Open3DIS: Open-vocabulary 3D Instance Segmentation with 2D Mask Guidance

Phuc Nguyen[1*] Tuan Duc Ngo[1,4*] Chuang Gan[2,4]
Evangelos Kalogerakis[4] Anh Tran[1] Cuong Pham[1,3] Khoi Nguyen[1]
[1]VinAI Research [2]MIT-IBM Watson AI Lab [3]Posts & Telecommunications Inst. of Tech. [4]UMass Amherst
{v.phucnda, v.anhtt152, v.khoindm}@vinai.io {tdngo, kalo}@cs.umass.edu
ganchuang@csail.mit.edu cuongpv@ptit.edu.vn
https://open3dis.github.io/

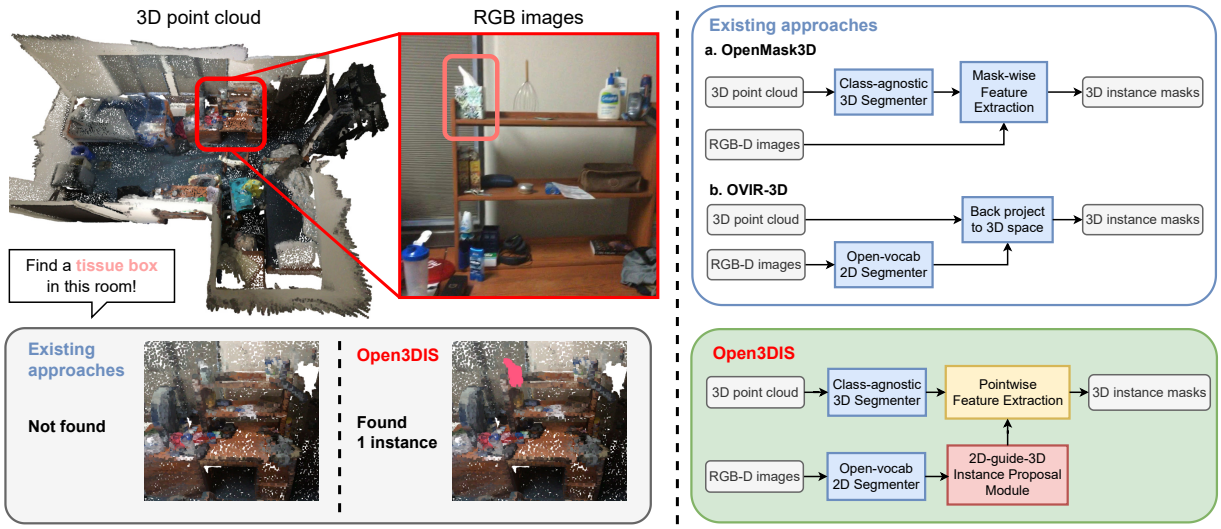


Figure 1. **Left:** While leading open-vocabulary 3D instance segmentation methods like OpenMask3D [61] and OVIR-3D [44] often struggle with small or ambiguous instances, particularly those from uncommon classes, Open3DIS excels in segmenting such cases. It outperforms existing methods by about $\sim$1.5$\mathbf{x}$ in average precision on ScanNet200 [55]. **Right:** Open3DIS aggregates proposals from both point cloud-based instance segmenters and 2D image-based networks. Our method incorporates novel components (red and yellow boxes) that perform aggregation and mapping of 2D masks to the point cloud across multiple frames, as well as 3D-aware feature extraction for effectively comparing object proposals to text queries.

Abstract

We introduce Open3DIS, a novel solution designed to tackle the problem of Open-Vocabulary Instance Segmentation within 3D scenes. Objects within 3D environments exhibit diverse shapes, scales, and colors, making precise instance-level identification a challenging task. Recent advancements in Open-Vocabulary scene understanding have made significant strides in this area by employing class-agnostic 3D instance proposal networks for object localization and learning queryable features for each 3D mask. While these methods produce high-quality instance proposals, they struggle with identifying small-scale and geometrically ambiguous objects. The key idea of our method is a new module that aggregates 2D instance masks across frames and maps them to geometrically coherent point cloud regions as high-quality object proposals addressing the above limitations. These are then combined with 3D class-agnostic instance proposals to include a wide range of objects in the real world. To validate our approach, we conducted experiments on three prominent datasets, including ScanNet200, S3DIS, and Replica, demonstrating significant performance gains in segmenting objects with diverse categories over the state-of-the-art approaches.

1. Introduction

This paper addresses the challenging problem of open-vocabulary 3D point cloud instance segmentation (OV-3DIS). Given a 3D scene represented by a point cloud, we seek to obtain a set of binary instance masks of any classes

*: Equal contribution

of interest, which may not exist during the training phase. This problem arises to overcome the inherent constraints of the conventional fully supervised 3D instance segmentation (3DIS) approaches [20, 21, 47, 57, 60, 63, 75, 77], which are bound by a closed-set framework – restricting recognition to a predefined set of object classes that are determined by the training datasets. This task has a wide range of applications in robotics and VR systems. This capability can empower robots or agents to identify and localize objects of any kind in a 3D environment using textual descriptions that detail names, appearances, functionalities, and more.

There are a few studies addressing the OV-3DIS so far [10, 11, 44, 61]. Most recently, [61] proposes the use of a pre-trained 3DIS model instance proposals network to capture the geometrical structure of 3D point cloud scenes and generate high-quality instance masks. However, this approach faces challenges in recognizing rare objects due to their incomplete appearance in the 3D point cloud scene and the limited detection capabilities of pre-trained 3D models for such infrequent classes. Another approach involves leveraging 2D off-the-shelf open-vocabulary understanding models [44, 73] to easily capture novel classes. Nevertheless, translating these 2D proposals from images to 3D point cloud scenes is challenging. This is because of the fact that 2D proposals capture only the visible portions of 3D objects and may also include irrelevant regions, such as the background. These two approaches are summarized in Fig. 1.

In this work, we introduce Open3DIS, a method for OV-3DIS that extends the understanding capability beyond predefined concept sets. Given an RGB-D sequence of images and the corresponding 3D reconstructed point cloud scene, Open3DIS addresses the limitations of existing approaches. It complements two sources of 3D instance proposals by employing a 3D instance network and a 2D-guide-3D Instance Proposal Module to achieve sufficient 3D object binary instance masks. The module (our key contribution) extracts geometrically coherent regions from the point cloud under the guidance of 2D predicted masks across multiple frames and aggregates them into higher-quality 3D proposals. Later, Pointwise Feature Extraction aggregates CLIP features for each instance in a multi-scale manner across multiple views, constructing instance-aware point cloud features for open-vocabulary instance segmentation.

To assess the open-vocabulary capability of Open3DIS, we conduct experiments on the ScanNet200 [55], S3DIS [1], and Replica [59] datasets. Open3DIS achieves state-of-the-art results in OV-3DIS, surpassing prior works by a significant margin. Especially, Open3DIS delivers a noteworthy performance improvement of $\sim$1.5 times compared to the leading method on the large-scale dataset ScanNet200.

In summary, the contributions of our work are as follows:

1. We present the "2D-guided 3D Proposal Module" creating precise 3D proposals by clustering cohesive point cloud regions using aggregated 2D instance masks from multi-view RGB-D images.
2. We introduce a novel pointwise feature extraction method for open-vocabulary 3D object proposals.
3. Open3DIS achieves state-of-the-art results on ScanNet200, S3DIS, and Replica datasets, exhibiting comparable performance to fully supervised methods.

2. Related Work

Open-vocabulary 2D scene understanding methods aim to recognize both base and novel classes in testing where the base classes are seen during training while the novel classes are not. Based on the types of recognition tasks, we can categorize them into open-vocabulary object detection (OVOD) [29, 43, 49, 64, 74, 76, 80], open-vocabulary semantic segmentation (OVSS) [9, 37, 39, 67, 69, 83], and open-vocabulary instance segmentation (OVIS) [19, 26, 62, 66, 78, 79]. A typical approach for handling the novel classes is to leverage a pre-trained visual-text embedding model, such as CLIP [51] or ALIGN [27] as a joint text-image embedding where base and novel classes co-exist, in order to transfer the models' capabilities on base classes to novel classes. However, these methods cannot trivially extend to 3D point clouds because 3D point clouds are unordered and imbalanced in density, and the variance in appearance and shape is much larger than that of 2D images.

Fully-supervised 3D Instance Segmentation (F-3DIS) aims to segment 3D point cloud into instances of training classes. Methods of F-3DIS can be categorized into three main groups: box-based [24, 71, 75], cluster-based [5, 12, 28, 63, 65], and dynamic convolution-based [20, 21, 42, 47, 57, 60, 68] techniques. Box-based methods detect and segment the foreground region inside each 3D proposal box to get instance masks. Cluster-based methods employ the predicted object centroid to group points to clusters or construct a tree or graph structure and subsequently dissect these into subtrees or subgraphs [25, 40]. For the third group, Mask3D [57] and ISBNet [47], proposed using dynamic convolution whose kernels, representative of different object instances, are convoluted with pointwise features to derive instance masks. In this paper, we use ISBNet as a 3D network, yet with necessary adaptations to output 3D class-agnostic proposals.

Open-vocabulary 3D semantic segmentation (OV-3DSS) and object detection (OV-3DOD) enable the semantic understanding of 3D scenes in an open-vocabulary manner, including affordances, materials, activities, and properties within unseen environments. This capability is highlighted in recent work [17, 23, 48] for OV-3DSS and [4, 45, 82] for OV-3DOD. Nevertheless, these methods cannot precisely locate and distinguish 3D objects with 3D instance masks, and thus cannot fully describe 3D object shapes.

