# Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance – Supplementary Material –

## 1. Implementation Details

### 1.1. Class-agnostic 3D Segmenter

We adopt the architecture from ISBNet [20] to serve as our class-agnostic 3D proposal network due to its publicly released implementation. This network processes $N$ points in a colored point cloud $\mathbf{P} \in \mathbb{R}^{N \times 6}$ and outputs a collection of $K$ binary 3D instance mask $\mathbf{M} \in \{0,1\}^{K \times N}$. At its core is a 3D UNet backbone a 3D UNet backbone [7], utilizing 3D sparse convolutions [6], which processes the input to produce a feature map $\mathbf{F}^{3D}$ of the point cloud. Subsequently, an instance-wise encoder, based on a sampling strategy, refines these features to produce instance-specific kernels and bounding box parameters. The final stage involves a box-aware dynamic convolution, which employs these instance kernels and mask features, augmented by the corresponding box predictions, to compute the binary mask for each instance.

During inference, we utilize the Intersection over Union (IoU) prediction score to filter out lower-quality masks, with a threshold of 0.2. This score is neutral regarding object classes—during training, the IoU prediction head is trained on the IoU values calculated between the predicted masks and their ground truth counterparts, which are determined by the Bipartite Matching algorithm. Next, we employ superpoints [15, 23] to refine the alignment of our proposals with the actual point cloud structure. This step ensures that our segmentation is consistent with the spatial organization of the point cloud. Lastly, we discard any small proposals that have fewer than 50 points.

### 1.2. Open-Vocabulary 2D Segmenter

In this study, we employ four 2D open-vocabulary instance segmenters: Grounded-SAM[1], DETIC [35], SEEM [36], and ODISE [30]. Here is a breakdown of how each of these segmenters is utilized:
(a) *For Grounded-SAM*, we utilize the Swin-B Grounding DINO decoder [17], which has been pretrained on various datasets including COCO [18], O365 [28], GoldG [13, 22], OpenImage [14], ODinW-35 [16], and RefCOCO [11]. This model is employed to generate bounding boxes from a given text prompt, with box and text thresholds both set to 0.4. Subsequently, these generated bounding boxes are passed through the ViT-L Segment Anything Model [12] to produce instance masks. To process every text query caption, we divide it into chunks, each containing 10 classes, accommodating the limitations of the 77-token decoder. Finally, we apply Non-Maximum-Suppression with an IoU threshold of 0.5 to obtain the ultimate bounding boxes.
(b) *For DETIC*, we follow [19] to use the Swin-B model pretrained on the ImageNet-21K dataset [3] with 21K classes as text queries. We set the confidence threshold at 0.5.
(c) *For SEEM*, we employ the Focal-T visual decoder, which is trained on RefCOCO and LVIS [9], with a logit score threshold of 0.4. Similar to Grounded-SAM, SEEM follows a query processing and post-processing procedure.
(d) *For ODISE*, we utilize the pre-trained label COCO version. This model is complemented by the Stable Diffusion [25] pre-trained on a subset of the LAION [27] dataset, along with Mask2Former [2] serving as the mask generator. We set the confidence threshold to 0.5.

### 1.3. S3DIS and Replica Datasets

(a) *For the S3DIS dataset*, which lacks original mesh data, we apply the superpoint-graph method from the Superpoint Transformer [24] to generate superpoints straight from the 3D point cloud data. For scenes having an extra large number of points (e.g. 1M points), we subsample the point cloud by a factor of 4 for efficient processing.
(b) *For the Replica dataset*, we adopt the mesh segmentation tool[2] based on Felzenszwalb and Huttenlocher's efficient graph-based image segmentation method [5] to create superpoints. The ground-truths for semantic and instance segmentation are provided by [29].

### 1.4. 3D Object Proposal Formation Process

The implementation details of the 3D Object Proposal Formation Process using the *Hierarchical merging order* and *Agglomerative merging strategy* are shown in Alg. 1. Having the 3D point cloud regions obtained from the merging

[1] https://github.com/IDEA-Research/Grounded-Segment-Anything

[2] https://github.com/ScanNet/ScanNet/tree/master/Segmentator

procedure across individual frames $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T\}$, the algorithm merges these independently fragmented regions (see Fig. 1) into well-formed ones recursively, resulting in high-quality augmented 3D proposals.

### 1.5. Point cloud - Image Projection

To establish the correspondence between a 3D point cloud and each frame of the RGB-D sequence $\mathbf{V}$, we employ the principles of pinhole camera projection. Given a 3D point cloud $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 6}$, and for a specific frame $t$, we consider its depth image $\mathbf{D}_t \in \mathbb{R}^{H\times W}$, intrinsic matrix $K_t \in \mathbb{R}^{3\times 3}$ and extrinsic matrix $[\mathbf{R}|\mathbf{c}]_t \in \mathbb{R}^{3\times 4}$, where $\mathbf{R}$ is a 3D rotation matrix and $\mathbf{c}$ is a 3D translation vector. The composite matrix of rotation and translation converts coordinate from the global frame (of the point cloud) to the camera's frame at time $t$. We compute the projection matrix that maps 3D points to 2D image coordinates as follows:

$$\Pi_t = \mathbf{K}_t \cdot [\mathbf{R}|\mathbf{c}]_t \tag{1}$$

Then the 2D projection of a 3D point $\mathbf{p}_i = [x_i^{(3d)}, y_i^{(3d)}, z_i^{(3d)}] \in \mathbf{P}$ is given by:

$$z_i^{(2d)} \cdot \begin{bmatrix} x_i^{(2d)} \\ y_i^{(2d)} \\ 1 \end{bmatrix} = \Pi_t \cdot \begin{bmatrix} x_i^{(3d)} \\ y_i^{(3d)} \\ z_i^{(3d)} \\ 1 \end{bmatrix} \tag{2}$$

where $z_i^{(2d)}$ is the projected depth value and $x_i^{(2d)}, y_i^{(2d)}$ is the 2D pixel coordinate. Next, we discard any points whose projections fall outside the image boundaries, defined by $x_i^{(2d)} \notin [0, W-1]$ or $y_i^{(2d)} \notin [0, H-1]$. To address occlusion within that viewpoint, we further filter out points where the difference between their projected depth and the actual depth recorded at the corresponding pixel in the depth image exceeds a certain depth threshold $\tau_{\text{depth}}$:

$$|z_i^{(2d)} - \mathbf{D}_t[\lfloor y_i^{(2d)} \rfloor, \lfloor x_i^{(2d)} \rfloor]| > \tau_{depth} \tag{3}$$

