Beyond Fixed Forms: Query-Aware Refinement for Open-Vocabulary 3D Instance Segmentation

Zhenghao Zhang
TU Munich
Munich, Germany
zhenghao.zhang@tum.de

Jie Hu
TU Munich
Munich, Germany
jie.hu@tum.de

Abstract

In the current state of research on open-vocabulary 3D instance segmentation, various methods have been proposed to obtain 3D instance masks with their semantic features. Although previous methods produce high-quality masks with meaningful features, the query phase remains unexploited and follows a simple retrieval approach. The final output can only be selected from fixed instance masks, which face challenges in accurately identifying small and geometrically ambiguous objects and handling queries with varying granularity. Addressing the lack of research in this particular direction, we introduce ***Beyond Fixed Forms (BeyondFF)****, a generic query-aware refinement process that can be applied to any off-the-shelf method for more comprehensive and well-aligned results. Our proposed method leverages semantic information within text queries to rediscover objects from multi-view RGB images and refine fixed-shape 3D instance masks to desired forms. Evaluations on ScanNet200 demonstrate that our method can qualitatively generate fine-grained masks that match corresponding text queries and bring significant performance gains on top of the state-of-the-art methods.*

1. Introduction

Understanding 3D scenes has always been a longstanding goal in research. The task of open-vocabulary 3D instance segmentation is formalized by Takmaz et al. [24] as segmenting and identifying 3D objects beyond the closed-vocabulary setting at instance level.

According to Yan et al. [25], current methods can be broadly divided following a two-stage scheme: 1. zero-shot 3D instance mask prediction. 2. open-vocabulary semantic queries. In the first stage, 3D instance masks are produced along with features to allow them to be identified in the second stage through text queries. Current methods [1, 6, 14, 16, 23–26] have demonstrated a decent ability to segment 3D instances without predefined categories. In the second stage, a simple mask retrieval process is performed, typically using cosine similarity.

However, a common trait of these methods is that the 3D instance masks remain fixed once generated. While these methods keep making progress in producing 3D instance masks in the first stage, the ability to accurately query small instances is limited by these fixed-shape masks. A small object might be shadowed by the instance mask of surrounding objects in the first stage and fail to be correctly identified with text queries. Also, the text queries serve purely as retrieval keys, their rich semantics are not exploited to their full potential.

To tackle these inherent limitations of the prediction-and-retrieval paradigm, we introduce Beyond Fixed Forms (BeyondFF), a process for rediscovering missed masks and refining fixed masks.

Our main contributions are listed as follows:

1. We introduce the idea of utilizing semantic information from text inputs to modify generated masks during the query process.
2. We propose an effective refinement module for combining results from different stages.
3. We demonstrate that incorporating BeyondFF into the existing model brings significant performance gain.

2. Related Works

2.1. Open-vocabulary 2D Understanding

Open-vocabulary 2D understanding is a class of tasks aiming at recognizing and localizing objects in 2D images without a fixed set of predefined categories. With the advancement in large vision-language pre-trained embedding models, e.g. CLIP [18], ALIGN[7], open-vocabulary object detection (OVOD) [3, 8, 10, 13, 27] and open-vocabulary semantic segmentation (OVSS) [2, 5, 9, 11, 17] tasks can be performed given free-form texts. In this work, we tackle the open-vocabulary 3D instance segmentation by leveraging the prior from open-vocabulary 2D models.

2.2. Open-Vocabulary 3D Instance Segmentation

Open-vocabulary 3D instance segmentation (OV-3DIS) focuses on identifying and segmenting instances in 3D scenes through open vocabulary. While high-quality 2D views are generally more readily available than detailed 3D reconstructions, some works adopt 2D instance segmentation techniques to process the frames, followed by back-projecting 2D masks into 3D point regions and employing various strategies to grow or aggregate these regions to form coherent 3D instance proposals [14, 25]. Other works utilize class-agnostic 3D backbones [15, 22] to generate zero-shot instance mask proposals in a 3D manner [6, 24]. Recent works [16, 23, 26] combine 2D and 3D modules for better mask prediction and more robust mask-feature association.