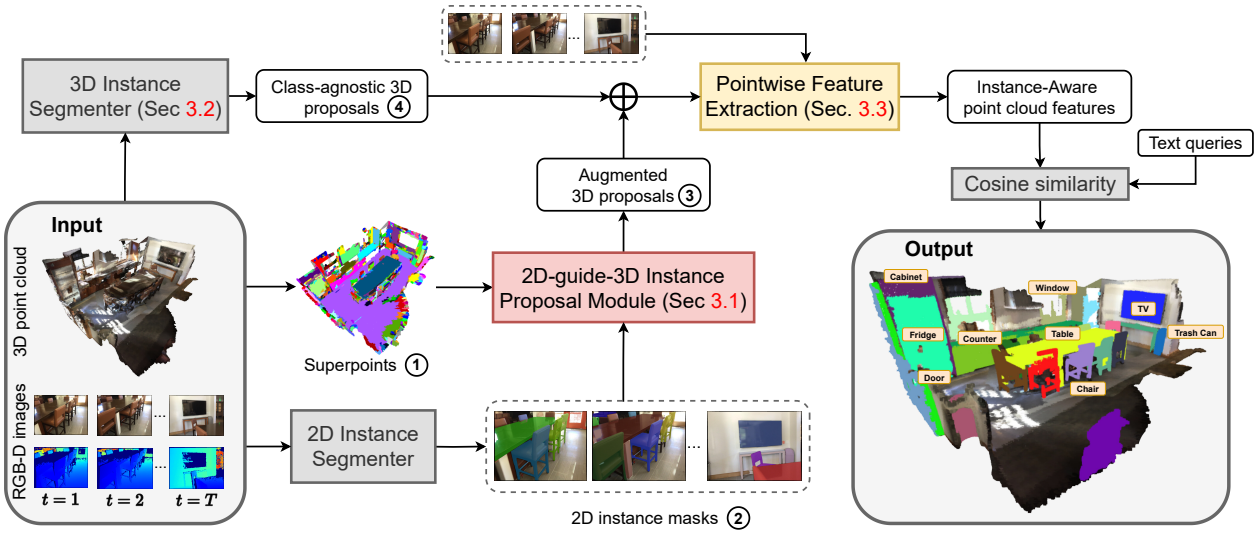


Figure 2. **Overview of Open3DIS**. A pre-trained class-agnostic 3D Instance Segmenter proposes initial 3D objects, while a 2D Instance Segmenter generates masks for video frames. Our 2D-guide-3D Instance Proposal Module (Sec. 3.1) combines superpoints and 2D instance masks to enhance 3D proposals, integrating them with the initial 3D proposals. Finally, the Pointwise Feature Extraction module (Sec. 3.3) correlates instance-aware point cloud CLIP features from multiview images with text embeddings to generate the ultimate instance masks.

Open-vocabulary 3D instance segmentation (OV-3DIS) concerns segmenting both seen and unseen classes (during training) of a 3D point cloud into instances. Methods of OV-3DIS can be split into 3 groups: open-vocabulary semantic segmentation-based, text description and 3D proposal contrastive learning based, and 2D open-vocabulary powered approaches. **The first group** includes OpenScene [48] and Clip3D [22] utilize clustering techniques such as DBScan on OV-3DSS results to generate 3D instance proposals. However, their quality relies on clustering accuracy and can lead to unreliable results for unseen classes. On the other hand, **the second group** comprising PLA [11], RegionPLC [72], and Lowis3D [10] focuses on training the 3D instance proposal network along with a contrastive open-vocabulary between the predicted proposals and their corresponding text captions. However, when growing the number of classes, these methods struggle to handle and may degrade their ability to distinguish diverse object classes. For **the final group**, OpenMask3D [61] uses a pre-trained 3DIS model to produce class-agnostic 3D proposals and classifies them based on their 2D mask projection CLIP score. However, the pre-trained 3DIS model faces challenges in identifying small or rare object categories with uncommon geometric structures. Meanwhile, OVIR-3D [44] and SAM3D [73] leverage pretrained 2D open-vocabulary models to produce 2D instance masks and back-project them onto the associated 3D point cloud. However, the 2D segmentation masks are not well-aligned with objects giving rise to low-quality 3D proposals where background points can be included in foreground objects. Nonetheless, the advantage of this group over other groups is in their leverage of 2D pretrained model on large-scale datasets such as CLIP [51] or SAM [32] which can be scaled to hundreds of classes as in Scannet200 [55]. Following the final group, Open3DIS generates high-quality 3D instance proposals by combining 3D masks from a 3DIS network with proposals produced by grouping geometrically coherent regions (superpoints) with the guidance of 2D instance masks. This complements the class-agnostic 3D instance proposals from 3D networks. Our method excels at capturing rare objects while preserving their 3D geometrical structures, achieving state-of-the-art performance in the OV-3DIS domain.

3. Method

Our approach processes a 3D point cloud and an RGB-D sequence, producing a set of 3D binary masks indicating object instances in the scene. We assume known camera parameters for each frame. Our architecture is depicted in Fig. 2. Similarly to prior work [11, 61, 72], we employ a *3DIS network module* to extract object proposals directly from the 3D point cloud. This module leverages 3D convolution and attention mechanisms, capturing spatial and structural relations for robust 3D object instance detection. Despite its advantages, sparse point clouds, sampling artifacts, and noise can lead to missed objects, especially for small objects e.g., the tissue box in Fig. 1.

Our approach integrates a novel *2D-guide-3D instance proposal* module, leveraging 2D instance segmentation networks trained on large image datasets to better capture smaller objects in individual images. However, resulting 2D masks may only capture parts of actual 3D object instances due to occlusions (Fig. 2 - (2)). To address this, we propose a strategy that constructs 3D object instance proposals by

	Recall	$\text{Recall}_{\text{head}}$	$\text{Recall}_{\text{com}}$	$\text{Recall}_{\text{tail}}$
Only 3D	61.63	81.92	53.68	12.06
Only 2D	68.61	76.66	74.73	34.68
2D and 3D	73.29	87.48	74.16	34.31

Table 1. Recall rate (%) of 2D, 3D, or combined proposals.

hierarchically aggregating and merging point cloud regions from back-projected 2D masks of the same object. To enhance the robustness and geometric homogeneity, we use "superpoints" [14] during the merging process. This yields complete object instances, complementing those extracted by 3DIS networks.

Detailed analysis in Tab. 1 exhibits the significant enhancement in recall rate, especially for *rare* classes, when integrating 2D and 3D proposals.

To enable open-vocabulary classification, we additionally employ a *point-wise feature extraction module* to construct a dense feature map across the 3D point cloud. In the following sections, we explain our modules in more detail, starting with the 2D-guide-3D Instance Proposal Module which constitutes our main contribution.

3.1. 2D-guide-3D Instance Proposal Module

This module takes as input a 3D point cloud $\mathbf{P} = \{\mathbf{p}_n\}_{n=1}^{N}$, where N is the number of points, and $\mathbf{p}_i \in \mathbb{R}^6$ includes 3D coordinates and RGB color. Additionally, it receives an RGB-D video sequence $\mathbf{V} = \{(\mathbf{I}_t, \mathbf{D}_t, \Pi_t)\}_{t=1}^{T}$, where each frame t contains RGB image $\mathbf{I}_t$, depth map $\mathbf{D}_t$, and camera matrix Π_t (i.e., the product of intrinsic and extrinsic matrices used for projecting 3D points onto the image plane). The output comprises K_1 binary instance masks represented in a $K_1 \times N$ binary matrix $\mathbf{M}_1$ (Fig. 2 - ③).

Superpoints. In a pre-processing step, we utilize the method of [14] to group points into geometrically homogeneous regions, termed superpoints (Fig. 2 - ①). This yields a set of U superpoints $\{\mathbf{q}_u\}_{u=1}^{U} \in \{0,1\}^{U \times N}$, where $\mathbf{q}_u$ is a binary mask of points. Superpoints enhance processing efficiency in the later stages of our pipeline and contribute to well-formed candidate object instances.

Per-frame superpoint merging. For all input frames, we utilize a pretrained 2D instance segmenter, employing Grounding-DINO [43] and SAM [33]. The network outputs a set of 2D masks (Fig. 2 - ②). For each 2D mask with index m (unique across all frames), we calculate the IoU $o_{u,m}$ with each superpoint $\mathbf{q}_u$ when projecting all points of $\mathbf{q}_u$ onto the image plane of mask m using the known camera matrix, excluding points outside the camera's field of view, and determining image pixels containing projected points. A superpoint is considered to have sufficient overlap with a 2D mask if the IoU is higher than a threshold $o_{u,m} > \tau_{iou}$.

However, 2D masks may include background regions or parts of nearby objects, making IoU alone insufficient to determine superpoints belonging to a 3D proposal. To address this, we leverage the 3D backbone of a 3D proposal network [47, 57] to extract per-point feature $\mathbf{F}^{3D} \in \mathbb{R}^{N \times D^{3D}}$ and measure feature similarity among these superpoints $\mathbf{q}_u$ whose features are determined by averaging their point features $\mathbf{f}_u^{3D} \in \mathbb{R}^{1 \times D^{3D}}$. For each 2D instance mask $\mathbf{m}_i^{2D}$, we initiate a point cloud region $\mathbf{r}_i$ with the superpoint having the largest IoU with the mask. We extend this region by merging with neighboring superpoints $\mathbf{q}_u$ that meet the overlapping condition (τ_{iou}) and also have the highesst cosine similarity $s_i^{\max} = \max_{u' \in \mathbf{r}_i} \cos(\mathbf{f}_{u'}^{3D}, \mathbf{f}_u^{3D})$ with those already in the region $\mathbf{r}_i$ above a threshold ($s_i^{\max} > \tau_{sim}$) (we will discuss the effect of all thresholds in our results section). The growth continues until no other overlapping or neighboring superpoints are found. Our superpoint merging procedure, compared to using points alone or other merging strategies (see Tab. 7), produces more well-formed point cloud regions corresponding to 2D masks per frame.

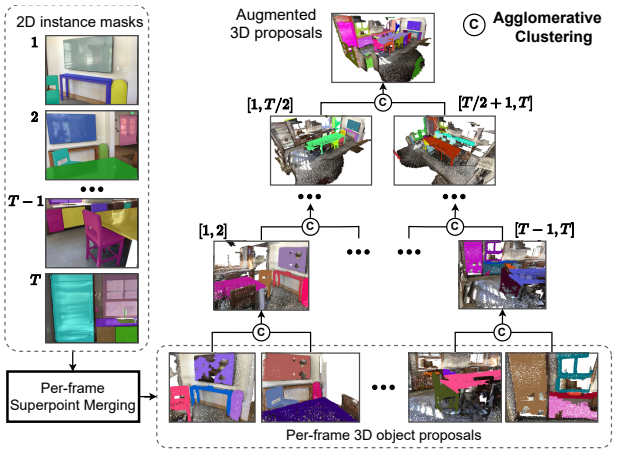


Figure 3. **2D-guide-3D Instance Proposal Module.** We generate initial 3D proposals using Per-frame Superpoint Merging, followed by hierarchical traversal across the RGB-D sequence to merge region sets between frames using Agglomerative clustering.

