Details Matter for Indoor Open-vocabulary 3D Instance Segmentation

Sanghun Jung[1,2†] Jingjing Zheng[2] Ke Zhang[2] Nan Qiao[2] Albert Y. C. Chen[2] Lu Xia[2] Chi Liu[2]
Yuyin Sun[2] Xiao Zeng[2] Hsiang-Wei Huang[1] Byron Boots[1] Min Sun[2,3] Cheng-Hao Kuo[2]
[1]University of Washington [2]Amazon Lab126 [3]National Tsing Hua University

Abstract

Unlike closed-vocabulary 3D instance segmentation that is often trained end-to-end, open-vocabulary 3D instance segmentation (OV-3DIS) often leverages vision-language models (VLMs) to generate 3D instance proposals and classify them. While various concepts have been proposed from existing research, we observe that these individual concepts are not mutually exclusive but complementary. In this paper, we propose a new state-of-the-art solution for OV-3DIS by carefully designing a recipe to combine the concepts together and refining them to address key challenges. Our solution follows the two-stage scheme: 3D proposal generation and instance classification. We employ robust 3D tracking-based proposal aggregation to generate 3D proposals and remove overlapped or partial proposals by iterative merging/removal. For the classification stage, we replace the standard CLIP model with Alpha-CLIP, which incorporates object masks as an alpha channel to reduce background noise and obtain object-centric representation. Additionally, we introduce the standardized maximum similarity (SMS) score to normalize text-to-proposal similarity, effectively filtering out false positives and boosting precision. Our framework achieves state-of-the-art performance on ScanNet200 and S3DIS across all AP and AR metrics, even surpassing an end-to-end closed-vocabulary method.

1. Introduction

The task of OV-3DIS [2, 8, 9, 22, 39, 49, 57, 59, 63] aims to predict 3D masks for individual objects in a 3D point cloud scene given open-vocabulary text queries (Fig. 1). OV-3DIS has diverse applications across domains, such as robotics, augmented reality, scene understanding, and 3D visual search. For example, in robotic tasks like indoor navigation and object manipulation, interpreting open-vocabulary queries and localizing corresponding objects in a 3D environment are crucial for effective performance.

Efforts have been made to tackle the task of OV-3DIS. A two-staged paradigm has been widely adopted across various works [2, 39, 49, 57]. They first generate the class-agnostic 3D proposals and then classify the pre-

Figure 1. **Examples of open-vocabulary predictions from our method in the ScanNet200 dataset [7].** Our method effectively retrieves instances based on functional descriptions (*e.g.*, drink water, heat mac & cheese) and object attributes (*e.g.*, red chair).

dicted proposals into open-vocabulary queries. While some works [2, 49] directly generate 3D proposals from point clouds using pretrained 3D networks [38, 45], other approaches [39, 57, 63] generate 3D proposals from images. For 3D proposal generation from images, they leverage vision foundation models (VFMs) [33, 43, 66] to ground object regions in each image frame. Each object region is then lifted to 3D point clouds and temporally aggregated across frames to find complete 3D masks.

While there have been various efforts based on this scheme, e.g., agglomerative clustering [39], progressive region growing [63], and graph clustering [57], these individual concepts are not mutually exclusive but complementary. This paper carefully combines the concepts and refines each step to address key challenges, achieving state-of-the-art (SoTA) performance in existing benchmarks. We generate 3D proposals from both images (i.e., image-based 3D proposals) and aggregated point clouds (i.e., point cloud-based 3D proposals). Image-based 3D proposal generation [37, 39, 63] involves many design choices in three steps: 1) frame-wise 2D object grounding, 2) lifting 2D predictions to 3D point clouds, and 3) 3D proposal aggregation across frames to find complete 3D masks. Finally, CLIP-based models [42] are used to classify the 3D proposals [39, 49, 57].

While we adopt this general paradigm, we refine each stage to effectively handle *missing details* in the existing literature. Also, we devise an additional iterative merging and removal step at the end of the proposal generation to suppress overlapped or partial proposals.

2D Object Grounding. We observe two representative types of wrong object predictions from VFMs: masks cover-

† The work is done during the internship at Amazon Lab126.

ing multiple objects and partial masks. While partial masks can be mitigated in later steps by merging or removal, wrong masks covering multiple instances can hardly be separated into individual instances. Thus, we sort 2D predictions in each frame by their size and remove the overlapped regions from the larger ones to minimize such cases.

2D to 3D Lifting. Following existing works [39, 60, 63], we use 3D superpoints [13] as a basic unit of point cloud operations. We aim to find a set of 3D superpoints corresponding to each 2D instance in the lifting step. We adopt two concepts from existing work [39]: frame-wise and instance-wise visibility scores to remove unconfident superpoints.