## 2. Additional Analysis

**Ablation study on the depth threshold** $\tau_{depth}$ is reported in Tab. 1. Overall, $\tau_{depth} = 0.1$ gives the best performance.

**Ablation study on the subsampling factors of RGB-D images** is shown in Tab. 2. By default, we subsample the number of images by a factor of 10. Increasing the subsampling factor to 20 or 40 slightly decreases the performance to 17.1 in AP scores. Reducing the number of images too much yields worse results. We also report the total runtime (in *hours*) to inference on the whole validation set of ScanNet200 in the last column.

**Algorithm 1** 3D Object Proposal Formation

**Input:** $T$ per-frame merged point cloud regions $\{\mathbf{r}_t\}_{t=1}^{T}$.
**Output:** Augmented 3D proposal set $\mathbf{r}$.

**function** HIERARCHICAL_TRAVERSE($s$: start, $e$: end)
  **if** $s = e$ **then**
    **return** $\mathbf{r}_s$ ▷ Look up in $\{\mathbf{r}_t\}_{t=1}^{T}$
  **else**
    $m \leftarrow \lfloor (s+e)/2 \rfloor$
    $\mathbf{r}_{\text{left}} \leftarrow$ HIERARCHICAL_TRAVERSE$(s, m)$
    $\mathbf{r}_{\text{right}} \leftarrow$ HIERARCHICAL_TRAVERSE$(m+1, e)$
    $\mathbf{r} \leftarrow (\mathbf{r}_{\text{left}} \cup \mathbf{r}_{\text{right}})$
    $\mathbf{C}_{\mathbf{r}} \leftarrow$ COST_MATRIX$(\mathbf{r})$ ▷ following Eq. (1) in the main paper
    $\mathbf{r} \leftarrow$ AGGLOMERATIVE_CLUSTERING$(\mathbf{r}, \mathbf{C}_{\mathbf{r}})$
    **return** $\mathbf{r}$
  **end if**
**end function**
$\mathbf{r} \leftarrow$ HIERARCHICAL_TRAVERSE$(1, T)$

| $\tau_{\text{depth}}$ | **AP** | **AP**$_{\text{head}}$ | **AP**$_{\text{com}}$ | **AP**$_{\text{tail}}$ |
|---|---|---|---|---|
| 0.2 | 17.4 | 17.7 | 15.6 | 19.3 |
| 0.1 | 18.2 | **18.9** | 16.5 | 19.2 |
| 0.05 | **18.7** | 17.7 | 16.4 | **22.8** |
| 0.025 | 17.7 | 17.6 | **17.6** | 18.6 |
| 0.01 | 16.7 | 16.3 | 13.8 | 21.2 |

Table 1. Ablation on the depth threshold $\tau_{\text{depth}}$.

| **Use 3D** | **Sub. factor** | **AP** | **AP**$_{\text{head}}$ | **AP**$_{\text{com}}$ | **AP**$_{\text{tail}}$ | **Time** (h) |
|---|---|---|---|---|---|---|
| ✓ | 10 (default) | **23.7** | **27.8** | 21.2 | 21.8 | 20 + 2.3 |
| | 10 (default) | **18.2** | **18.9** | 16.5 | 19.2 | 20 + 2 |
| | 20 | 17.9 | 17.9 | 16.5 | **19.6** | 10 + 1 |
| | 40 | 17.4 | 17.3 | **16.7** | 18.5 | 5 + 0.5 |
| | 80 | 16.5 | 16.7 | 15.4 | 17.1 | |
| | 160 | 13.2 | 12.4 | 12.4 | 15.2 | |
| | 320 | 9.0 | 8.6 | 8.0 | 10.7 | |

Table 2. Study on the subsampling factors of RGB-D images.

**Class-agnostic evaluation on ScanNet200 [26] and ScanNet++ [33]** We further examine the quality of mask proposals generated by Open3DIS on the ScanNet200 and ScanNet++ datasets. In ScanNet200, employing the 3D backbone ISBNet, Open3DIS (2D + 3D) demonstrates superior performance over existing methods in producing high-quality 3D proposals, as depicted in Tab. 3. In ScanNet++, unlike previous methods, we utilize only 100 subsampled 2D RGB-D frames per 3D scene (for computational efficiency). The results using solely 2D data exhibit promising outcomes, as illustrated in Tab. 4.

| Method | AP | $AP_{50}$ | $AP_{25}$ | AR | $AR_{50}$ | $AR_{25}$ |
|---|---|---|---|---|---|---|
| Superpoint | 5.0 | 12.7 | 38.9 | | | |
| DBSCAN [4] | 1.6 | 5.5 | 32.1 | | | |
| OVIR-3D [19] (Detic) | 14.4 | 27.5 | 38.8 | | | |
| Mask Clustering [31] (CropFormer) | 17.4 | 33.3 | 46.7 | | | |
| ISBNet [20] (3D) | 40.2 | 50.0 | 54.6 | 66.8 | 80.4 | 87.4 |
| **Ours** (Grounded SAM) | 29.7 | 45.2 | 56.8 | 49.0 | 70.0 | 83.2 |
| **Ours** (3D + Grounded SAM) | 34.6 | 43.1 | 48.5 | 66.2 | 81.6 | 91.4 |
| **Ours** (SAM) | 31.5 | 45.3 | 51.1 | 61.2 | 87.1 | 97.5 |
| **Ours** (3D + SAM) | 41.5 | 51.6 | 56.3 | 74.8 | 90.9 | 97.8 |