3D object proposal formation. To create 3D object proposals, one option is to utilize the point cloud regions obtained from the merging procedure across individual frames. However, this results in fragmented proposals, capturing only parts of object instances, as the regions correspond to 2D masks from single views (Fig. 2 - ②). To address this, we merge point cloud regions from different frames in a bottom-up manner, creating more complete and coherent 3D object masks. Agglomerative clustering combines region sets from pairs of frames until no compatible pairs remain. The resulting set includes merged and standalone regions, which can be matched with other region sets from subsequent frames. In the following paragraphs, we discuss three crucial design choices in this process: (a) the matching score between region pairs, (b) the matching process between sets of regions, and (c) the order of frames or

region sets used in matching and merging.

Matching score. For a pair of point cloud regions $(\mathbf{r}_i, \mathbf{r}_j)$, we define a matching score based on (a) feature similarity and (b) overlap degree. Their feature-based similarity s' is measured through cosine similarity between the regions' feature vectors $\mathbf{f}_i^{\text{3D}}$, or $s'_{i,j} = \cos(\mathbf{f}_i^{\text{3D}}, \mathbf{f}_j^{\text{3D}})$, which are in turn computed as the average of their point features. While this measures if the regions belong to the same object's shape, it may yield high similarity for duplicate instances with the same geometry. To address this, we also consider the degree of overlap, expressed as the IoU $o'_{i,j} = \text{IoU}(\mathbf{r}_i, \mathbf{r}_j)$ between the two regions $\mathbf{r}_i, \mathbf{r}_j$, which is expected to be high for overlapping regions of the same instance. Two regions are considered matching if their feature-based similarity and IoU score satisfy $s'_{i,j} > \tau_{sim}$ and $o'_{i,j} > \tau_{iou}$ (same thresholds used during per-frame superpoint merging). Our approach, incorporating matching scores based on point cloud deep features and geometric structures, results in more coherent and well-defined point cloud regions compared to other strategies (see Tab. 7).

Agglomerative clustering process. To merge region sets $\{\mathbf{r}_i\}_{i=1}^{I}$ and $\{\mathbf{r}_j\}_{j=1}^{J}$ from different frames into a unified set $\{\mathbf{r}_l\}_{l=1}^{L}$, where $L \leq I + J$, we employ Agglomerative clustering [46]. We begin by concatenating them into a single "active set" $\{\mathbf{r}_l\}_{l=1}^{I+J}$. We compute the each entry $c_{i,j}$ of the binary cost matrix $\mathbf{C}$ of size $(I + J) \times (I + J)$ as:

$$c_{i,j} = \mathbb{1}\left(o'_{i,j} > \tau_{iou}\right) \odot \mathbb{1}\left(s'_{i,j} > \tau_{sim}\right), \qquad (1)$$

where $\mathbb{1}(\cdot)$ is the indicator function, $\odot$ is the AND operator.

The agglomerative clustering procedure iteratively merges regions within the "active set" according to the cost matrix $\mathbf{C}$ and continues to update this matrix until no further merges are possible - indicated by the absence of any positive elements in $\mathbf{C}$.

Merging order. We explored two merging strategies: a *sequential* order, where region sets are merged between consecutive frames, and the resulting set is further merged with the next frame, and a *hierarchical* order, which involves merging region sets between non-consecutive frames in separate passes. The hierarchical approach forms a binary tree, with each level merging sets from consecutive pairs of the previous level (see Fig. 3). Details and performance analysis are presented in the Experiments section.

3.2. 3D Instance Segmentation Network

Network design. This network directly processes 3D point clouds to generate 3D object instance masks. We employ established 3D instance segmentation networks like Mask3D [57] and ISBNet [47] as our backbone. For each object candidate, the kernel computed from sampled points and their neighbors is convolved with point-wise features to predict the binary mask. In our open-vocabulary scenario, we exclude semantic labeling heads, focusing solely on the binary instance mask head. The output consists of K_2 binary masks in a $K_2 \times N$ binary matrix $\mathbf{M}_2$ (see Fig. 2 - ④).

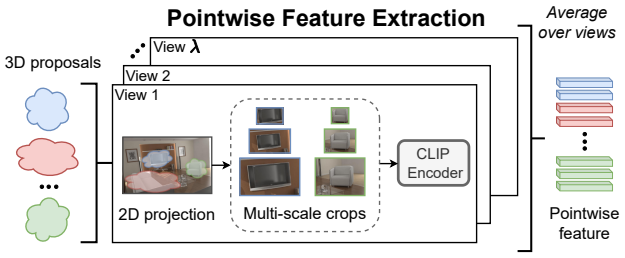


Figure 4. **Pointwise Feature Extraction**. Each 3D proposal is projected to top-λ views to obtain CLIP features in a multi-scale manner. The final proposal feature is averaged across views.

Combining object instance proposals. We simply append the proposals of set $\mathbf{M}_2$ to $\mathbf{M}_1$ to form the final set of K proposals $\mathbf{M}$ with the size of $K \times N$. Note that we apply NMS here to remove near-duplicate proposals with the overlapping IoU threshold τ_{dup}.

3.3. Pointwise Feature Extraction

In the final stage of our pipeline, we compute a feature vector for each 3D object proposal from our combined proposal set. This per-proposal feature vector serves various instance-based tasks, such as comparison with text prompts in the CLIP space [51]. Unlike prior open-vocabulary instance segmentation methods [61], which use a top-λ frame/view approach, we employ a more "3D-aware" pooling strategy. This strategy accumulates feature vectors on the point cloud, considering the frequency of each point's visibility in each view (see Fig. 4). *Our rationale is that points more frequently visible in the top-λ views should contribute more to the proposal's feature vector.*

Let $\mathbf{f}_{\lambda,k}^{\text{CLIP}} \in \mathbb{R}^{D^{\text{CLIP}}}$ be the 2D CLIP image feature of k-th instance in λ-th view, $\nu_\lambda \in \{0,1\}^N$ be the visibility map of view λ, and $\mathbf{m}_k^{\text{3D}} \in \{0,1\}^N$ be the k-th proposal binary mask in $\mathbf{M}$. We obtain the pointwise CLIP feature $\mathbf{F}^{\text{CLIP}} \in \mathbb{R}^{N \times D^{\text{CLIP}}}$ as:

$$\mathbf{F}^{\text{CLIP}} = \left\| \sum_k \left(\sum_\lambda (\nu_\lambda * \mathbf{f}_{\lambda,k}^{\text{CLIP}}) * \mathbf{m}_k^{\text{3D}} \right) \right\|_2, \qquad (2)$$

where $*$ is the element-wise multiplication (broadcasting if necessary) and $\|x\|_2$ is the L2 normalized vector of x.

The final classification score between a text query ρ and a 3D mask $\mathbf{m}_k^{\text{3D}}$ is the average cosine similarity score between its CLIP text embedding $\mathbf{e}_\rho$ and all points within the mask, particularly:

$$s_{k,\rho}^{\text{CLIP}} = \frac{1}{|\mathbf{m}_k^{\text{3D}}|} \sum_n \cos(\mathbf{F}^{\text{CLIP}} * \mathbf{m}_k^{\text{3D}}, \mathbf{e}_\rho), \qquad (3)$$

where $|\mathbf{m}_k^{\text{3D}}|$ is the number of points in the k-th mask.

4. Experiments

4.1. Experimental Setup

Datasets. We mainly conduct our experiments on the challenging dataset ScanNet200 [55], comprising 1,201 training and 312 validation scenes with 198 object categories. This dataset is well-suited for evaluating real-world open-vocabulary scenarios with a long-tail distribution. Additionally, we conduct experiments on Replica [59] (48 classes) and S3DIS [2] (13 classes) for comparison with prior methods [10, 11]. Replica has 8 evaluation scenes, while S3DIS includes 271 scenes across 6 areas, with Area 5 used for evaluation. We follow the categorization approach from [11] for S3DIS. Notably, we omit experiments on ScanNetV2 [7] due to its relative ease compared to ScanNet200 and identical input point clouds.

Evaluation metrics. We evaluate using standard AP metrics at IoU thresholds of 50% and 25%. Additionally, we calculate mAP across IoU thresholds from 50% to 95% in 5% increments. For ScanNet200, we report category group-specific AP_{head}, AP_{com}, and AP_{tail}.

Implementation Details. To process ScanNet200 and S3DIS scans efficiently, we downsampled the RGB-D frames by a factor of 10. Our approach utilizes the Grounded-SAM framework[1], integrating Grounding-DINO [43] and Segment Anything [31]. We employ the dataset class names as text prompts for generating 2D instance masks, followed by NMS with $\tau_{dup} = 0.5$ to handle overlapping instances. Our PyTorch implementation includes a 2D-guide-3D Instance Proposal Module, generating superpoints from [36, 52]. In Pointwise Feature Extraction, each proposal is projected into all viewpoints, and we select the top λ=5 views with the largest projected points. For CLIP, we use the *ViT-L/14* variant trained on OpenAI's WIT dataset [51].

4.2. Comparison to prior work

Setting 1: ScanNet200. The quantitative evaluation of the ScanNet200 dataset is summarized in Tab. 2. Following [61], we utilize the class-agnostic 3D proposal network trained on the ScanNet200 training set, then test the OV-3DIS on the validation set. Employing our 2D-Guided-3D Instance Proposal Module, Open3DIS achieves 18.2 and 19.2 in AP and AP_{tail}. We outperform OVIR-3D [44] and OpenMask3D [61] by margins of $+5.2$ and $+2.8$ in AP, and surpass all other methods, even the fully-supervised approaches in the AP_{tail} metric. This emphasizes the effectiveness of our 2D-Guide-3D Instance Proposal Module, which is effective in crafting precise 3D instance masks independently of any 3D models. Combining with class-agnostic 3D proposals from ISBNet boosts our performance to 23.7, 29.4, and 32.8 in AP, AP_{50}, and AP_{25} — reflecting a 1.5x enhancement in AP compared to prior methods. Impressively, our method competes closely with fully supervised techniques, attaining approximately 96% and 88% of the AP scores of ISBNet and Mask3D, and excelling in the AP_{com} and AP_{tail}. This performance underscores the advantages of merging 2D and 3D proposals and demonstrates our model's adeptness at segmenting rare objects.