Tracking-based 3D Proposal Aggregation. We progressively enlarge the lifted 3D superpoints of instances by tracking them sequentially, analogous to OVIR-3D [37]. However, ours has unique features to improve the limitations of existing works. First, we adopt a superpoint-level intersection over union (sIOU) metric instead of a point-level IOU. This effectively reduces memory usage and computation time. Also, we apply frame-wise sIOU comparison to match a new observation to existing tracklets (i.e., a list of tracked 2D instances and their lifted 3D superpoints). Specifically, we compare a new observation with each tracked instance in tracklets to find a match. We observe that such frame-wise comparisons induce robustness to wrong 2D predictions and noisy projections compared to tracklet-wise comparisons [37] (i.e., using a representative 3D mask for each tracklet by aggregating 3D superpoints of tracked instances).

Iterative Merging/Removal. We suppress overlapped or partial proposals by merging and removing them. We iteratively merge proposals if they have large overlaps. We refine merged proposals using multi-view consensus [37, 57] after every merge iteration. After the merging step, we remove partial masks if they are included in other proposals.

Instance Classification. We classify the aggregated 3D proposals into open-vocabulary queries. While existing works [39, 49] leverage CLIP [42] for classification, they can be contaminated by co-visible objects or be sensitive to irregularly shaped objects. Instead, we adopt Alpha-CLIP [48] to obtain an object-centric representation by attending object regions using alpha-channel masks. Additionally, we introduce a Standardized Maximum Similarity (SMS) score as a proxy for uncertainty to reduce false positives. The maximum similarity score is standardized using scene-specific statistics, and proposals with low SMS scores are removed from classification. Such a classification strategy helps enhance precision, delivering SoTA performance on benchmarking datasets.

Our contributions are summarized as follows:

- We carefully combine the existing concepts and refine 3D proposal generation by removing overlaps in 2D predictions and applying robust 3D tracking for aggregation.
- We introduce an additional iterative merging/removal step after aggregation to suppress false positives coming from overlapped or partial 3D proposals.
- We take advantage of object-centric feature representation by replacing CLIP with Alpha-CLIP and further reduce false positive 3D proposals by measuring the Standardized Maximum Similarity (SMS) score.
- We demonstrate significant improvements over SoTA methods on ScanNet200 and S3DIS datasets across AP and AR metrics.

2. Related Work

2.1. Closed-vocabulary 3D Instance Segmentation

This task aims to predict 3D instance segmentation masks by assuming a closed set of classes. Several methods [11, 19, 36, 58, 62] have proposed to predict bounding boxes and segment out the instance in each of the bounding boxes. Another group of approaches [4, 10, 16, 25, 30, 34, 52, 54] builds the instances from point embeddings by using graphs or clustering algorithms. Lastly, the most recent line of work [17, 29, 38, 45] adopts transformer architecture [51] or dynamic convolution [24, 50] to predict the 3D instance proposals from the point cloud. Mask3D [45] utilizes transformer architecture along with sparse convolution, demonstrating state-of-the-art performance. Another work, ISBNet [38], proposes to use improved kernel generation and bounding-box-guided dynamic convolutions. In our paper, Mask3D and ISBNet are used as our 3D instance segmentation networks, while other networks are also applicable to ours.

2.2. Open-vocabulary 2D Grounding

One limitation of closed-vocabulary studies is that they hardly generalize to new environments since they cannot identify novel classes that did not appear in the training set. Addressing such concerns, open-vocabulary 2D grounding aims to identify novel classes by adopting 2D foundation models or adapting a model to new scenes to discover novel classes. Three different categories exist under this task: open-vocabulary object detection [5, 26, 35, 41, 53, 61, 64–66], open-vocabulary semantic segmentation [3, 14, 32, 33, 55], and open-vocabulary instance segmentation [43, 56, 67]. Most works [5, 14, 15, 32, 33, 66] propose to align their representations to those of pre-trained vision-language models (VLMs) such as CLIP [42]. In our work, we utilize Grounded SAM [43] as our 2D instance grounding method, which utilizes both Grounding DINO [35] and Segment-Anything Model (SAM) [28]. Grounding DINO detects object bounding boxes with given open-vocabulary queries, and SAM predicts the instance mask in each bounding box.