Most existing works primarily focus on producing high-quality, zero-shot 3D instance masks. However, these methods often use CLIP[18] to generate features for mask proposals and treat the query process merely as a retrieval task, selecting pre-generated 3D instance masks that correspond to the text query [14, 16, 23, 24]. In contrast, we focus on utilizing the semantics of the text query and propose BeyondFF, a query process that enables query-aware refinement.

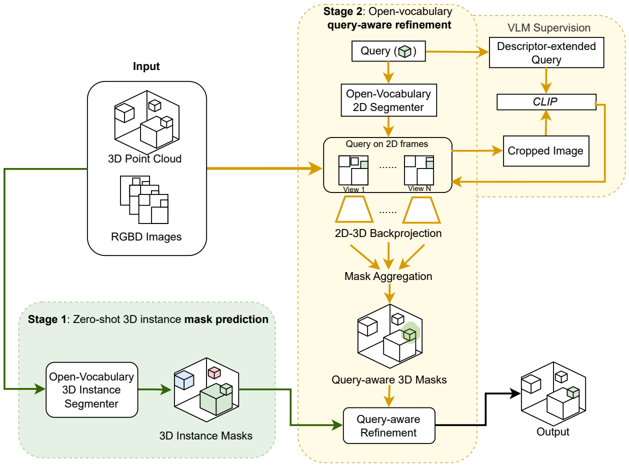

Figure 1. **An overview of our approach.** Our pipeline follows the two-stage schema. In stage 1, we predict zero-shot 3D instance masks using any off-the-shelf method. In stage 2, we refine stage 1 masks with the help of query semantics. We first take the query as the input for 2D detection and segmentation models to rediscover missed objects on RGB frames. We also introduced a VLM supervision mechanism to prevent false predictions using variants of CLIP [20]. The 2D masks are then back-projected and aggregated into query-aware 3D masks. Finally, the instance masks from stage 1 and stage 2 are combined to form the final results consistent with the query.

3. Method

Similar to prior works, we adopt the two-stage schema for our architecture (see Section Sec. 1). Our method is a generic query-aware refinement process designed for stage 2, that can be integrated into any existing approaches.

An overview of BeyondFF is illustrated in Fig. 1. Our approach takes a 3D point cloud $P = \{p_1, p_2, \cdots, p_N\}$ and a series of RGB-D frames $I = \{i_1, i_2, \cdots, i_T\}$ as input. We assume all camera parameters are known in advance.

In stage 1, we use off-the-shelf OV-3DIS methods to generate 3D instance masks. Methods with a 3D backbone are preferred for extra geometry understanding. In stage 2, we leverage open-vocabulary 2D detection and segmentation models to locate instances of our interest in RGB frames. To ensure the detected objects are aligned with the query, we additionally employ a vision-language model (VLM) supervision module to filter false predictions. After obtaining 2D masks, we perform 2D to 3D back-projection and mask aggregation to get query-aware 3D instance masks. In the final refinement, we combine the query-aware 3D instance masks generated from stage 2 and the 3D instance masks generated in stage 1. Details of the proposed method are presented below.

3.1. Zero-shot 3D Instance Mask Prediction

The first stage of our approach involves generating 3D masks $M_{3D}^{s1} = \{m_1^{s1}, m_2^{s1}, \cdots, m_k^{s1}\}$ with their corresponding text-aligned features $F^{s1} = \{f_1^{s1}, f_2^{s1}, \cdots, f_k^{s1}\}$. It is pertinent to point out that this can be done by any OV-3DIS method that takes a 3D point cloud P and RGB-D frames I as inputs. Because the proposed refinement process is a pure 2D process, we recommend using 3D or 2D-3D combined methods for instance information (see Sec. 2.2).