Table 3. Class-agnostic evaluation on ScanNet200 [26] (updated on 2024, Mar. 19th).

| Method | AP | $AP_{50}$ | $AP_{25}$ | AR | $AR_{50}$ | $AR_{25}$ | NOTE |
|---|---|---|---|---|---|---|---|
| ISBNet [20] (3D) | 6.2 | 10.1 | 16.2 | 10.9 | 16.9 | 25.2 | pretrained Scannet200 |
| SAM3D [32] | 7.2 | 14.2 | 29.4 | | | | |
| SAM-guided Graph Cut [8] | 12.9 | 25.3 | 43.6 | | | | |
| Segment3D [10] | 12.0 | 22.7 | 37.8 | | | | |
| SAI3D [34] (SAM) | 17.1 | 31.1 | 49.5 | | | | |
| **Ours** (SAM) | 18.5 | 33.5 | 44.3 | 35.6 | 63.7 | 82.7 | 100 frames per scene |
| **Ours** (SAM) | 20.7 | 38.6 | 47.1 | 40.8 | 75.7 | 91.8 | all frames per scene |

Table 4. Class-agnostic evaluation on ScanNet++ [33] (updated on 2024, Mar. 19th).

To assess the quality of class-agnostic masks in the 2D context, we utilize all masks generated by the 2D-G-3DIP module without any postprocessing, which typically yields high recall albeit at the cost of precision. In the case of 3D masks, we select the top 100 masks from ISBNet based on their confidence scores. Subsequently, to evaluate the Open-Vocab capability, the class-agnostic masks undergo postprocessing by selecting the top k (where k ranges approximately between 300 and 600) masks with the highest CLIP scores. Final confidence score set to 1.0 (OpenMask3D).

# 3. Qualitative Results

## 3.1. Constructing 3D proposals from a single image

In order to acquire high-quality 3D augmented proposals, it is essential to guarantee the effective elevation of 2D masks from a single image to a 3D scene. The extensive overlap of 2D masks often covering multiple objects and the sensitivity of pairing points with pixels due to imperfect camera calibration are the main factors contributing to the poor performance of prior point-based approaches that rely solely on geometric Intersection over Union (IoU). In Fig. 2, SAM3D [32] masks are dispersed over a wide area, while OVIR-3D [19] masks are noisy and fragmented into parts. Open3DIS, however, addresses these issues by considering the superpoints and merging them using averaged 3D deep features. Our method achieves consistency in 3D and 2D, yielding significantly cleaner 3D point cloud regions of corresponding masks on a single 2D image.

## 3.2. Reason for Using Superpoints in 2D-G-3DIP

We have opted to utilize 3D Superpoints as the representation for our innovative 2D-G-3DIP module. The choice of 3D Superpoints is motivated by their remarkable ability to precisely encapsulate the shape and boundary of objects within a 3D scene. Essentially, when we examine an object within the 3D environment, we find that a subset of 3D Superpoints can accurately and completely cover that object's shape, as visually demonstrated in Fig. 3.

Despite the potential imperfections introduced by Depth sensors, previous methods [19, 32] have typically relied on Point Cloud - Image Projection techniques to generate *Point-wise 3D instance masks*. However, this approach often yields a sparse set of 3D proposals, and some points may be obscured, resulting in incomplete masks see in Fig. 5.

In contrast, our Open3DIS takes a distinct approach. We assign weights to groups of points, specifically 3D Superpoints, and harness the power of 3D deep features and geometric Intersection over Union (IoU) calculations. This unique combination allows us to produce *Superpoint-wise 3D instance masks* that are significantly more detailed and precise than what previous methods could achieve. These masks offer a finer-grained representation of object in-

stances in 3D scenes, even in the presence of occlusions and imperfections.

### 3.3. More Qualitative Results on ScanNet200, Replica, and S3DIS

**ScanNet200.** We present visualizations of Open3DIS applied to the extensive Scannet200 dataset. In Fig. 4, we display scenes that have been processed by Open3DIS alongside their corresponding Instance Ground Truth (Instance GT). Despite the considerable size of the Scannet200 dataset, it is important to note that the ground truth annotations may overlook certain relatively small objects within the scenes. These omitted objects are represented by black points, indicating instances that have not been labeled. Open3DIS utilizes both 2D and 3D segmenters to generate comprehensive 3D instance masks, ensuring that even significantly small objects are covered. Although we continue to use the Scannet200 dataset for evaluation purposes, primarily due to its inclusion of a wide range of object classes, we anticipate that Open3DIS will demonstrate notably superior performance when applied to finer-grained 3D instance segmentation datasets.

In comparison to other methods, as depicted in Fig. 5 with a closer look, Open3DIS excels in producing finer 3D masks that effectively cover objects with complex and ambiguous geometric structures. On the other hand, OVIR-3D relies on 2D segmenters and directly extends 2D masks to 3D scenes through point-based Intersection over Union (IoU) matching. This approach results in suboptimal mask quality, despite its capability to discover rare object classes. In contrast, OpenMask3D employs a 3D instance segmenter and evaluates each 3D instance using the CLIP model. While this approach may offer benefits in certain scenarios, it compromises the generality of Open-Vocabulary 3D Instance Segmentation (Open-Vocabulary 3DIS). Particularly, OpenMask3D may struggle to identify rare object classes when expanding the number of classes during training.

Tab. 3 in the main paper provides an illustration of these differences. OpenMask3D, when trained on Scannet20, achieves an Average Precision (AP) score of 12.6, whereas Open3DIS surpasses the state-of-the-art method with an impressive AP score of 19.0. This substantial performance gap underscores Open3DIS's superiority in handling diverse and challenging 3D instance segmentation tasks.