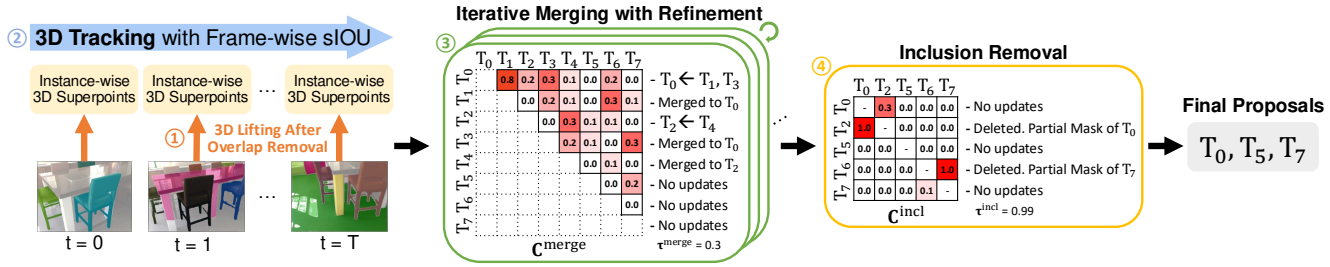

Figure 2. **Overview of image-based 3D proposal generation.** We first remove overlaps between 2D predictions within each frame and lift them to 3D point cloud using a camera projection matrix. Afterward, we aggregate 3D-lifted predictions across frames using a frame-wise sIOU metric with tracking. These 3D proposals are further iteratively merged and refined to progressively merge similar proposals based on a predefined threshold (τ^{merge}). At last, we remove partial masks if their inclusion ratios to other proposals are higher than a predefined threshold (τ^{incl}). For further details about merging and removal, see the last paragraph of Sec. 3.1.

2.3. Open-vocabulary 3D Instance Segmentation

This task aims to address the open-vocabulary grounding problems in 3D point clouds. Several studies [18, 23, 27, 31, 40] have proposed to align the point embeddings to the CLIP embeddings. However, they often require clustering algorithms to find the instances, heavily relying on the accuracy of clustering algorithms. Also, some of them are trained on specific datasets, which may limit their generalization. On the other hand, other studies [2, 22, 39, 49, 57] adopt a two-stage scheme, which first generates class-agnostic masks and then classifies the instances. OpenMask3D [49] utilizes Mask3D [45] to generate the class-agnostic instances in the 3D point cloud and project the instances to 2D images to extract their CLIP embeddings. Also, OpenYOLO3D [2] classifies 3D proposals generated by Mask3D [45] using open-vocabulary 2D object detector [5]. However, while they demonstrate promising results, using a pre-trained 3D instance segmentation model often fails at detecting novel or "tail" classes since they do not or rarely appear during training. Thus, OVIR-3D [37] and SAI3D [63] utilize image-based 3D proposals as an alternative, and Open3DIS [39] uses both image-based and point cloud-based proposals to improve the recall of tail classes.

3. Method

The task of OV-3DIS is to predict a list of 3D instance masks $\mathbf{m} \in \{0,1\}^{K\times N}$ that correspond to a list of user queries $\mathcal{Q}$ from a sequence of images $\mathcal{I}$ and point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$. K denotes the number of 3D proposals, and N denotes the number of points in the point cloud. We generate proposals from both images and point clouds. Our image-based proposal generation is composed of four steps: 2D object grounding, 2D-to-3D lifting, 3D proposal aggregation, and iterative merging/removal. Fig. 2 illustrates the latter three steps in detail with examples. For point cloud-based 3D proposals, we utilize pre-trained 3D instance segmentation models [38, 45] and discard the class predictions, retaining only the class-agnostic masks. We concatenate the proposals from both modalities and classify them into one of the open-vocabulary queries using Alpha-CLIP with SMS-based filtering. We will present our image-based proposal generation in Sec. 3.1, and then elaborate on instance classification in Sec. 3.2.

3.1. Image-based Proposal Generation

Leveraging VFMs [28, 35, 43], image-based proposals provide a complementary approach for detecting novel classes not covered during the training of the 3D instance segmentation models [38, 45]. To generate these proposals, we: 1) ground 2D objects and remove overlapping regions, 2) lift 2D predictions to 3D superpoints, 3) aggregate 3D proposals over frames using tracking, 4) refine 3D proposals, 5) iteratively merge and remove redundant proposals.

2D Object Grounding and Overlap Removal. We use Grounded SAM [43] to segment 2D instance masks in each image from open-vocabulary queries. For each frame, predicted 2D instance masks are sorted by their size, and overlapping regions of larger masks are removed (i.e., overlap removal). This step mitigates the issue of masks frequently capturing multiple objects, which can lead to 3D proposals spanning multiple instances. Although overlap removal may result in partial 3D proposals, we found that separating 3D masks containing multiple instances into distinct masks is far more challenging than starting with multiple partial 3D masks for each instance and merging/removing them after aggregation.

2D Instance to 3D Superpoints Lifting. For 3D point cloud lifting of 2D pixels, we leverage camera matrices, i.e., the multiplication of an intrinsic and extrinsic matrix for each frame. We adopt two concepts from Open3DIS [39]: frame-wise visibility ratio $r_t(\mathbf{s})$ and instance-wise visibility ratio $c_{t,\cdot}(\mathbf{s})$ for filtering out superpoints.

