moreover, inaccuracies in mask generation, particularly those arising from blurred RGB images, can critically degrade the quality of segmentation when larger masks unduly influence the merging process.

Given these factors, our method prioritizes calculating the *OR* by concentrating on larger masks to ensure that only those with substantial similarity are combined. This careful approach preserves the detail within smaller masks and lessens the chances of errors when merging into larger masks. Additionally, we refine the resolution of the 3D mask area by voxelizing merged points at a quarter of the original voxel scale. This step is key to preserving clear group dominance within the 3D space, which in turn, enhances the dominant group voting process—a critical component for the next stage of segmentation.

Dominant Voting. After merging t frames into a comprehensive 3D mask $m_g^{3D} \in P$, with each mask having unique identifiers g for the entire scene's 3D masks, we identify mesh segmented areas, denoted as $ms_a \in P$, where a represents each distinct segmented area. These areas are

TABLE I

3D INSTANCE SEGMENTATION RESULTS ON SCANNET200 [10].

Model	Open-Vocab	3D Proposal	Map Type	AP	AP_{50}	AP_{25}	head (AP)	common (AP)	tail (AP)
Mask3D [19]		Supervised	Per-Instance	26.9	36.2	41.4	39.8	21.7	17.9
OpenMask3D [5]	✓	Mask3D [19]	Per-Instance	15.4	19.9	23.1	17.1	14.1	14.9
OpenScene [3]	✓	-	Per-Voxel	6.6	10.2	14.8	7.2	5.8	6.9
SAM3D [6]	✓	None	Per-Instance	8.4	13.1	18.7	9.3	7.0	9.1
Ours	✓	None	Per-Instance	**11.9**	**17.4**	**23.2**	**12.5**	**10.5**	**12.7**

TABLE II

3D INSTANCE SEGMENTATION RESULTS ON REPLICA DATASET [11].

Model	AP	AP_{50}	AP_{25}
Mask3D [19]	5.8	8.5	10.7
OpenMask3D [5]	13.1	18.4	24.2
OpenScene [3]	7.3	9.4	11.2
Ours	**14.2**	**19.6**	**28.1**

determined through [9], with the calculation of ms_a based on surface normals. This method positions ms_a as the foundational units for our 3D instance segmentations, marking them as the primary representatives. Subsequently, each area in ms_a is assigned the most dominant group id g, determined through a voting process within the area, formalized by the following equation:

$$ms_{a'} = \text{vote}(ms_a \cap m_g^{3D}) \tag{4}$$

Finally, with ms_a assigned the dominant group id g, it becomes the definitive outcome of our 3D instance segmentation effort, with each $ms_{a'}$ representing an individual instance.

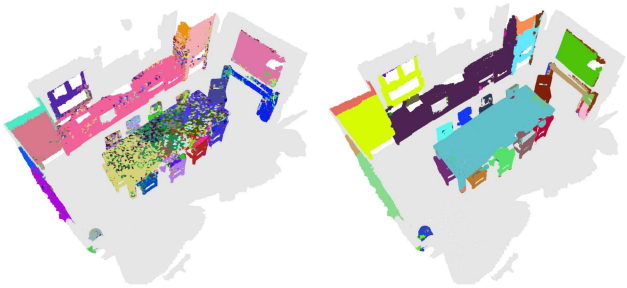
C. Per-Mask Open-Vocabulary Embedding

During the process of obtaining the final 3D instance segmentation ms_a, we assign 3D mask scores to each group g within $m_{t,g}^{3D}$. These scores are calculated based on the pixel counts in the RGB image I_t^{RGB} and the point count in the point cloud P, as defined by:

$$\text{Score}_{t,g} = \alpha \cdot \text{PixelCount}(I_t^{RGB}, g) + \beta \cdot \text{PointCount}(P, g) \tag{5}$$

where α and β are weighting factors that modulate the contributions of pixel and point counts to the total score of each 3D mask.