To assess the generalizability of our approach, we conducted an additional experiment where the class-agnostic 3D proposal network is substituted with the one trained solely on the ScanNet20 dataset. We then categorized the ScanNet200 instance classes into two groups: the *base* group, consisting of 51 classes with semantics similar to ScanNet20 categories, and the *novel* group of the remaining classes. We report the AP_{novel}, AP_{base}, and AP in Tab. 3. Our proposed Open3DIS achieves superior performance compared to PLA [11], OpenMask3D [61], with large margins in both *novel* and *base* classes. Notably, PLA [11], trained with contrastive learning techniques, falls in a setting with hundreds of novel categories.

Setting 2: Replica. We further evaluate the zero-shot capability of our method on the Replica dataset, with results detailed in Tab. 4. Considering that several Replica categories share semantic similarities with ScanNet200 classes, to maintain a truly zero-shot scenario, we omitted the class-agnostic 3D proposal network for this dataset (using proposals from 2D only). Under this constraint, our approach still outperforms OpenMask3D [61] and OVIR-3D [44] by margins of $+5.0$ and $+7.0$ in AP, respectively.

Setting 3: S3DIS. In line with the setting of PLA [11], we trained a fully-supervised 3DIS model on the *base* classes of the S3DIS dataset, followed by testing the model on both *base* and *novel* classes. The results are shown in Tab. 5, where we report the performance in terms of AP^{B}_{50} and AP^{N}_{50}, representing the AP_{50} for the *base* and *novel* categories, respectively. Open3DIS significantly outperforms existing methods in AP^{N}_{50}, achieving more than double their scores. This remarkable performance underscores the efficacy of our approach in dealing with unseen categories, with the support of the 2D foundation model.

Our qualitative results with arbitrary text queries. We visualize the qualitative results of text-driven 3D instance segmentation in Fig. 5. Our model successfully segments instances based on different kinds of input text prompts, involving object categories that are not present in the labels, object's functionality, object's branch, and other properties.

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

Method	Setting	3D Proposal	AP	AP_{50}	AP_{25}	AP_{head}	AP_{com}	AP_{tail}
ISBNet [47]	Fully-sup		24.5	32.7	37.6	38.6	20.5	12.5
Mask3D [57]			26.9	36.2	41.4	39.8	21.7	17.9
OpenScene [48] + DBScan [13]†	Open-vocab	None	2.8	7.8	18.6	2.7	3.1	2.6
OpenScene [48] + Mask3D [57]		Mask3D [57]	11.7	15.2	17.8	13.4	11.6	9.9
SAM3D† [73]		None	6.1	14.2	21.3	7.0	6.2	4.6
OVIR-3D† [44]		None	13.0	24.9	32.3	14.4	12.7	11.7
OpenMask3D [61]		Mask3D [57]	15.4	19.9	23.1	17.1	14.1	14.9
Ours (only 2D)	Open-vocab	None	18.2	26.1	31.4	18.9	16.5	19.2
Ours (only 3D)		ISBNet [47]	18.6	23.1	27.3	24.7	16.9	13.3
Ours (2D and 3D)		ISBNet [47]	**23.7**	**29.4**	**32.8**	**27.8**	**21.2**	**21.8**

Table 2. OV-3DIS results on **ScanNet200**. Methods with † are adapted and evaluated on ScanNet200. Our proposed method achieves the highest AP, outperforming previous methods in all metrics. The best results are in **bold** while the second best results are underscored.

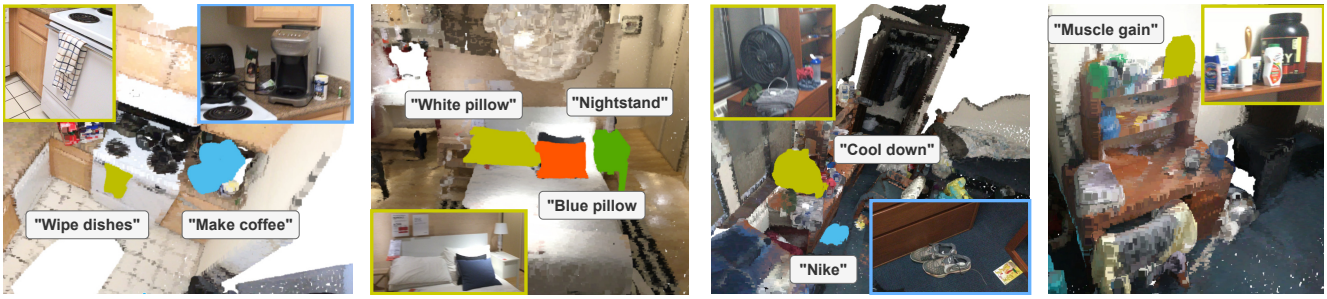


Figure 5. Qualitative results of our method on zero-shot instance segmentation. We query instance masks using arbitrary text prompts involving object categories that are not present in the ScanNet200 labels. For each scene, we showcase the instance that has the highest similarity score to the query's embedding. These visualizations underscore the model's zero-shot capability, as it successfully identifies and segments objects that were never encountered during the training phase.

Method	Pretrain	AP_{novel}	AP_{base}	AP
OpenMask3D	ScanNet200	15.0	16.2	15.4
Ours		22.6	26.7	23.7
PLA (Base 15)	ScanNet20	0.3	10.8	3.2
PLA (Base 20)		0.3	15.8	4.5
OpenScene + Mask3D		7.6	11.1	8.5
OpenMask3D		11.9	14.3	12.6
Ours		**16.5**	**25.8**	**19.0**

Table 3. OV-3DIS results on **ScanNet200** dataset, using the class-agnostic 3D proposal network trained on ScanNet20.

Method	3D Proposal	AP	AP_{50}	AP_{25}
OpenScene + Mask3D	Mask3D	10.9	15.6	17.3
OpenMask3D	Mask3D	13.1	18.4	24.2
OVIR-3D †	None	11.1	20.5	27.5
Ours	None	**18.1**	**26.7**	**30.5**

Table 4. OV-3DIS results on **Replica** dataset.†We adopt the source code of [44] to this dataset.

4.3. Ablation study

To validate the design choices of our proposed method, we have carried out a series of ablation studies on the validation set of the ScanNet200 dataset.

Method	B8/N4		B6/N6	
	AP^B_{50}	AP^N_{50}	AP^B_{50}	AP^N_{50}
LSeg-3D [11]	58.3	0.3	41.1	0.5
PLA [11]	59.0	8.6	46.9	9.8
Lowis3D [10]	58.7	13.8	**51.8**	15.8
Ours	**60.8**	**26.3**	50.0	**29.0**

Table 5. OV-3DIS results on S3DIS in terms of AP^B_{50} and AP^N_{50}.

Setting	AP	AP_{head}	AP_{com}	AP_{tail}
A1: OpenScene (distill)	3.3	5.5	2.4	1.7
A2: OpenScene (fusion)	17.5	21.5	17.1	13.3
A3: OpenScene (ensemble)	5.6	6.4	4.8	5.7
B: Mask-wise Feature	22.2	25.9	19.3	21.4
C: Point-wise Feature	**23.7**	**27.8**	**21.2**	**21.8**

Table 6. Comparing between extracting per-mask and per-point features for classification using Open3DIS instance proposal set.

Study on different kinds of features for open-vocabulary classification is presented in Tab. 6. In the first three rows (setting A1-A3), we employ the pointwise feature map extracted by OpenScene [48] to perform classification on our

Use Superpoint	Filtering Cond.	AP	AP_{head}	AP_{com}	AP_{tail}
✓	Deep. Feature	**18.2**	**18.9**	**16.5**	**19.2**
✓	None	15.9	16.5	14.3	17.0
✓	Euclid Dist.	16.0	16.4	14.1	17.6
	None	12.0	12.6	11.2	12.2

Table 7. Ablation on different configurations of the 2D-G-3DIP.

Merging Strat.	Merging Ord.	AP	AP_{head}	AP_{com}	AP_{tail}
Hungarian	Sequential	13.2	13.9	11.3	14.7
Hungarian	Hierarchical	16.1	16.1	13.3	19.4
Agglomerative	Sequential	16.9	17.8	16.1	18.0
Agglomerative	Hierarchical	**18.2**	**18.9**	**16.5**	**19.2**

Table 8. Ablation on different merging configurations.

3D Seg.	AP	AP_{head}	AP_{com}	AP_{tail}
Mask3D [57]	**23.7**	26.4	**22.5**	**21.9**
ISBNet [47]	**23.7**	**27.8**	21.2	21.8

Table 9. Ablation on different 3D segmenters.

3D proposals. Of these, the *fusion* approach, which directly projects CLIP features from 2D images onto the 3D point cloud, yields the highest results, 17.5 in AP. In setting B, we adopt a strategy akin to [61], extracting features for each mask by projecting the 3D proposals onto the top-λ views, which attains an AP of 22.2. Surpassing these, our Point-wise Feature Extraction (setting C) achieves the best AP score of 23.7, substantiating our design choice.

Study on the 2D-guide-3D Instance Proposal Module is in Tab. 7. Our proposed approach (row 1), utilizing superpoints to merge 3D points into regions and filter outliers based on cosine similarity in feature space, achieves an AP of 18.2. Disabling this filtering notably reduces AP by 2.3. Comparatively, a more basic method (row 3) relying on Euclidean distance to eliminate outlier superpoints yields an AP of 16.0, showing the lesser effectiveness of Euclidean distance for noise filtering. Our baseline (last row), grouping 3D points solely based on 2D masks, significantly decreases AP to 12.0, underscoring the necessity of superpoint merging for effective 3D proposal creation.

We study different merging configurations, including *merging strategy* and *merging order* in Tab. 8. We compare the proposed Agglomerative clustering and the Hungarian matching. Specifically, we first establish a partial matching between two sets of regions, then matched pairs are merged into new refined regions, and unmatched ones remain the same. Using Hungarian matching yields inferior results relative to agglomerative clustering, with a drop of $\sim$2.0 in AP. Adopting the sequential merging order leads to a slight decrease by $\sim$1.0 in AP in performance. The best results are achieved when agglomerative clustering is paired with the hierarchical merging order.