Specifically, $r_t(\mathbf{s})$ denotes the visibility ratio of superpoint $\mathbf{s}$ with respect to image $\mathbf{I}_t$. This ratio is defined as the proportion of 3D points within the superpoint whose projections are visible in the image. Similarly, $c_{t,i}(\mathbf{s})$ indicate the ratio of visible superpoint $\mathbf{s}$ supported (*i.e.,* overlapped) by 2D mask of the i-th instance in image $\mathbf{I}_t$. This is defined as the proportion of visible 3D points within the instance

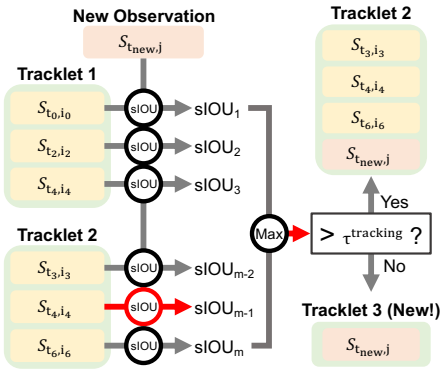

Figure 3. **Matching tracklets with a new observation.** We conduct frame-wise sIOU comparisons between a new observation and each tracked instance in tracklets. If the maximum sIOU exceeds a predefined threshold, we update the corresponding tracklet; otherwise, a new tracklet is initialized.

mask relative to the total number of visible 3D points in the image. Based on these definitions, we define two sets of 3D superpoints: one for 3D superpoints visible in the image and another for 3D superpoints visible within a specific instance mask:

$$\begin{aligned} \mathcal{S}_t &= \{\mathbf{s} \mid r_t(\mathbf{s}) > \tau^{\text{img}}\} \\ \mathcal{S}_{t,i} &= \{\mathbf{s} \mid r_t(\mathbf{s}) > \tau^{\text{img}} \text{ and } c_{t,i}(\mathbf{s}) > \tau^{\text{inst}}\}, \end{aligned} \tag{1}$$

where τ^{img} and τ^{inst} are predefined thresholds for filtering out 3D superpoints with low visibility and/or support from the i-th 2D segment. This formulation ensures that only 3D superpoints with sufficient visibility and/or support are included for further processing.

Tracking-based 3D Proposal Aggregation. We aggregate 2D instance masks and their corresponding 3D superpoints by tracking them over frames. We maintain a list of tracklets, where each tracklet records a list of tracked 2D instance masks and their lifted 3D superpoints. Note that each tracklet corresponds to a single 3D instance proposal after aggregation.

Tracklets are initialized using the lifted 3D superpoints of 2D instances from the first image. Afterward, we associate new observations from the next frames with existing tracklets using frame-wise sIOU metrics. Specifically, we compute the sIOU between the lifted 3D superpoints of the new observation and each tracked instance in each tracklet (see Fig. 3). If the highest sIOU exceeds a predefined threshold τ^{tracking}, the new observation is assigned to the corresponding tracklet for update. Otherwise, a new tracklet is created for this new instance. When measuring sIOU between two sets, we only consider co-visible superpoints in both image frames. Formally noting, given the i-th instance mask from an image $\mathbf{I}_{t_a}$ and the j-th instance mask from $\mathbf{I}_{t_b}$, we denote their corresponding 3D superpoints within each instance mask as $\mathcal{S}_{t_a,i}$ and $\mathcal{S}_{t_b,j}$ respectively and a co-visible 3D superpoints set between images

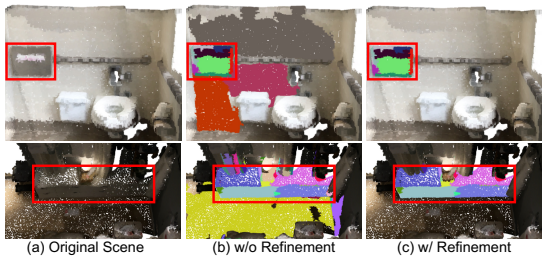

Figure 4. **Effectiveness of 3D proposal refinement.** Red boxes indicate the object of interest, and segments of different colors denote 3D superpoints. Without refinement, the 3D instance proposal often extends beyond the object boundaries due to noisy 2D-to-3D projections or inaccurate mask predictions. With refinement, irrelevant 3D superpoints are removed, and our method successfully removes 3D superpoints that do not belong to the object, resulting in geometrically consistent and precise predictions.

as $\text{Vis}_{t_a,t_b} = \mathcal{S}_{t_a} \cap \mathcal{S}_{t_b}$. sIOU between these two instance masks from two frames is defined as:

$$\text{sIOU} = \frac{|(\mathcal{S}_{t_a,i} \cap \mathcal{S}_{t_b,j})|}{|(\mathcal{S}_{t_a,i} \cup \mathcal{S}_{t_b,j}) \cap \text{Vis}_{t_a,t_b}|}, \tag{2}$$

where $|\cdot|$ denotes the cardinality.

3D Proposal Refinement. After tracking, we have an additional refinement step for the 3D proposal in each tracklet by removing 3D superpoints that are infrequently visible in the tracked 2D instances across multiple views. We adopt a concept from MaskClustering [57] and OVIR-3D [37] and calculate a *superpoint-level* multi-view consensus rate. As illustrated in Fig. 4, removing superpoints with low visibility effectively refines the proposal to have a tight, semantic-aligned boundary. For each superpoint in a tracklet, the multi-view consensus rate is defined as the ratio of tracked frames in which the superpoint appears within the instance mask to the total number of frames where it is visible. 3D superpoints with a consensus rate below a predefined threshold τ^{ref} are removed from the tracklet, ensuring only reliable superpoints are retained.