In the merging process, if two group ids are combined, the higher score, $\text{Score}_{t,g}$, is preserved for the unified group g. In the finalization stage of the 3D instance segmentation, the highest $\text{Score}_{t,g}$ within each group is leveraged to select targeted frames from I_t^{RGB}. This selection process enables the extraction of 2D image crops corresponding to the g mask area. Subsequently, these crops are embedded with CLIP features for open-vocabulary inference, enhancing the semantic richness and applicability of the segmentation in diverse analytical contexts.



(a) Raw Depth Input (b) Supplemented Depth Input

Fig. 4. **Impact of Using Supplemented Depth Data.** Comparison of results from raw depth data (a) against supplemented depth data (b) for 3D map construction, specifically examining outputs produced prior to the implementation of the dominant voting process.

IV. EXPERIMENTS

This section outlines our comprehensive assessment of the proposed 3D instance segmentation approach. We begin by detailing the datasets and evaluation metrics employed in our analysis (Sec. IV-A). Subsequent sections delve into statistical analysis of segmentation performance (Sec. IV-B), presentation of qualitative outcomes (Sec. IV-C), a comparative ablation study analyzing the effects of different depth data types on segmentation performance (Sec. IV-D), and real-world experiments to demonstrate the practical applicability of our method (Sec. IV-E).

A. Experimental Setting

Datasets. Our evaluation leverages the ScanNet200 [10] and Replica [11] datasets, offering a diverse set of environments for assessing our method's efficacy. Specifically, the validation subset of ScanNet200, comprising 312 scenes annotated across 200 categories, serves as the primary benchmark for our 3D instance segmentation task. The categories are divided into head (66 categories), common (68 categories), and tail (66 categories) subsets based on label frequency, facilitating an in-depth analysis across varying data distributions. Furthermore, the inclusion of the Replica dataset, with its detailed office and room scenes, aids in evaluating our method's zero-shot ability across different settings.

Metrics. To quantify the segmentation accuracy, we adopt the average precision (AP) metric, a standard in 3D instance segmentation evaluations. AP calculations are conducted at two mask overlap thresholds—50% and 25%—and results are averaged over a range from 0.5 to 0.95 in increments