2D Seg.	AP	AP_{head}	AP_{com}	AP_{tail}
SEEM [83]	21.5	26.5	19.6	18.0
ODISE [70]	21.6	26.0	19.5	19.1
Detic [81]	22.2	26.8	20.0	19.2
Grounded-SAM	**23.7**	**27.8**	**21.2**	**21.8**

Table 10. Ablation on different 2D segmenters.

τ_{iou} threshold	0.3	0.5	0.7	0.9	0.95
AP	17.7	17.8	18.0	**18.2**	16.9
$\mathbf{AP}_{50}$	25.4	25.8	25.9	**26.1**	24.1

Table 11. Ablation on τ_{iou}.

τ_{sim} threshold	0.5	0.7	0.8	0.9	0.95
AP	14.2	14.6	17.2	**18.2**	16.2
$\mathbf{AP}_{50}$	21.0	21.8	25.1	**26.1**	23.8

Table 12. Ablates τ_{sim}.

View Selection	Top 1	Top 5	Top 10	Top 20	All
AP	21.2	**23.7**	22.6	22.5	22.5
$\mathbf{AP}_{50}$	27.3	**29.4**	28.7	29.0	29.1

Table 13. Ablation on top-λ view selection.

Ablation Study on Segmenters. Our comparative analysis of various *class-agnostic 3D segmenters* and *open-vocabulary 2D segmenters* is presented in Tab. 9 and 10. The findings reveal that utilizing either ISBNet [47] or Mask3D [57] leads to similar levels of performance, achieving an AP of 23.7. Incorporating 2D instance masks from SEEM [83], Detic [81] or ODISE [70] leads to a slight decrease in AP by $\sim$1.4, which we attribute to the less refined outputs produced by these models.

Ablation study on different values of visibility threshold and similarity threshold. We report the performance of our version using only proposals from the 2D-G-3DIP with different values of the visibility threshold and similarity threshold in Tab. 11 and 12. Using τ_{iou}=0.9 and τ_{sim}=0.9 yields optimal results.

Study on different values of viewpoints is illustrated in Tab. 13. Relying only on the viewpoint with the highest number of projected points reduces the AP score to 21.2. Conversely, raising the number of views to 10 or more also yields worse results, likely due to the presence of inferior, occluded 2D masks. λ=5 reports the best performance.

5. Discussion

Limitations. Our Class-agnostic 3D Proposal and 2D-guided-3D Instance Proposal Module currently operate independently, with their outputs being combined to obtain

the final 3D proposal set. A better-integrating strategy, where these modules enhance each other's performance in a synergistic fashion, would be an interesting future direction.

Conclusions. This paper has introduced Open3DIS, the pioneering method that combines 3D class-agnostic proposals with 2D open-vocabulary segmentation to tackle open-vocabulary 3D instance segmentation. Our 2D-Guided-3D Instance Proposal Module generates top-tier 3D proposals by integrating 2D instance masks from RGB-D images. We significantly outperform existing OV-3DIS approaches across benchmarks (around 50% on ScanNet200, 80% on S3DIS, and 40% on Replica) and exhibit comparable performance to fully supervised approaches like ISBNet or Mask3D. Additionally, our method excels in segmenting objects in 3D spaces based on diverse textual descriptions, unlocking new capabilities for machine comprehension and interaction within intricate 3D environments.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 2

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 6

[3] 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. 15, 22

[4] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *arXiv preprint arXiv:2310.02960*, 2023. 2

[5] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 2

[6] 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. 13

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

[8] 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. 13

[9] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[10] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *arXiv preprint arXiv:2308.00353*, 2023. 2, 3, 6, 7

[11] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 6, 7

[12] Shichao Dong, Guosheng Lin, and Tzu-Yi Hung. Learning regional purity for instance segmentation on 3d point clouds. In *European Conference on Computer Vision*, pages 56–72. Springer, 2022. 2

[13] 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. 7

[14] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 4, 13

[15] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 13

[16] 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. 13

[17] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023. 2

[18] 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. 13

[19] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19498–19507, 2023. 2

[20] Tong He, Chunhua Shen, and Anton van den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 354–363, 2021. 2

[21] Tong He, Wei Yin, Chunhua Shen, and Anton van den Hengel. Pointinst3d: Segmenting 3d instances by points. In *Computer Vision–ECCV 2022: 17th European Conference,*

Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, pages 286–302. Springer, 2022. 2

[22] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313*, 2023. 3

[23] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2023. 2

[24] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 2

[25] Le Hui, Linghua Tang, Yaqi Shen, Jin Xie, and Jian Yang. Learning superpoint graph cut for 3d instance segmentation. In *Advances in Neural Information Processing Systems*, 2022. 2

[26] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 2

[27] 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. 2

[28] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 2

[29] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, 2023. 2

[30] 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. 13

[31] 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. 6, 13

[32] 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. 3

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[34] 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. 13

[35] 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. 13

[36] 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. 6, 13

[37] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[38] 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. 13

[39] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[40] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2783–2792, 2021. 2

[41] 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. 13

[42] Jiaheng Liu, Tong He, Honghui Yang, Rui Su, Jiayi Tian, Junran Wu, Hongcheng Guo, Ke Xu, and Wanli Ouyang. 3d-queryis: A query-based framework for 3d instance segmentation. *arXiv preprint arXiv:2211.09375*, 2022. 2

[43] 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. 2, 4, 6, 13

[44] 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, 2, 3, 6, 7, 13, 14, 15

[45] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[46] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011. 5

[47] 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. 2, 4, 5, 7, 8, 13

[48] 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. 2, 3, 7, 15

[49] Chau Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. *arXiv preprint arXiv:2310.17109*, 2023. 2

[50] 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. 13

[51] 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. 2, 3, 5, 6

[52] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. *arXiv preprint arXiv:2306.08045*, 2023. 6, 13

[53] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 13

[54] 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. 13

[55] 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. 1, 2, 3, 6, 15, 22

[56] 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. 13

[57] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 2, 4, 5, 7, 8

[58] 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. 13

[59] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 6

[60] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 2

[61] 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, 3, 5, 6, 7, 8, 13, 15

[62] Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M Patel, and Ran Xu. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23539–23549, 2023. 2

[63] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on 3d point clouds. In *CVPR*, 2022. 2

[64] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023. 2

[65] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 2

[66] Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. *arXiv preprint arXiv:2301.00805*, 2023. 2

[67] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 2

[68] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 235–252. Springer, 2022. 2

[69] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2

[70] 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. 8, 13

[71] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Advances in Neural Information Processing Systems*, pages 6737–6746, 2019. 2

[72] Jihan Yang, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. *arXiv preprint arXiv:2304.00962*, 2023. 3

[73] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2, 3, 7, 14, 15

[74] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 2

[75] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 2

[76] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 2022. 2

[77] Cheng Zhang, Haocheng Wan, Shengqiang Liu, Xinyi Shen, and Zizhao Wu. Pvt: Point-voxel transformer for 3d deep learning. *arXiv preprint arXiv:2108.06076*, 2021. 2

[78] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 2

[79] Zhuowen Tu Zheng Ding, Jieke Wang. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning*, 2023. 2

[80] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2

[81] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 8, 13

[82] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*, 2023. 2

[83] 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. 2, 8, 13

6. Implementation Details

6.1. Class-agnostic 3D Segmenter

We adopt the architecture from ISBNet [47] 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 [16], utilizing 3D sparse convolutions [15], 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 [36, 52] 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.

6.2. Open-Vocabulary 2D Segmenter

In this study, we employ four 2D open-vocabulary instance segmenters: Grounded-SAM[2], DETIC [81], SEEM [83], and ODISE [70]. Here is a breakdown of how each of these segmenters is utilized:
(a) *For Grounded-SAM*, we utilize the Swin-B Grounding DINO decoder [43], which has been pretrained on various datasets including COCO [41], O365 [58], GoldG [34, 50], OpenImage [35], ODinW-35 [38], and RefCOCO [30]. 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 [31] 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 [44] to use the Swin-B model pretrained on the ImageNet-21K dataset [8] 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 [18], 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 [54] pre-trained on a subset of the LAION [56] dataset, along with Mask2Former [6] serving as the mask generator. We set the confidence threshold to 0.5.

6.3. S3DIS and Replica Datasets

(a) *For the S3DIS dataset*, which lacks original mesh data, we apply the superpoint-graph method from the Superpoint Transformer [53] 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[3] based on Felzenszwalb and Huttenlocher's efficient graph-based image segmentation method [14] to create superpoints. The ground-truths for semantic and instance segmentation are provided by [61].

6.4. 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 procedure across individual frames $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T\}$, the algorithm merges these independently fragmented regions (see Fig. 6) into well-formed ones recursively, resulting in high-quality augmented 3D proposals.

6.5. 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:

$$\Pi_t = \mathbf{K}_t \cdot [\mathbf{R}|\mathbf{c}]_t. \tag{4}$$

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:

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

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

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$(1, T)$

$$z_i^{(2d)} \cdot \begin{bmatrix} x_i^{(2d)} \\ y_i^{(2d)} \\ 1 \end{bmatrix} = \Pi_t \cdot \begin{bmatrix} x_i^{(3d)} \\ y_i^{(3d)} \\ z_i^{(3d)} \\ 1, \end{bmatrix} \tag{5}$$

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 τ_{depth}:

$$|z_i^{(2d)} - \mathbf{D}_t[\lfloor y_i^{(2d)} \rfloor, \lfloor x_i^{(2d)} \rfloor]| > \tau_{depth} \tag{6}$$

7. Additional Analysis

Ablation study on the depth threshold τ_{depth} is reported in Tab. 14. Overall, using $\tau_{depth} = 0.1$ gives the best performance.

Ablation study on the subsampling factors of RGB-D images is shown in Tab. 15. 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. in AP scores. Reducing the number of images too much yields worse results.

τ_{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 14. Ablation on the depth threshold τ_{depth}.

Subsampling factor	**AP**	**AP**$_{\text{head}}$	**AP**$_{\text{com}}$	**AP**$_{\text{tail}}$
10 (default)	**18.2**	**18.9**	16.5	19.2
20	17.9	17.9	16.5	**19.6**
40	17.4	17.3	**16.7**	18.5
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 15. Study on the subsampling factors of RGB-D images.

8. Qualitative Results

8.1. 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. 7, SAM3D [73] masks are dispersed over a wide area, while OVIR-3D [44] 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.

8.2. 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. 8.

Despite the potential imperfections introduced by Depth sensors, previous methods [44, 73] 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. 10.