**Replica.** The qualitative results of our approach on the Replica dataset are visualized in Fig. 6a.

**S3DIS.** The qualitative results of our approach on the S3DIS dataset are visualized in Fig. 6b.

### 3.4. Open-Vocabulary Scene Exploration

We showcase the remarkable Open-Vocabulary scene exploration capabilities of Open3DIS on the ARKitScenes [1] (Fig. 7a) and ScanNet200 [26] (Fig. 7b) datasets, which are notable for containing a vast array of scenes featuring diverse and rare objects. Specifically, we demonstrate the system's ability to query instance objects based on various attributes such as material, color, affordances, and usage. We intentionally exclude the Class-agnostic 3D Segmenter component, thereby pushing our method toward a near Zero-Shot Instance Segmentation approach. Remarkably, in challenging scenarios, such as identifying objects like a Post-it note, a picture of a horse, or a bottle of olive oil, Open3DIS outperforms other methods [19, 21, 29, 32] significantly. Some of these methods struggle to detect these objects, let alone locate them accurately. *Please see the supplementary video for a live demo.*

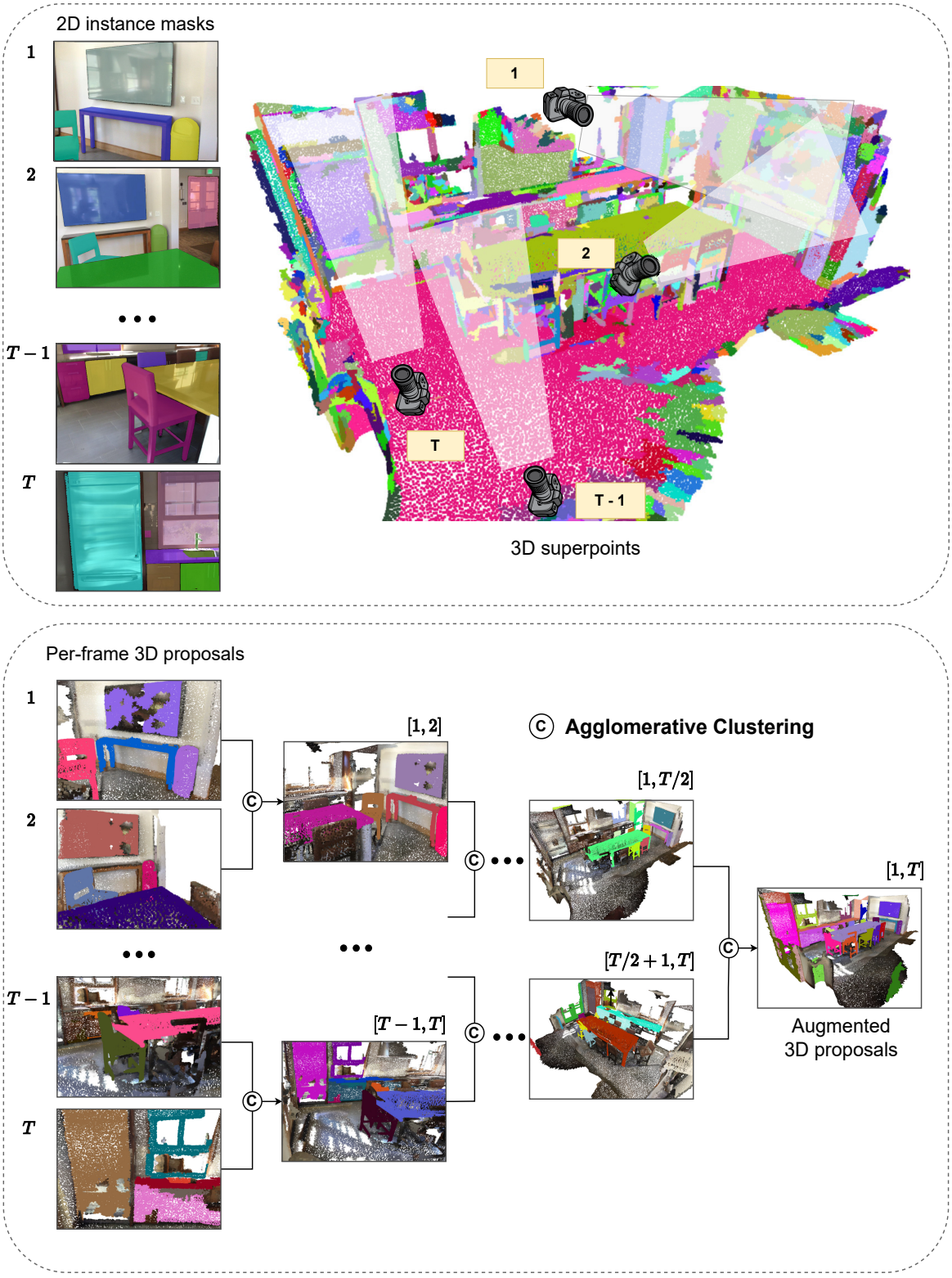



Figure 1. (Top) The 2D-G-3DIP module utilizes 2D per-frame instance masks to generate per-frame 3D proposals by leveraging 3D superpoints. (Bottom) Our proposed hierarchical merging. These proposals are considered point cloud regions and undergo a hierarchical merging process across multiple views, resulting in the final Augmented 3D proposals (Best viewed in color).