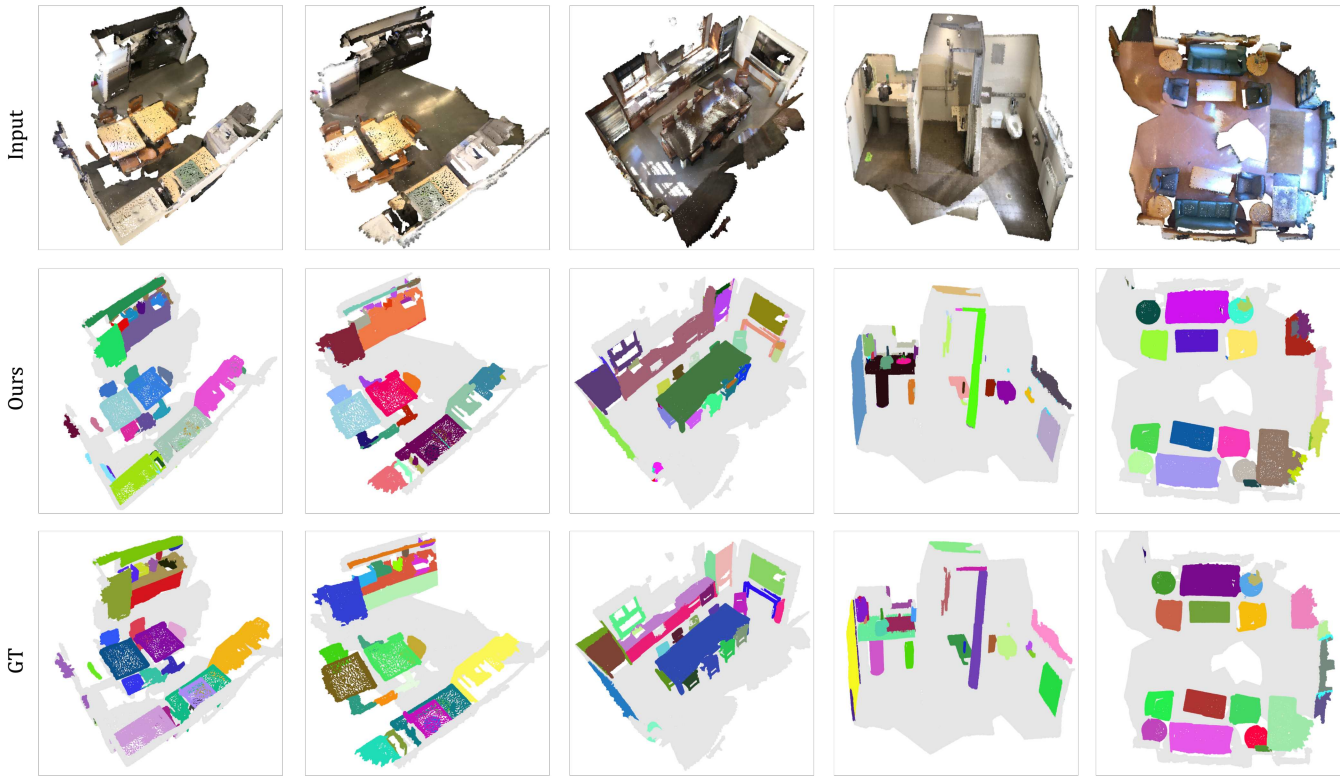


Fig. 5. **Qualitative Analysis of Segmentation Performance.** The images show input scenes, Ground Truth (GT) annotations for each scene, and our model's results demonstrating the effectiveness of open-vocabulary instance segmentation.

of 0.05, as per ScanNet's evaluation protocol [10]. Each predicted mask in our model is ascribed a uniform prediction confidence score of 1.0 for simplicity in metric computation.

Implementation details. Our system processes posed RGB-D pairs from ScanNet200 and Replica datasets, analyzing 1 frame out of every 10 in RGB-D sequences. To obtain object-level masks, we employ CropFormer [7] as our 2D mask predictor. For open-vocabulary feature extraction, we use CLIP ViT-H. We voxelize the reconstructed point cloud with a radius of 2cm and merge voxelization with a radius of 0.05cm for ScanNet and Replica. We perform post-processing to refine the output 3D mask by using the nearest neighbor algorithm [21] and DBSCAN [22] to separate disconnected point clusters and remove irrelevant ones.

B. Quantative Results

Evaluation on ScanNet200. We present our evaluation results for the ScanNet200 dataset, detailing AP, AP50, and AP25 metrics for the dataset's categories in Tab. I. Our method outperforms traditional per-voxel mapping approaches, highlighting the efficacy of our zero-shot 3D instance segmentation strategy. Notably, methods such as Mask3D [19] and OpenMask3D [5] exhibit better performance, benefiting from direct learning on the ScanNet dataset.

While supervised methods generally excel in the 'head' and 'common' categories, leveraging the extensive training on these more frequent categories, they tend to falter in the 'tail' category, where instances are less common. This contrast underscores a limitation in the adaptability of supervised approaches to the long-tail distribution of dataset categories. In comparison, our method maintains consistent performance across all categories—head, common, and tail—demonstrating its robustness and general applicability for 3D instance segmentation across a diverse range of scenes.

Evaluation on Replica. We performed a zero-shot evaluation of our method on the Replica dataset to assess its generalization potential. The results, detailed in Tab. II, show our method significantly outperforming the baseline models. This performance disparity highlights the limitations of models reliant on ScanNet-specific training when faced with the Replica dataset, a different environment. Such a decline in performance underscores the importance of further developing zero-shot 3D instance mapping techniques. Our findings advocate for methods with higher adaptability and underscore the need for models that can seamlessly generalize across a variety of datasets, ensuring wide-ranging applicability and a deeper understanding of diverse environments.