Iterative 3D Proposal Merge and Removal. While our overlap removal step in the 2D grounding step effectively removes masks spanning multiple instances, it may generate partial masks of instances. For example, as shown in Fig. 5, a single object may be decomposed into multiple partial 3D proposals, each covering only part of the object. To address this, we merge these partial proposals into a complete 3D representation. Moreover, we apply this merging iteratively so that we can progressively enlarge instances at each iteration. Also, each merge is followed by proposal refinement using multi-view consensus to exclude noisy superpoints from later merging. Conversely, when a larger 3D proposal contains smaller, redundant proposals, we remove the redundant ones to ensure higher precision.

Suppose we have K tracklets, each with a 3D mask proposal represented as $\mathbf{m}_k \in \{0,1\}^N$, derived from the

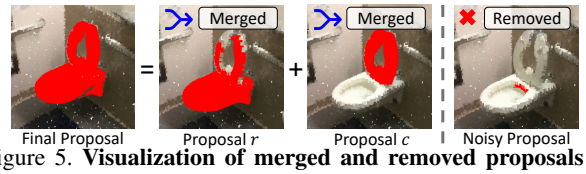

Figure 5. **Visualization of merged and removed proposals in the ScanNet200 dataset.** Overlapping and noisy proposals often emerge after instance tracking. We effectively handle these issues by merging duplicate proposals and eliminating noisy ones, ensuring high-quality proposals.

tracked 3D superpoints within each tracklet. For each merging iteration, we compute IOU between a pair of 3D proposals, constructing a cost matrix $\mathbf{C}^{\text{merge}} \in [0,1]^{K\times K}$ that is a strictly upper-triangular matrix. For each 3D proposal, we identify other proposals with an IOU exceeding a predefined threshold τ^{merge} and merge them into the current 3D proposal. This process is repeated until no further merges are possible. After each 3D proposal merge, we also merge their corresponding tracklets and refine the resulting 3D proposal using the multi-view consensus rate. More detailed implementation can refer to supplementary materials.

After merging proposals, we remove smaller 3D proposals that are contained within larger ones. Given two 3D proposals $\mathbf{m}_r, \mathbf{m}_c \in \{0,1\}^N$, we define an inclusion rate of $\mathbf{m}_r$ within $\mathbf{m}_c$ as $r^{\text{incl}}(\mathbf{m}_r, \mathbf{m}_c)$, which is the proportion of $\mathbf{m}_r$ included in $\mathbf{m}_c$, with a value of 1 indicating that $\mathbf{m}_r$ is fully contained within $\mathbf{m}_c$. Using this ratio, we construct an inclusion cost matrix $\mathbf{C}^{\text{incl}} \in [0,1]^{K\times K}$, which is a full matrix since the inclusion ratio is asymmetric. For each 3D proposal, if its inclusion rate with respect to any other proposal exceeds a predefined threshold τ^{incl}, the proposal is removed. Unlike the merge process, this filtering step is applied only once.

3.2. Open-Vocabulary Instance Classification

For instance classification, we concatenate class-agnostic proposals from image-based and point cloud-based methods and apply non-maximum suppression (NMS) with an IOU threshold of 0.95. We prioritize the point cloud-based proposals over the image-based ones since they have fewer false positives.

Feature Extraction. Previous work [39, 49] leverages CLIP [42] to extract visual features from cropped image regions using projected instance bounding boxes. However, this approach has notable limitations (Fig. 6): (a) Resizing the crop to a square aspect ratio distorts the object's original geometry, hindering CLIP's ability to capture its geometric characteristics. (b) Visual features are contaminated by co-visible objects (e.g., bookshelves, tables), leading to poor predictions from CLIP. To address these issues, we adopt Alpha-CLIP [48] to enforce an object-centric focus to the model. Alpha-CLIP incorporates object masks as an additional input to guide the model's attention. The object masks are generated using SAM by projecting the predicted 3D

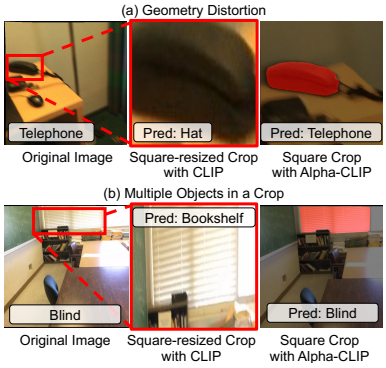

Figure 6. **Failure cases of using CLIP for instance classification.** CLIP fails when the shape of the object gets distorted or when other objects are also present within the crop.

proposals onto images and querying with bounding boxes or subsampled points. We also apply a square crop during preprocessing to preserve the object's geometry while ensuring compatibility with the model's input requirements.