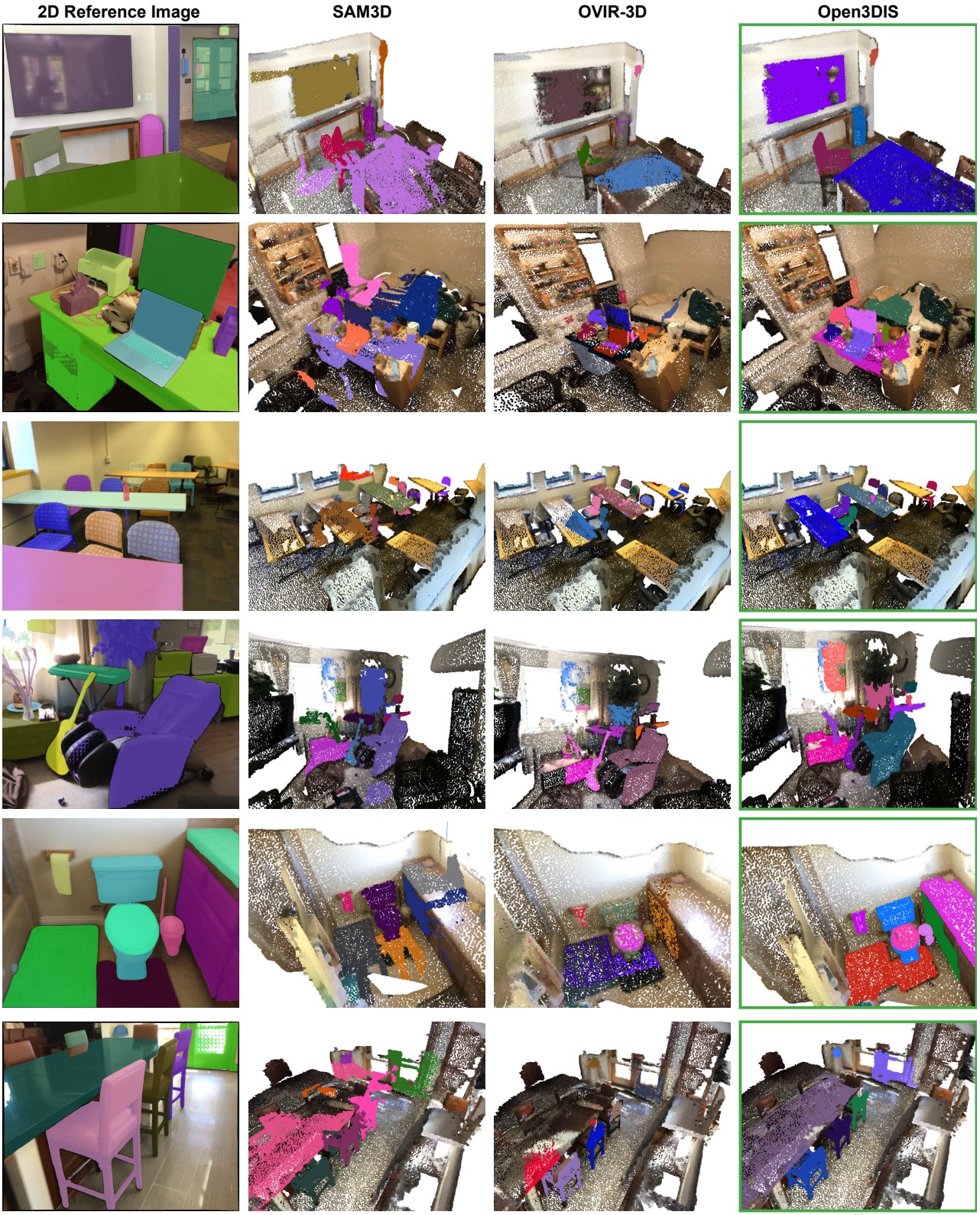



Figure 2. Qualitative results of our method compared to others in Constructing 3D proposals from 2D masks of an image. Each row shows one example, including the input 2D reference image, other 2D lifting methods, and our Open3DIS (**only 2D**) (Best viewed in color).