C. Qualitative Results

As shown in Fig. 4, incorporating additional depth information from point clouds is crucial for the transition from 2D to 3D segmentation techniques. While precise 2D segmentation is foundational, its effectiveness may be diminished when depth images are affected by reflective objects, potentially hindering the 2D to 3D conversion process. Therefore, enriching our depth data with scenes reconstructed from

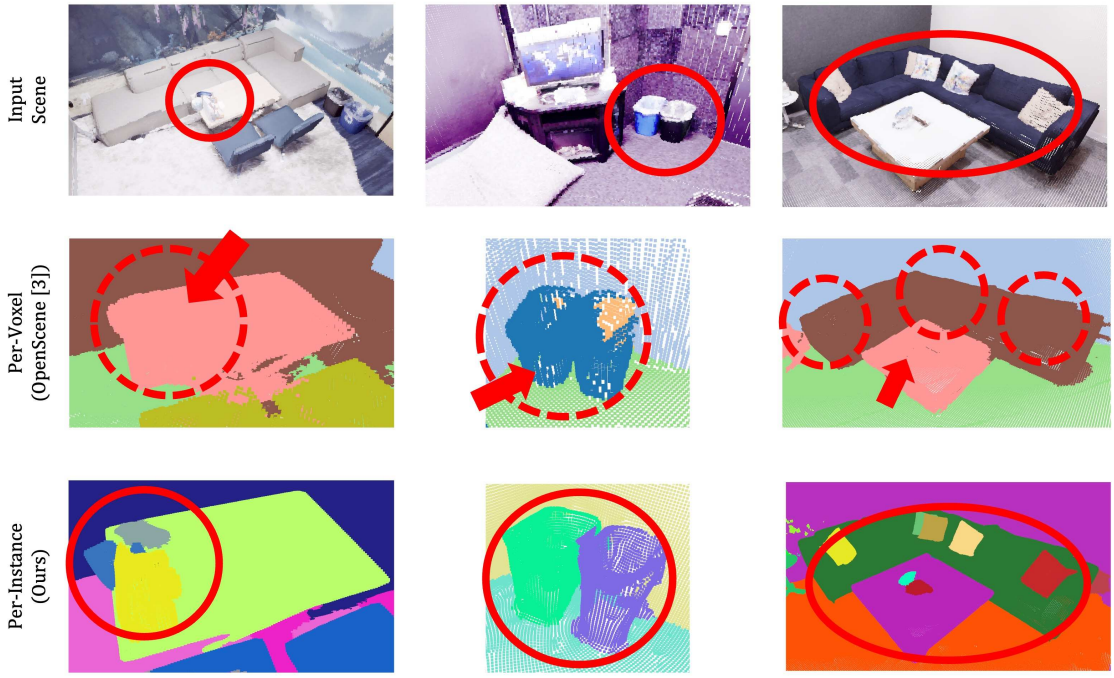


Fig. 6. **Comparative Analysis of Segmentation Results.** Per-Voxel method [3] vs. Per-Instance method (Ours). Unlike the per-voxel method, our approach segments nearby objects, such as items on a desk, an attached trash bin, and cushions on a sofa.

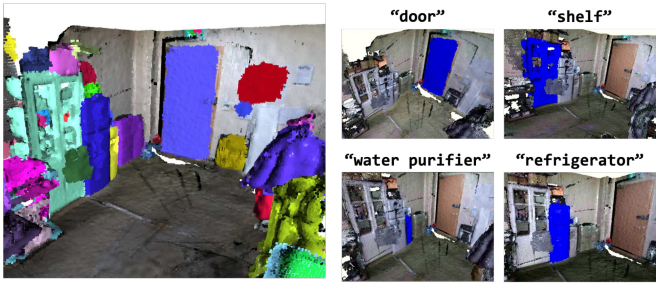
TABLE III
EFFECT OF DIFFERENT DEPTH IMAGE TYPES ON 3D INSTANCE SEGMENTATION [10].

Depth Type	AP	AP_{50}	AP_{25}
Raw depth	11.1	17.1	22.5
Synthetic depth	10.1	15.9	20.7
Supplemented depth	**11.9**	**17.4**	**23.2**

point clouds plays a vital role in achieving reliable 2D to 3D mapping.

Further, in Fig. 5, we present qualitative examples of our approach applied to the task of open-vocabulary 3D instance segmentation. Leveraging its zero-shot capabilities, OV-MAP generates high-quality 3D instance segmentation maps where every instance is amenable to open-vocabulary queries. Remarkably, OV-MAP achieves this level of accuracy without relying on models trained specifically on the ScanNet200 dataset, showcasing its effectiveness and the potential for wide applicability in diverse settings.