We adopt a similar approach to OpenMask3D [49] for visual embedding extraction. Given a 3D proposal and the visual encoder from Alpha-CLIP, we project the proposal onto all 2D images and select a subset of images with the highest visibility for multiscale visual feature extraction. Let $\mathbf{f}^l_{v,k}$ represent the CLIP feature extracted at a scale level l from the v-th image for the k-th 3D proposal. The final L2-normalized feature $\mathbf{F}_k$ for this proposal is computed as:

$$\mathbf{F}_k = \sum_{v\in\mathcal{V}_k} \sum_{l\in\mathcal{L}} \mathbf{f}^l_{v,k} \cdot \alpha_{v,k}, \tag{3}$$

where $\mathcal{V}_k$ denotes the set of images with top visibility for the k-th 3D proposal, and $\alpha_{v,k} \in [0,1]$ is the visibility ratio of the k-th 3D proposal in image $\mathbf{I}_v$. This ratio is defined as the number of visible points in the image divided by the total number of points in the 3D proposal. Finally, given K proposals and C text queries, we compute the cosine similarity between the visual features of the proposals and the text features, resulting in a similarity matrix $\mathbf{L} \in [-1,1]^{K\times C}$.

3D Proposal Filtering with SMS. We further suppress unconfident proposals by using the CLIP similarity score as a proxy for uncertainty. However, CLIP scores are not normalized across different text embeddings, making it challenging to apply a single filtering threshold for all queries. To address this, we standardize the maximum similarity scores (i.e., SMS score) within each text embedding to obtain relative scores. Specifically, for each query q_c, we compute the mean and variance of the similarity scores as $\mu_c = \frac{1}{K}\sum_k \mathbf{L}_{k,c}$ and $\sigma_c^2 = \frac{1}{K}\sum_k (\mathbf{L}_{k,c} - \mu_c)^2$, where K is the total number of proposals. Next, for each proposal k, we identify the maximum similarity value $\mathbf{L}_{k,c_{\max}}$ across all queries and standardize it using the corresponding statistics: $c_k^{\text{SMS}} = \frac{\mathbf{L}_{k,c_{\max}} - \mu_{c_{\max}}}{\sigma_{c_{\max}}}$. Proposals with an SMS score below a predefined threshold τ^{SMS} are removed from the predictions, ensuring more reliable filtering.

Eval. Protocol	Methods	3D Proposals		mAP	mAP_{50}	mAP_{25}	mAP_{head}	mAP_{common}	mAP_{tail}
		Image-based	Point cloud-based						
Fully Supervised	ISBNet [38]	✗	✓	24.5	32.7	37.6	38.6	20.5	12.5
	Mask3D [45]	✗	✓	26.9	36.2	41.4	39.8	21.7	17.9
Top-1	SAM3D† [60]	✓	✗	9.8	15.2	20.7	9.2	8.3	12.3
	OVIR-3D† [37]	✓	✗	9.3	18.7	25.0	9.8	9.4	8.5
	SAI3D† [63]	✓	✗	12.7	18.8	24.1	12.1	10.4	16.2
	Ours (2D Only)	✓	✗	**21.5**	**31.2**	**37.7**	**18.8**	**19.6**	**26.9**
	OpenIns3D [22]	✗	✓	8.8	10.3	14.4	16.0	6.5	4.2
	OpenMask3D [49]	✗	✓	15.4	19.9	23.1	17.1	14.1	14.9
	OpenYOLO3D [2]	✗	✓	21.9	28.3	31.7	25.6	21.1	18.5
	Ours (3D Only)	✗	✓	**24.2**	**31.8**	**36.4**	**27.2**	**22.3**	**23.1**
	OpenScene [40]	✓	✓	11.7	15.2	17.8	13.4	11.6	9.9
	Ours (2D + 3D)	✓	✓	**25.8**	**32.5**	**36.2**	**26.3**	**23.2**	**28.2**
Top-K	Open3DIS [39]	✓	✗	18.2	26.1	31.4	18.9	16.5	19.2
	Ours (2D Only)	✓	✗	**25.4**	**37.4**	**44.4**	**23.4**	**23.5**	**30.2**
	Open3DIS [39]	✗	✓	18.6	23.1	27.3	24.7	16.9	13.3
	OpenYOLO3D [2]	✗	✓	24.7	31.7	36.2	27.8	24.3	21.6
	Ours (3D Only)	✗	✓	**29.0**	**37.6**	**42.8**	**33.0**	**28.1**	**25.3**
	Open3DIS [39]	✓	✓	23.7	29.4	32.8	27.8	21.2	21.8
	Ours (2D + 3D)	✓	✓	**32.7**	**41.4**	**45.3**	**34.5**	**30.7**	**33.1**

Table 1. **OV-3DIS results on the ScanNet200 validation set [7].** Top-1 evaluation protocol refers to assigning one predicted class per instance mask, and Top-K evaluation protocol [2, 39] refers to allowing multiple predicted classes per instance mask. We evaluate methods under three settings: image-based 3D proposals only (*i.e.,* 2D only), point cloud-based 3D proposals only (*i.e.,* 3D only), and a combination of both (*i.e.,* 2D+3D). In all three settings across different protocols, our method achieves the SoTA performance, significantly outperforming other methods. †numbers are adopted from SAI3D [63].