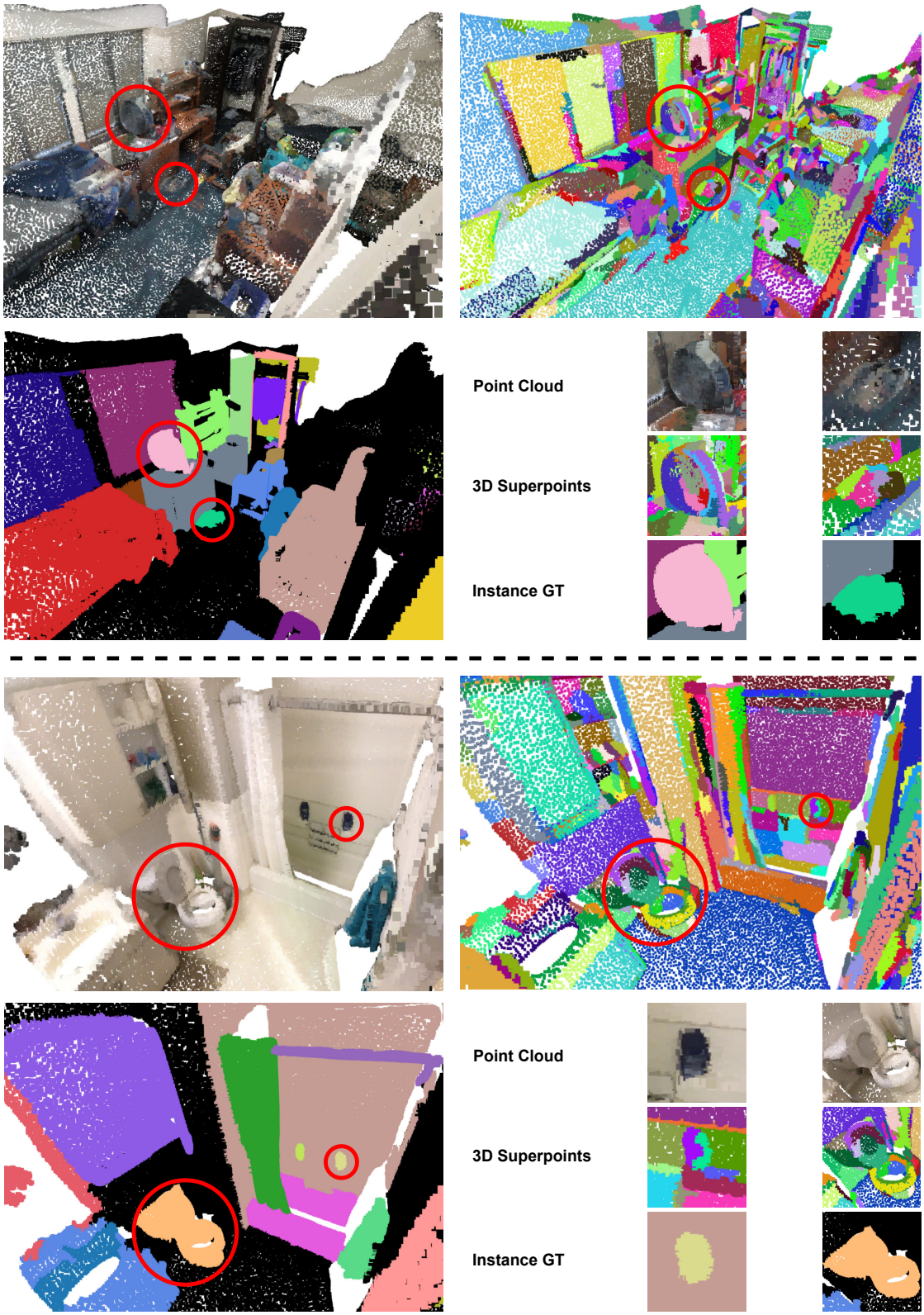



Figure 3. Two examples (separated by the dashed line) illustrating the reason for using the 2D-G-3DIP module when creating point cloud regions, with a focus on accurately covering object instances indicated by the Red circles (Best viewed in color).


Figure 4. Qualitative results of our method on the ScanNet200 dataset. Each row shows one example, including the input RGB point cloud, instance ground truth, and our predictions (Best viewed in color).