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

8.3. More Qualitative Results on ScanNet200, Replica, and S3DIS

ScanNet200. We present visualizations of Open3DIS applied to the extensive ScanNet200 dataset. In Fig. 9, 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. 10 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. 11a.

S3DIS. The qualitative results of our approach on the S3DIS dataset are visualized in Fig. 11b.

8.4. Open-Vocabulary Scene Exploration

We showcase the remarkable Open-Vocabulary scene exploration capabilities of Open3DIS on the ARKitScenes [3] (Fig. 12a) and ScanNet200 [55] (Fig. 12b) 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 [44, 48, 61, 73] significantly. Some of these methods struggle to detect these objects, let alone locate them accurately. *Please see the supplementary video for a live demo.*

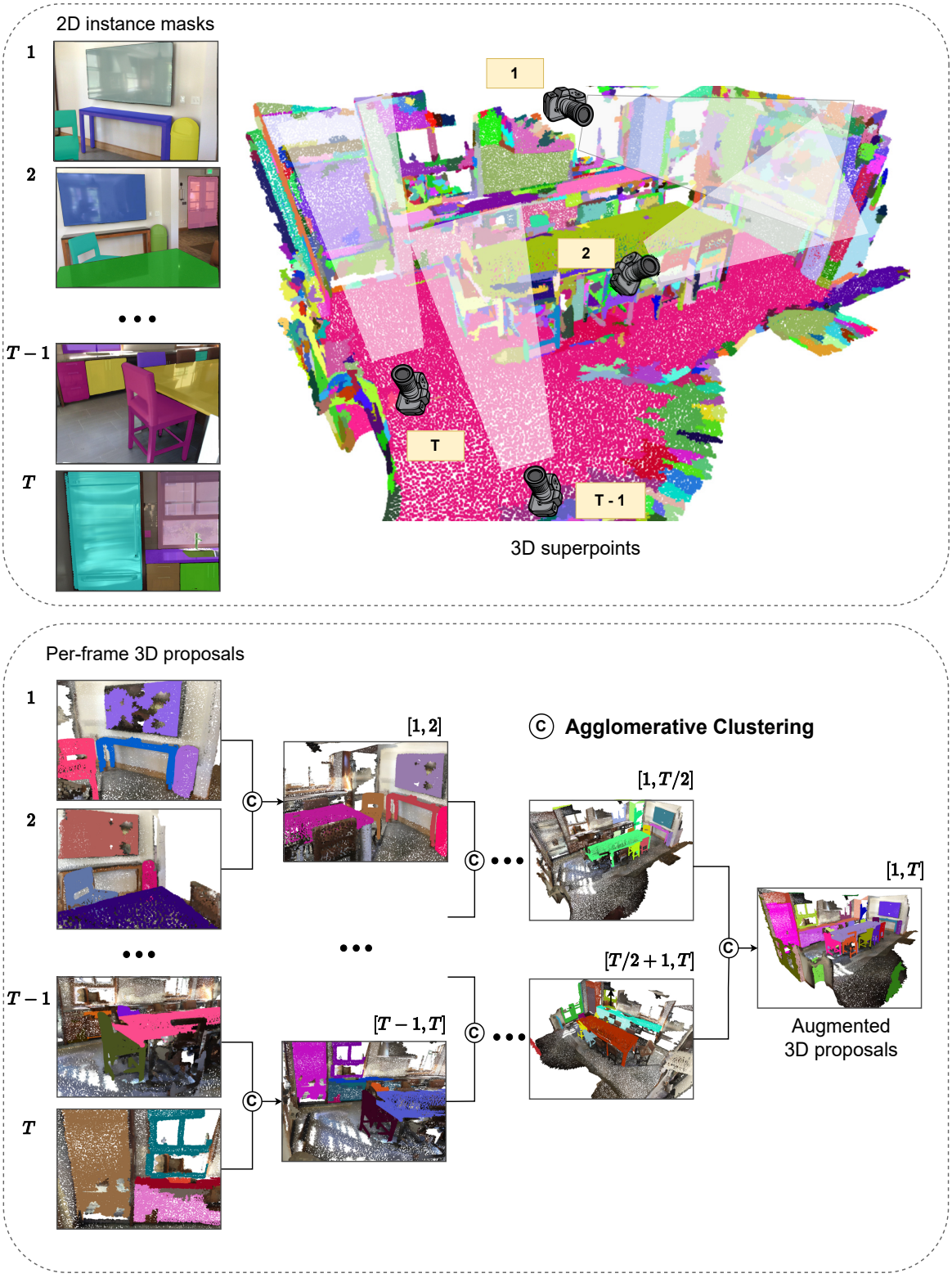


Figure 6. (Top) The 2D-G-3DIP module utilizes 2D per-frame instance masks to generate per-frame 3D proposals by leveraging 3D superpoints. (Bottom) Our proposed hierarchical merging. These proposals are considered point cloud regions and undergo a hierarchical merging process across multiple views, resulting in the final Augmented 3D proposals (Best viewed in color).

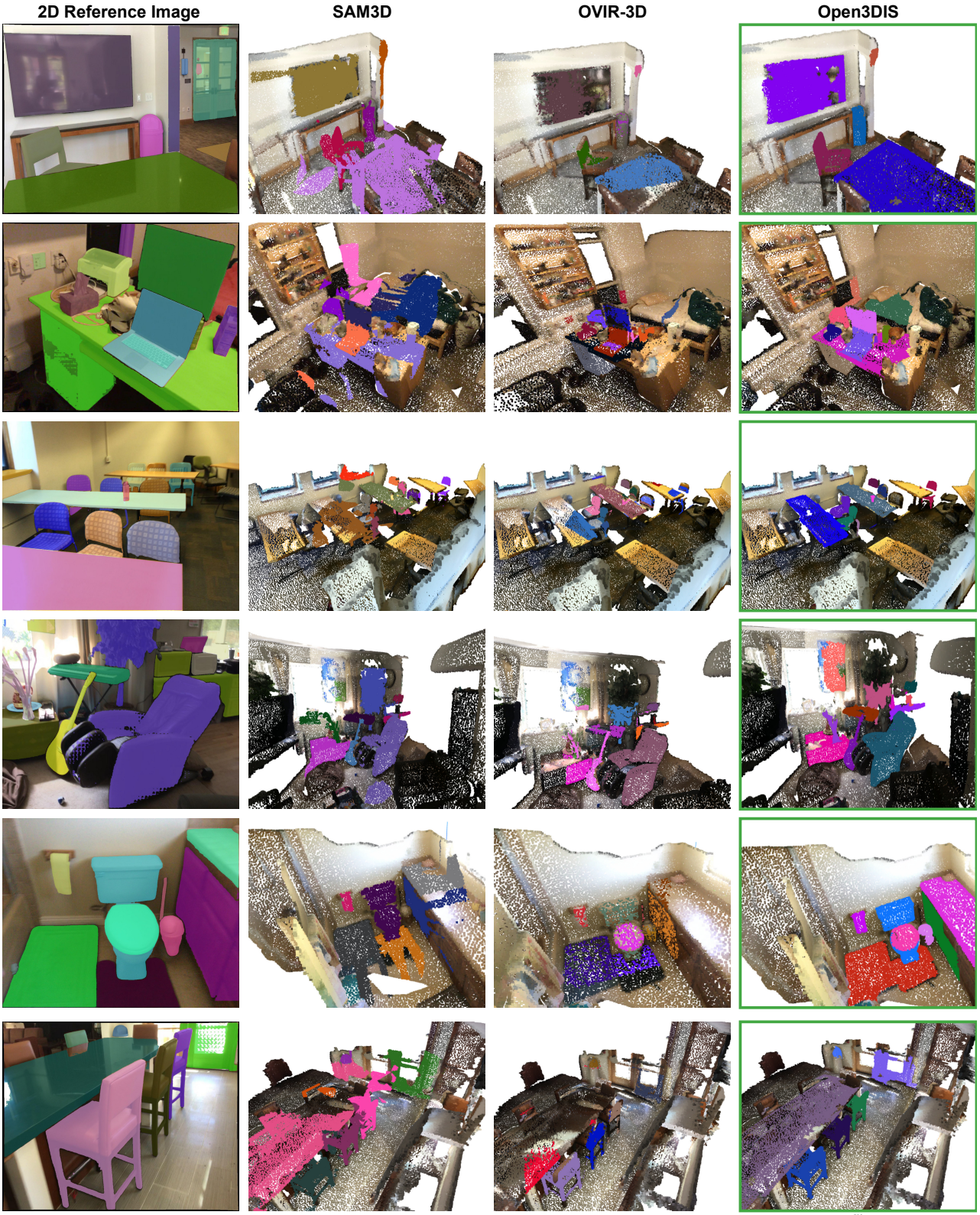


Figure 7. Qualitative results of our method compared to others in Constructing 3D proposals from 2D masks of an image. Each row shows one example, including the input 2D reference image, other 2D lifting methods, and our Open3DIS (**only 2D**) (Best viewed in color).