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on three datasets: ScanNet200 [7], S3DIS [1], and Replica [47]. **ScanNet200** is a real-world dataset comprising diverse indoor environments with 200 object categories. It includes 1,201 scenes in the training set and 312 scenes in the validation set. Object categories are divided into head, common, and tail classes based on their frequency. We validate our method and baselines on the validation set, reporting performance for each category group (*i.e.,* head, common, tail) as well as the overall performance. **S3DIS** consists of 271 scenes from 6 different areas, with Area 5 used for our evaluation. Although the dataset includes 13 classes, we exclude "stuff" categories such as floor, ceiling, wall, and clutter from our evaluation to focus on object-centric performance. **Replica** is a synthetic dataset created from digital replicas of real-world scenes, featuring 48 object classes across 8 different scenes. On this dataset, we assess generalization performance by evaluating a 3D instance segmentation model [45] trained on ScanNet200.

Evaluation Metrics. We measure mean average precision (mAP) and mean average recall (mAR) at IOU thresholds of 25% and 50%. Additionally, we measure mAP and mAR across IOU thresholds ranging from 50% to 95% with 5% increments. For class-agnostic evaluations, we calculate AP and AR on those IOU ranges.

Evaluation Protocols. We found existing literature adopts different evaluation strategies. Several works [22, 40, 49, 63] assign one class prediction per each 3D instance (i.e., Top-1), while other works [2, 39] allows multiple predictions per each 3D instance by selecting Top-K predictions (e.g., top 300 / 600) over class predictions of all instances, based on their prediction scores. For fair comparisons, we evaluate our method on both evaluation settings on ScanNet200 and Replica. For the S3DIS dataset, we only evaluate by using the Top-1 strategy. Further details can be found in the supplementary materials.

Implementation Details. For the ScanNet200 dataset, we downsample the number of image frames by a factor of 5 to reduce computational load. We follow the same setting of OpenMask3D [49] for multi-scale CLIP feature extraction, i.e., 3 scale levels with an expansion ratio of 0.2. We set the following thresholds for both ScanNet200 and S3DIS: $\tau^{\text{img}} = 0.1$, $\tau^{\text{inst}} = 0.3$, $\tau^{\text{tracking}} = 0.3$, $\tau^{\text{merge}} = 0.3$, $\tau^{\text{ref}} = 0.4$, and $\tau^{\text{incl}} = 0.99$. For the Replica dataset, we adjust τ^{merge} to 0.7 and disable multiview consensus ratio-based filtering, as Replica is a synthetic dataset without projection errors. Additionally, we observe a distribution shift in the CLIP visual representations, as CLIP is trained on real-world data while Replica consists of synthetic data. This shift enlarges the gap between the visual and text embeddings of CLIP. To address this, inspired by prior work on handling distributional gaps [6, 20, 21, 44, 46], we perform dimension reduction by computing the first principal axis of the visual CLIP features and removing its contribution from both visual and text embeddings. We use the template "*a blurry photo of* {*CLASS_NAME*} *in a room.*" Further details are provided in the supplementary materials.

4.2. Quantitative Results

ScanNet200. We adopt Mask3D [45] trained on the ScanNet200 training set for point cloud-based proposals. Table 1 shows that our method outperforms all baselines in all three

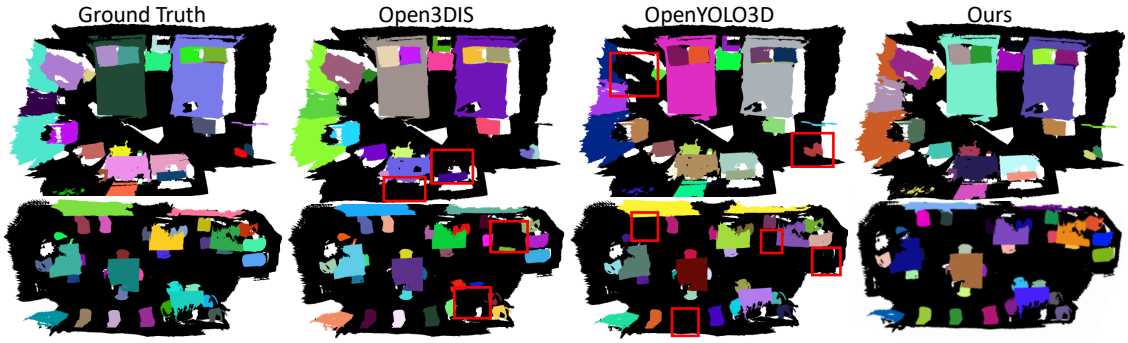

Figure 7. **Qualitative comparisons on the ScanNet200 dataset.** Black regions indicate empty predictions (*no object*), while red boxes highlight objects missed by other methods but successfully detected by ours. 3D instance masks are colored randomly.