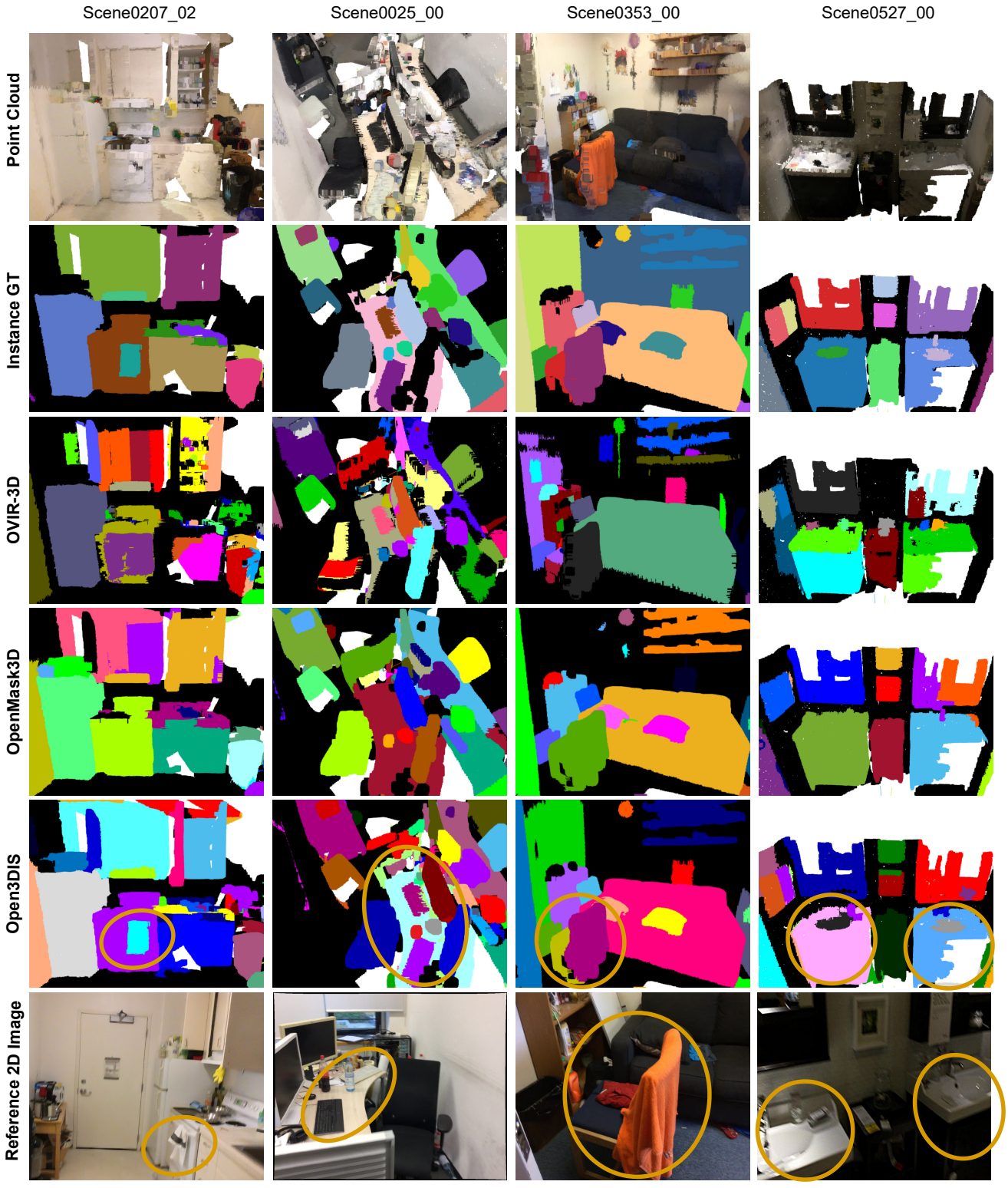
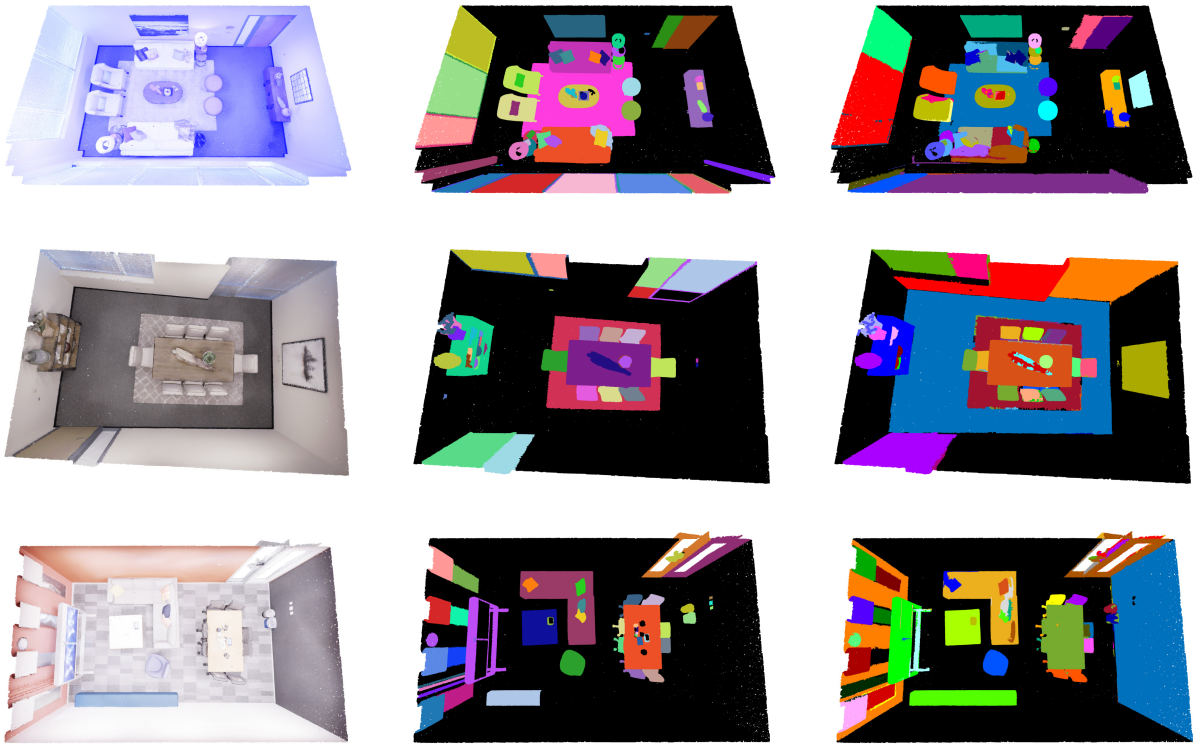
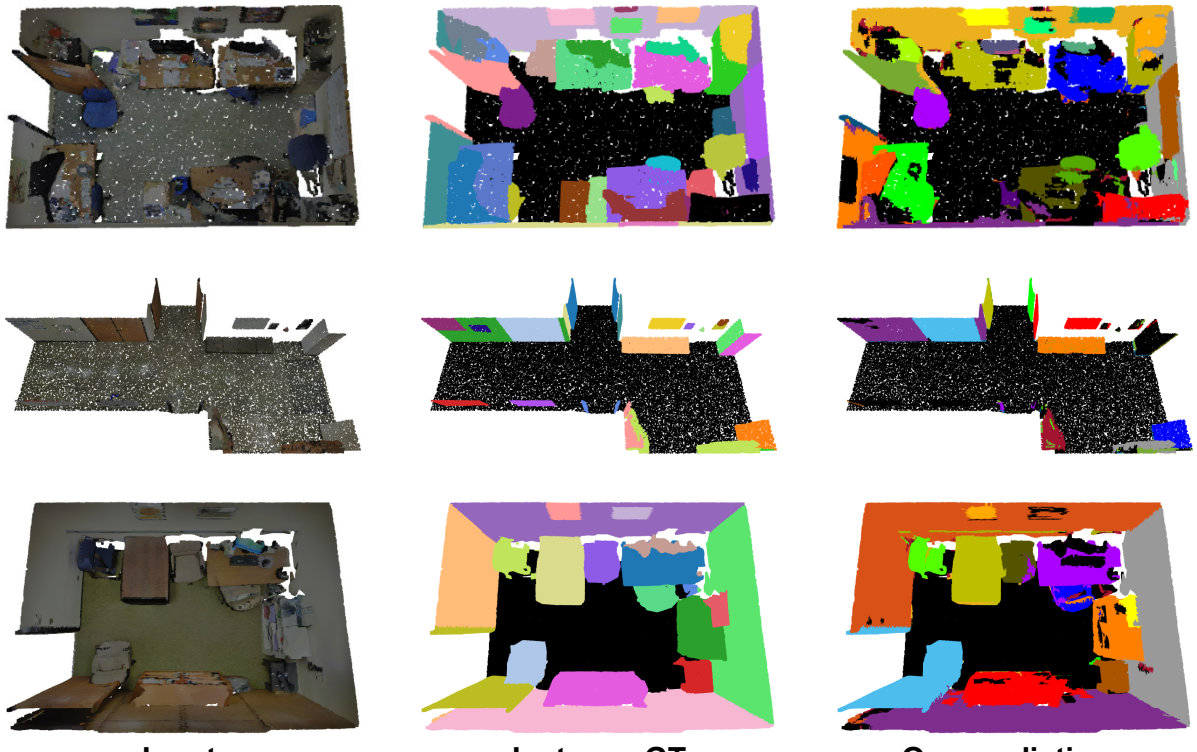



Figure 5. Qualitative results of our method compared to others on ScanNet200 dataset. Each column shows one example in Orange ellipses demonstrating that Open3DIS performs better than others (Best viewed in color).