3.2. Query-aware 2D Instance Segmentation

Object detection and segmentation. For a given text query q in stage 2, we utilize a pre-trained 2D instance segmenter to identify and segment masks on RGB frames. A 2D instance segmenter is composed of a detection model and a segmentation model. The detection model first predicts bounding boxes $B = \{b_1, b_2, \cdots, b_j\}$. The bounding boxes then go through VLM supervision to ensure consistency with the text query, returning filtered bounding boxes in form $B' = \{b_1, b_2, \cdots, b_{j'}\}$ (details explained below). From filtered bounding boxes B', the segmentation model segments masks $M_{2D}^{s2} = \{m_{2D,1}^{s2}, m_{2D,2}^{s2}, \cdots, m_{2D,j'}^{s2}\}$.

Though the proposed method does not depend on a certain choice of 2D instance segmenter, the 2D detecter must have the ability to precisely locate the detected object. Precise bounding boxes are crucial for later segmentation to recover instance shape information for objects that are either shadowed or missed in stage 1. During the development of our pipeline, Grounding DINO [13] shows outstanding capability of outputting minimum bounding boxes that contain the queried objects. Some other options are also tested. For instance, YOLOWorld[3] is an appealing option due to its good performance and fast inference (∼7-8x faster compared to Grounding DINO[13]). However, YOLOWorld tend to predict larger bounding boxes that contain whole objects from the COCO dataset [12] (e.g., the bounding box of the whole bed for query "headboard"). Therefore we decided to use Grounding DINO and SAM [19] despite their slow inference speed.

VLM supervision. We leverage the pre-trained visual-text embedding model CLIP [18]. Concretely, we take inspiration from a CLIP variant, WaffleCLIP [20], which provides more accurate and robust text embedding using feature ensembling. We extend the query q with LLM-generated descriptors or randomized words and characters, calculating the text embeddings of all extended queries $E_q = \{e_{q1}, e_{q2}, \cdots, e_{qx}\}$. The final text feature e_q is the mean of E_q. If the similarity between the text feature e_q and the corresponding bounding box feature e_b is lower than a predefined threshold Θ_{clip} (default 0.2), the bounding box will be filtered. By adding the VLM supervision, we not only decrease the false predictions from the 2D detector but also increase the overall speed as fewer boxes are passed to later processes.

3.3. 2D-3D Instance Fusion

2D-to-3D projection and aggregation. As the camera parameters are known, 2D masks M_{2D}^{s2} can easily be projected to 3D masks M_{3D}^{s2} using intrinsics and poses. The aggregation is done by considering 3D intersection over union $IoU(m_{3D,a}^{s2}, m_{3D,b}^{s2})$. 3D masks with IoU value higher than the threshold Θ_{iou_agg} (default 0.2) will be aggregated to a mask.

Filtering. Filtering consists of two parts: a global point-wise detection rate filtering and a mask-wise overlap resolution. They are both based on the intuition that correct predictions occur more frequently than false/inaccurate predictions.

Detection rate is introduced by Lu et al. [14], which is the frequency of a point being considered as part of the instance c_p^{ins} over the frequency of it being visible c_p^{vis}, i.e. $r_p^{det} = c_p^{ins}/c_p^{vis}$ Points with r_p^{det} smaller than the threshold Θ_{dr} (adaptive threshold, default bottom 30%) are filtered.

Overlap resolution is done by comparing the number of initial projected masks. Only the mask aggregated from the highest number of initial masks can keep the overlapped area, while all other intersected masks must remove the overlapped area from their masks.

After filtering over points, any mask with more than $\Theta_{percentage}$ (default 60%) points filtered, or less than Θ_{small} (default 10) points left will be filtered completely. Then we get query-aware 3D masks M_{3D}^{s2} from stage 2.

3.4. Refinement

After getting masks from stage 1 M_{3D}^{s1} and stage 2 M_{3D}^{s2}, we proceed as described in Algorithm 1. For every stage 2 mask, we calculate the IoU values with all stage 1 masks and find the best match stage 1 mask with the highest IoU. If a valid stage 1 mask is found with IoU lower than Θ_{iou} (default 0.45), it indicates that the stage 2 mask finds something different and therefore can be used to refine the original stage 1 mask. Otherwise, the stage 1 mask already accurately highlights the queried object, so no refinement is needed. In the end, we also use query q to retrieve masks from M_{3D}^{s1} and append them to the final refined results M^{ref}.