Furthermore, Fig. 6 showcases the distinguishable advantages of our method over the traditional per-voxel mapping approach [3]. The per-voxel method often conflates adjacent instances, leading to a less accurate representation of the scene. In contrast, our utilization of a 2D class-agnostic model for segmentation demonstrates superior performance in distinguishing between individual instances. This approach not only enhances the clarity of the mapping results but also underscores the effectiveness of class-agnostic models in achieving precise instance segmentation within 3D spaces.



(a) Segmentation Result (b) Open-Vocabulary Queries

Fig. 7. **Real-World Map Creation Validation.** (a) 3D instance segmentation on real-world data. (b) Open-vocabulary query results showing segmented objects.

D. Ablation Study

To evaluate the impact of different depth data types on 3D mask generation, we conducted an ablation study comparing raw, synthetic, and supplemented depth images. As shown in Table III, the use of supplemented depth data, which combines both raw and synthetic sources, significantly improves performance across all metrics. For instance, the AP increases from 11.1 (raw) to 11.9 (supplemented), with corresponding gains in AP50 and AP25. This highlights the effectiveness of using enhanced depth data to improve 2D-to-3D mask projections, resulting in more accurate 3D instance segmentation.

E. Real-World Experiments

To validate the real-world applicability of our approach, we captured RGB-D data from an actual indoor environment and processed it to generate detailed point clouds of the scene. Our method was then applied to accurately identify and segment individual objects based on open-vocabulary queries. The results of the instance segmentation are shown in Fig. 7(a), while the corresponding open-vocabulary query results are illustrated in Fig. 7(b). Notably, objects such as 'water purifier,' 'door,' 'shelf,' and 'refrigerator' were correctly segmented and matched purely from open-vocabulary descriptions, without relying on predefined labels. These results highlight that our OV-MAP method not only achieves accurate instance segmentation but also excels in handling open-vocabulary inputs, demonstrating its effectiveness and potential for use in diverse, real-world environments.

V. Conclusion

In this paper, we presented OV-MAP, an open-vocabulary zero-shot 3D instance segmentation map for mobile robots. By employing a class-agnostic segmentation model for accurate 2D-to-3D mask projection and incorporating a 3D mask voting mechanism, our method achieves zero-shot 3D instance mapping. It demonstrates superior performance through rigorous testing on the ScanNet200 and Replica datasets. This work represents a significant advancement in open-vocabulary instance segmentation for 3D scenes and has potential applications in mobile robotics.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[2] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615, IEEE, 2023.

[3] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

[4] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, A. Maalouf, S. Li, G. Iyer, S. Saryazdi, N. Keetha, *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," *arXiv preprint arXiv:2302.07241*, 2023.

[5] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," *arXiv preprint arXiv:2306.13631*, 2023.

[6] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, "Sam3d: Segment anything in 3d scenes," *arXiv preprint arXiv:2306.03908*, 2023.

[7] L. Qi, J. Kuen, W. Guo, T. Shen, J. Gu, J. Jia, Z. Lin, and M.-H. Yang, "High-quality entity segmentation," *arXiv preprint arXiv:2211.05776*, 2022.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, pp. 167–181, 2004.

[10] D. Rozenberszki, O. Litany, and A. Dai, "Language-grounded indoor 3d semantic segmentation in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[11] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[12] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv:2201.03546*, 2022.

[13] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

[14] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan, *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19446–19455, 2023.

[15] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision*, pp. 540–557, Springer, 2022.

[16] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Clip on wheels: Zero-shot object navigation as object localization and exploration," *arXiv preprint arXiv:2203.10421*, vol. 3, no. 4, p. 7, 2022.

[17] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Pla: Language-driven open-vocabulary 3d scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7010–7019, 2023.

[18] H. Ha and S. Song, "Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models," *arXiv preprint arXiv:2207.11514*, 2022.

[19] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8216–8223, IEEE, 2023.

[20] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

[21] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 international conference on intelligent computing and control systems (ICCS)*, pp. 1255–1260, IEEE, 2019.

[22] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.