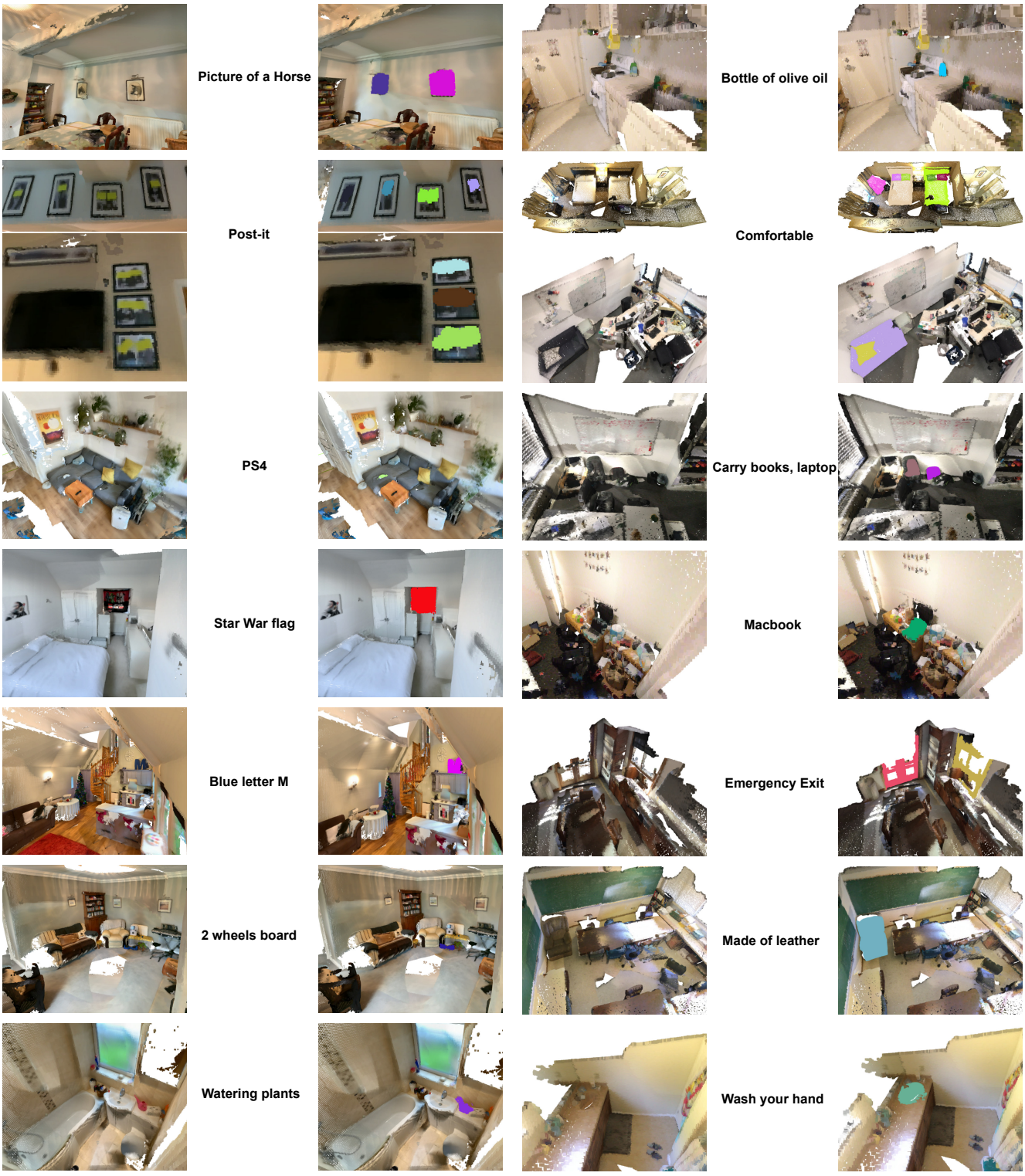



(a) Qualitative results on Replica




(b) Qualitative results on S3DIS

Figure 6. Qualitative results of our method on the Replica (Top) and S3DIS (Bottom) datasets. Each row shows one example, including the input RGB point cloud, instance ground truth, and our predictions (Best viewed in color).




(a) ARKitScenes (b) Scannet200

Figure 7. Open-Vocabulary exploration on **ARKitScenes** [1] (Left) and **Scannet200** [26] (Right) with Open3DIS (2D only). The middle column presents the text queries, the original point cloud is displayed on the left column, and colored regions represent 3D instance proposals on the right column. (Best viewed in color, zoom-in is advised).

## References


[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 4, 11

[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] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[5] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 1

[6] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 1

[7] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 1

[8] Haoyu Guo, He Zhu, Sida Peng, Yuang Wang, Yujun Shen, Ruizhen Hu, and Xiaowei Zhou. Sam-guided graph cut for 3d instance segmentation. *arXiv preprint arXiv:2312.08372*, 2023. 3

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[10] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. *arXiv*, 2023. 3

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[14] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[15] Loic Landrieu and Mohamed Boussaha. Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7440–7449, 2019. 1

[16] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022. 1

[17] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. 2023. 1

[18] 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. 1

[19] 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. 1, 3, 4

[20] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 1, 3

[21] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[22] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[23] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. *arXiv preprint arXiv:2306.08045*, 2023. 1

[24] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. 2023. 1

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[26] 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. 2, 3, 4, 11

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[28] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[29] Aycca 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, 4

[30] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 1

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

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

[33] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[34] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. *arXiv preprint arXiv:2312.11557*, 2023. 3

[35] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1

[36] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 1