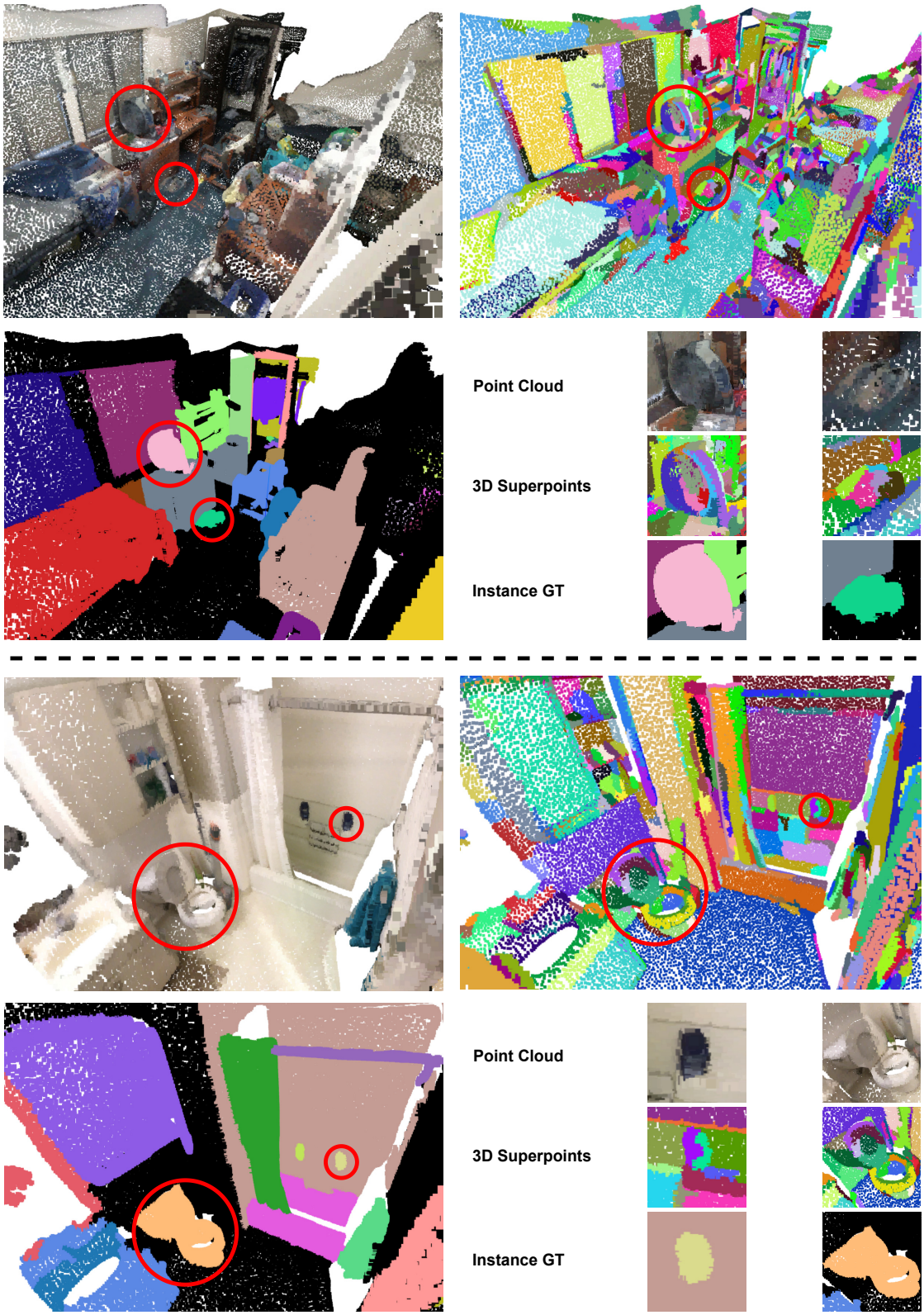


Figure 8. Two examples (separated by the dashed line) illustrating the reason for using the 2D-G-3DIP module when creating point cloud regions, with a focus on accurately covering object instances indicated by the Red circles (Best viewed in color).

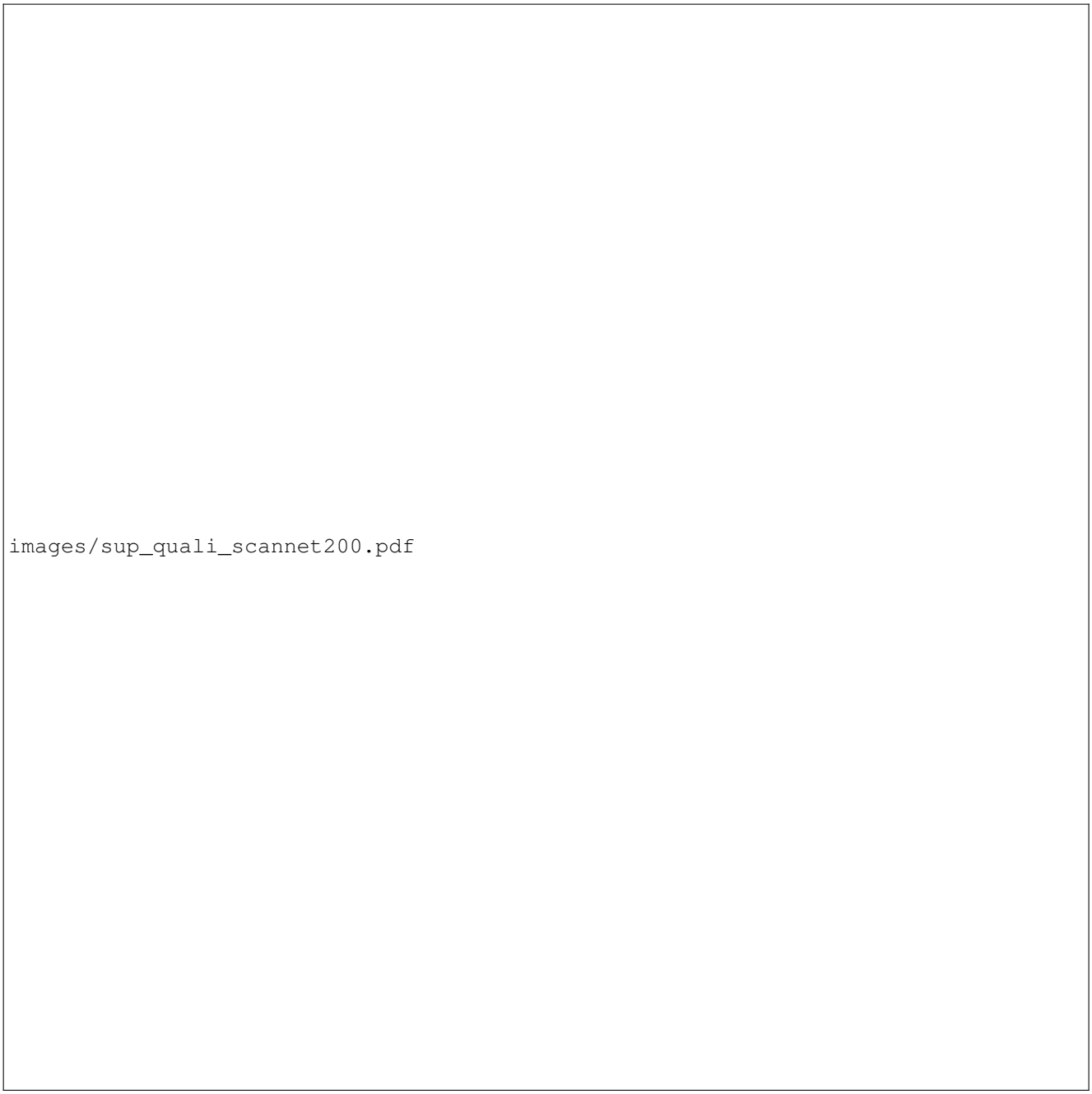


Figure 9. Qualitative results of our method on the ScanNet200 dataset. Each row shows one example, including the input RGB point cloud, instance ground truth, and our predictions (Best viewed in color).