Algorithm 1 Refinement using query-aware masks

```
Input: $q$, $M^{s1}$, $M^{s2}$, $\Theta_{iou}$
Output: $M^{ref}$
$M^{ref} \leftarrow \{\}$
$M^{match} \leftarrow \{\}$
for $m^{s2} \in M^{s2}$ do
    Calculate IoUs between $m^{s2}$ and $M^{s1}$
    Find $m^{s1}$ with highest $IoU$
    $M^{match} \leftarrow M^{match} \cup \{m^{s1}\}$
    if $IoU < \Theta_{iou}$ then
        $M^{ref} \leftarrow M^{ref} \cup \{m^{s2} \cap m^{s1}\}$
    else
        $M^{ref} \leftarrow M^{ref} \cup \{m^{s1}\}$
    end if
end for
Use query $q$ to retrieve $M^q$ from $M^{s1}$
$M^{ref} \leftarrow M^{ref} \cup (M^q - M^{match})$
return $M^{ref}$
```

4. Evaluation

Owing to the training-free nature of BeyondFF, we directly perform experiments on ScanNet [4] dataset, which features richly annotated 3D indoor scans and is widely recognized in the research community for its diversity and comprehensiveness, making it an ideal choice for evaluating our pipeline.

Constrained by limited time and resources, we selected a subset of 142 scenes ending with "00" from the ScanNet dataset for evaluation. This subset maintains the diversity of scene types and preserves the class distribution of the ScanNet200 [21] classes, ensuring a representative evaluation.

We downsample the RGB-D frames of ScanNet scenes by a factor of 10 to ensure efficient processing. We run our pipeline using the default values specified in section Sec. 3. After running the pipeline, we set the confidence score for every output 3D instance mask to 1.0 following the setup of Open3DIS [16].

We adopt Average Precision (AP), AP50, and AP25 as evaluation metrics, and analyze the results according to class frequencies (Head, Common, Tail, Base, Novel). To ensure a fair comparison, we evaluate Open3DIS under the same settings. Subsequently, we compare the AP between the vanilla Open3DIS and Open3DIS + BeyondFF.

The inference time depends on the availability of the occurrence frequency class in the scene. Specifically, scenes with numerous instances of the queried class take longer to process, whereas scenes with fewer or no instances of the queried class are processed faster, because no subsequent segmentation needs to be performed if no boxes are detected. The inference time of 2D detection and segmentation ranges from approximately 10 seconds for scenes with 1000-2000 RGB-D frames and few or no frames containing instances of the queried class, to 70 seconds for scenes with 5000-6000 frames and abundant instances of the queried class. On average, it takes 35 seconds to process a class in a scene. The 2D-3D instance fusion and refinement takes less than 1 second in total. All experiments are conducted using a single RTX A5000 GPU with 24GB VRAM.

4.1. Quantative Results

As demonstrated in Tab. 1, adding BeyondFF to Open3DIS increases the overall mAP across different IoU thresholds (50% 95% with 5% intervals) by a margin of 4 and yields significant improvements of AP under 50% and 25% IoU thresholds. [1]

A more detailed analysis in Tab. 1 shows that the improvements are consistent across different scannet200 class groups, as well as across AP under different IoU thresholds. This indicates that BeyondFF effectively leverages the semantics of text query, resulting in more true positive mask predictions in the output.

4.2. Experiment Setup

We show qualitative results of masks after BeyondFF refinement in Fig. 2. It not only is able to rediscover true instances that are shadowed in stage 1 prediction, but also demonstrates good open-vocabulary understanding ability by precisely delineate desired instances using arbitrary text queries.