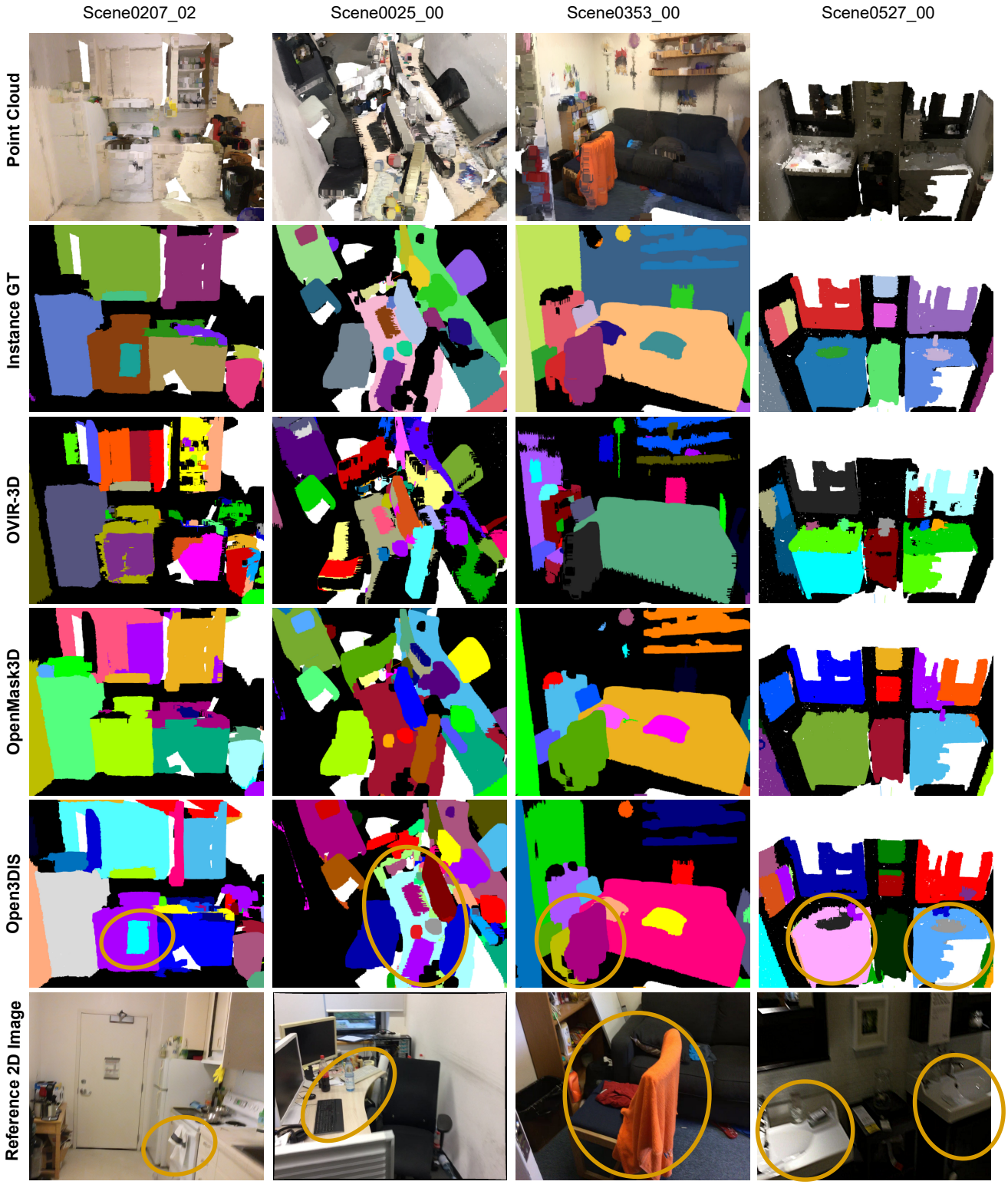
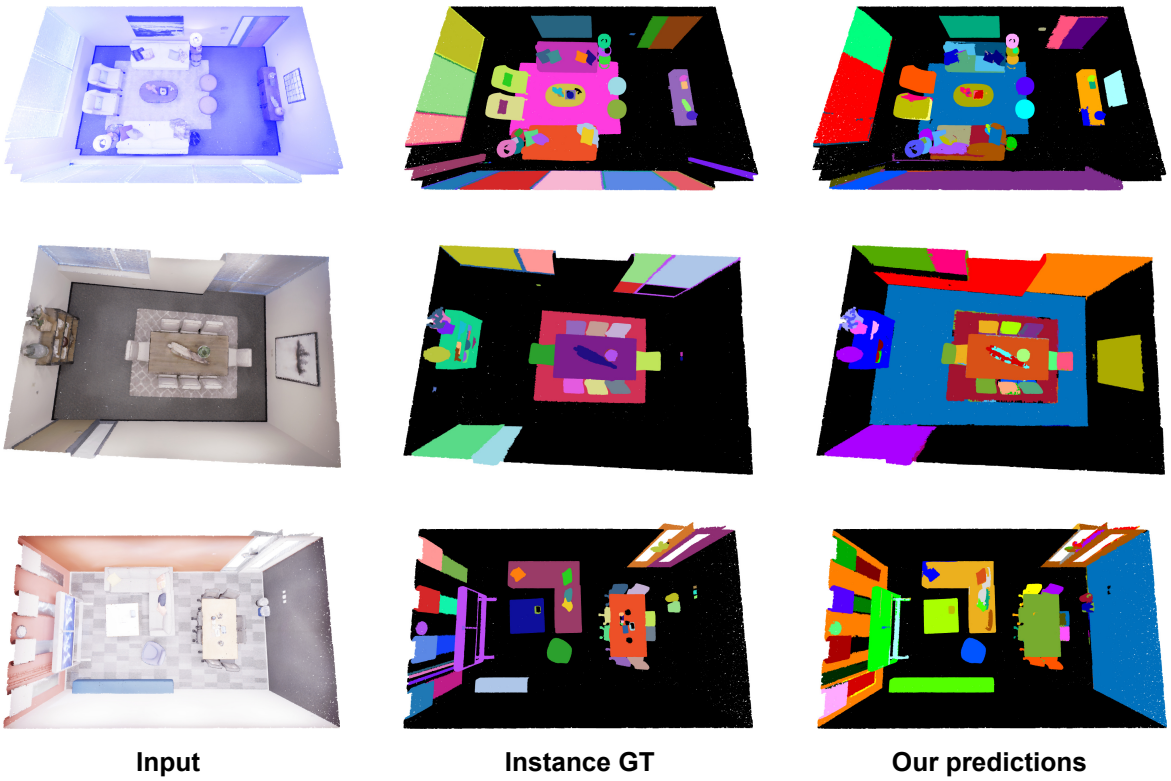
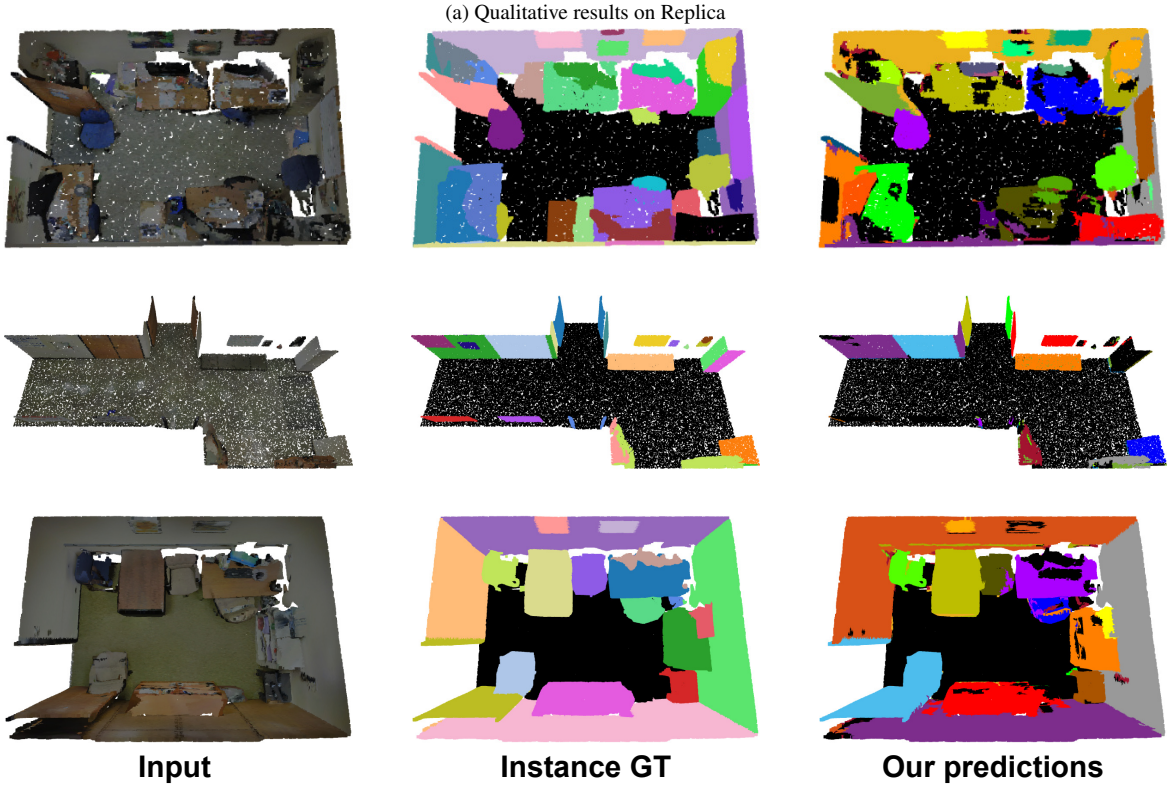


Figure 10. Qualitative results of our method compared to others on ScanNet200 dataset. Each column shows one example in Orange ellipses demonstrating that Open3DIS performs better than others (Best viewed in color).

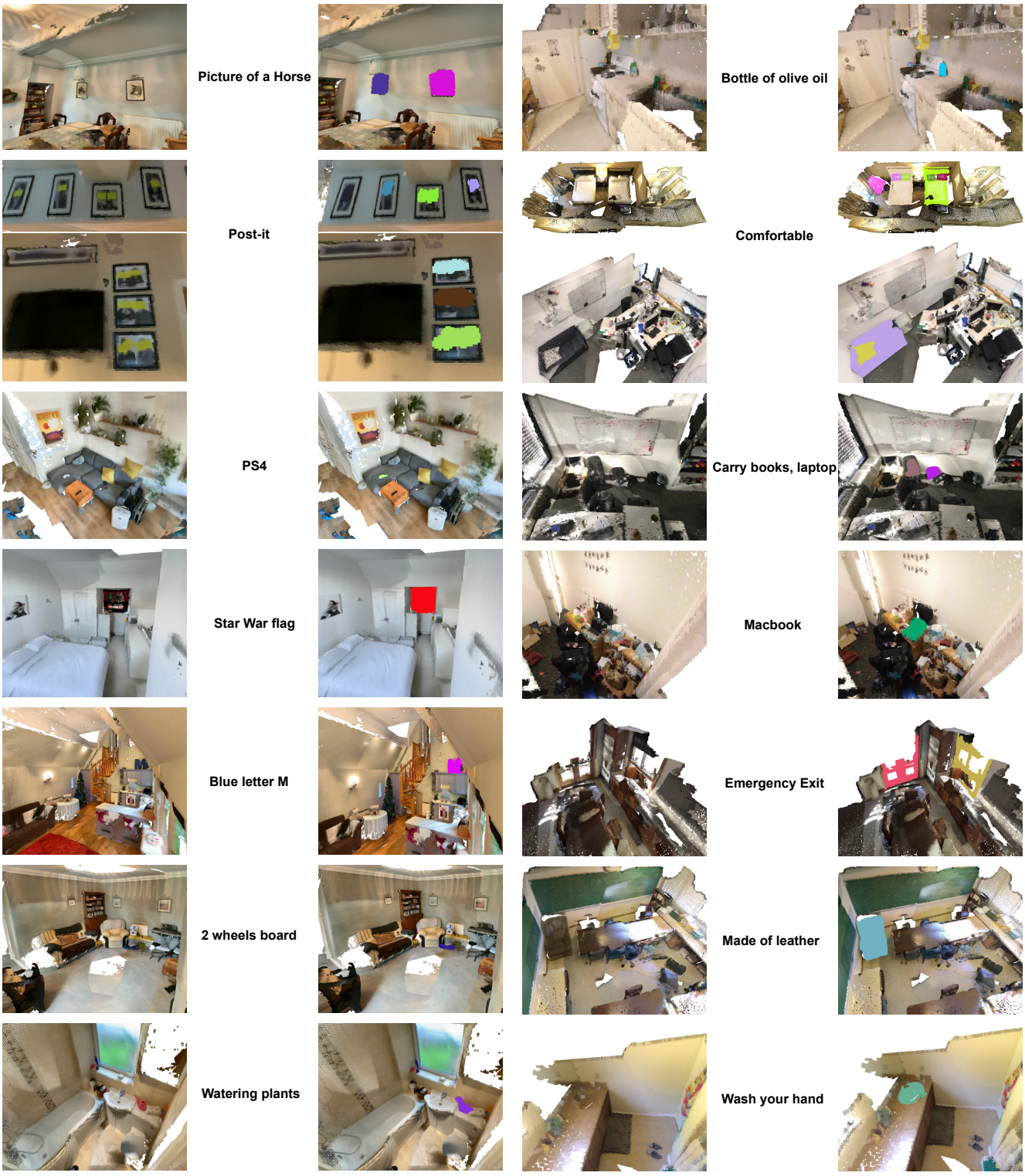


(a) Qualitative results on Replica



(b) Qualitative results on S3DIS

Figure 11. 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 12. Open-Vocabulary exploration on **ARKitScenes** [3] (Left) and **ScanNet200** [55] (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).