5. Conclusion and Limitation

In this work, we address the inherent limitations of the prediction-and-retrieval paradigm in the open-vocabulary 3D instance segmentation task by introducing the idea of modifying predicted masks using query semantics. We propose Beyond Fixed Forms, a novel pipeline that facilitates this idea by exploiting the semantic information from text queries to rediscover and refine pre-generated fixed-shape masks. Beyond Fixed Forms has the potential to be applied on top of any approach for better query-related results.

Limitations. Our refinement process is a pure 2D approach, thus lacking a geometric and global understanding of the whole scene.Moreover, inference time is the major performance bottleneck of our method, making BeyondFF hard to use in time-constrained scenarios. This is mainly due to the slow inference speed of the current choice for 2D instance segmenter. With the emergence of better and faster models, a real-time refinement process with 3D understanding combined would be an interesting future direction.

Classes	Methods	AP	AP50	AP25
Overall	Open3DIS [16]	23.7	28.2	31.2
	Open3DIS + BFF	**27.4**	**33.3**	**39.6**
Head	Open3DIS	27.0	32.5	35.5
	Open3DIS + BFF	**29.4**	**36.5**	**42.0**
Common	Open3DIS	21.3	24.9	26.5
	Open3DIS + BFF	**26.9**	**32.0**	**38.6**
Tail	Open3DIS	22.4	26.6	31.1
	Open3DIS + BFF	**25.4**	**30.8**	**37.8**
Base	Open3DIS	23.6	28.7	32.4
	Open3DIS + BFF	**27.8**	**34.6**	**43.2**
Novel	Open3DIS	23.7	28.0	30.8
	Open3DIS + BFF	**27.2**	**32.8**	**38.3**

Table 1. Comparison of Open3DIS and Open3DIS + BeyondFF across different ScanNet200 classes groups. Results only tested on 120 classes.

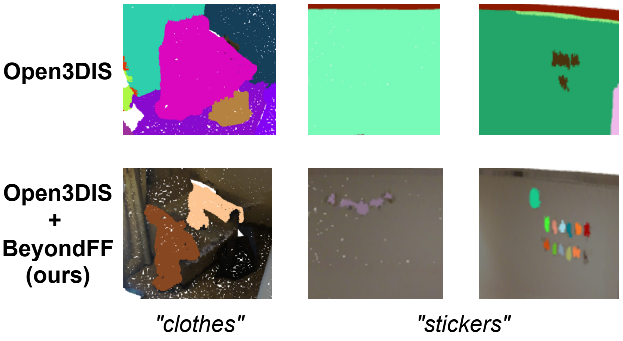

Figure 2. **Qualitative results in Scannet200.** We show that our method can rediscover shadowed objects,e.g., clothes from a sofa on scene0435 (left), stickers from the wall on scene0695 (middle). And our method can refine inaccurate stage 1 masks, such as stickers on scene0695 (right).

References

[1] Mohamed El Amine Boudjoghra, Angela Dai, Jean Lahoud, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan,

[1] Results tested on 120 classes, distributed as follows: Head:40, Common:40, Tail:40, Base:90, Novel:30

and Fahad Shahbaz Khan. Open-yolo 3d: Towards fast and accurate open-vocabulary 3d instance segmentation, 2024. 1

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[3] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 1, 3

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[5] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1

[6] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *arXiv preprint arXiv:2309.00616*, 2023. 1, 2

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[8] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11144–11154, 2023. 1

[9] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 1

[10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[11] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 3

[14] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data, 2023. 1, 2, 3

[15] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution, 2023. 2

[16] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4

[17] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 1

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[19] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

[20] Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts, 2023. 2, 3

[21] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[22] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 2

[23] Hanchen Tai, Qingdong He, Jiangning Zhang, Yijie Qian, Zhenyu Zhang, Xiaobin Hu, Yabiao Wang, and Yong Liu. Open-vocabulary sam3d: Understand any 3d scene, 2024. 1, 2

[24] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[25] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation, 2024. 1, 2

[26] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes, 2024. 1, 2

[27] 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*, pages 350–368. Springer, 2022. 1