Methods	mAP	mAP_{50}	mAP_{25}	mAR	mAR_{50}	mAR_{25}
Open3DIS† [39]	24.5	36.2	49.3	30.5	43.4	55.3
Ours (2D Only)	**26.9**	**41.4**	**51.9**	**35.3**	**52.5**	**62.5**
Open3DIS† [39]	26.6	33.5	39.2	34.2	41.7	47.4
OpenYOLO3D† [2]	28.1	37.3	44.6	37.8	46.3	52.0
Ours (3D Only)	**30.4**	**39.4**	**47.4**	**38.7**	**47.9**	**55.1**
Open3DIS† [39]	28.9	37.0	43.1	44.1	54.5	61.4
Ours (2D + 3D)	**31.3**	**43.5**	**50.4**	**48.2**	**65.1**	**72.9**

Table 2. **OV-3DIS results on S3DIS [1].** †numbers are obtained using their official codes. Top-1 evaluation protocol is used.

settings, *i.e.,* 2D-only, 3D-only, and 2D+3D across all evaluation protocols. In 2D-only evaluations, our method outperforms previous SoTA methods, SAI3D and Open3DIS, by 8.8% and 7.2% , respectively. Notably, our image-based method excels at predicting tail classes, where other comparing methods often struggle. Note that tail classes refer to *less frequent, rare* classes in the training set, not based on their object sizes. For 3D-only evaluations, we use the same point cloud-based 3D proposals for both our method and OpenYOLO3D. The resulting 2.3% and 4.3% mAP improvements over OpenYOLO3D demonstrate the effectiveness of our classification method. Furthermore, using both image-based and point cloud-based proposals, we outperform the previous SoTA, Open3DIS [39], by 9.0% in mAP.

S3DIS. We train ISBNet [38] on Area 1∼4, 6 and adopt its predictions on Area 5 as our point cloud-based proposals for all the baselines. As reported in Table 2, our method consistently outperforms the baselines by a large margin in each experiment setting: 2D-only, 3D-only, and 2D+3D. Using both image-based and point cloud-based proposals boosts the recall significantly, improving 2D-only and 3D-only methods by 12.9% and 9.5%, respectively. Note that we use thing classes only since our task is 3D instance segmentation. The results for both stuff and thing classes are available in the supplementary materials.

Replica. The results are summarized in Table 3. This experiment aims to assess the generalizability of our method by adopting a ScanNet200-trained 3D instance segmentation model for point cloud-based proposals. Our approach consistently outperforms other methods within the same category under both 2D-only and 2D+3D settings, achieving

Eval.	Methods	mAP	mAP_{50}	mAP_{25}	mAR	mAR_{50}	mAR_{25}
Top-1	OVIR-3D [37]	11.1	20.5	27.5	-	-	-
	Ours (2D Only)	**20.8**	**32.4**	**38.5**	28.5	43.1	49.9
	OpenMask3D [49]	13.1	18.4	24.2	-	-	-
	OpenYOLO3D [2]	**23.7**	**28.6**	**34.8**	**26.6**	**31.9**	**38.5**
	Ours (3D Only)	22.0	26.7	32.5	**26.6**	31.5	37.0
	Ours (2D + 3D)	**22.6**	**31.7**	**37.7**	**33.9**	**46.5**	**53.6**
Top-K	Open3DIS† [39]	18.2	25.9	31.0	32.3	46.2	54.9
	Ours (2D Only)	**21.6**	**32.6**	**39.8**	**39.6**	**59.5**	**71.3**
	Open3DIS† [39]	16.0	19.4	23.5	29.2	35.4	42.5
	Ours (3D Only)	**18.9**	**24.4**	**32.1**	**34.0**	**43.9**	**57.5**
	Open3DIS† [39]	18.4	23.8	28.2	33.0	42.6	50.0
	Ours (2D + 3D)	**25.7**	**34.9**	**42.3**	**48.8**	**66.3**	**79.7**

Table 3. **OV-3DIS results on Replica [47].** †numbers are obtained using their official codes.

superior results on both mAP and mAR. In the 3D-only setting, our method significantly surpasses OpenMask3D and Open3DIS, which adopt CLIP [42] for predictions. However, it lags behind OpenYOLO3D [2] in terms of mAP, which does not use CLIP for instance classification. We hypothesize that the domain gap between real-world data and synthetic data from Replica may degrade the performance of Alpha-CLIP. Surprisingly, our 2D-only method achieves higher mAP_{50} and mAP_{25} than 3D only methods where the masks are generated from ScanNet200-trained 3D networks. This highlights the exceptional generalization capability of our 2D-only approach. Additionally, our 2D+3D method attains the highest mAR across all settings.

4.3. Qualitative Results

Fig. 7 presents qualitative comparisons on the ScanNet200 dataset. Red boxes indicate instances missed by Open3DIS and OpenYOLO3D, while our method successfully detects all objects. These visual results are consistent with the recall metrics: Open3DIS and OpenYOLO3D achieve the mAR of 43.3% and 47.7%, respectively, whereas our method significantly outperforms both with an mAR of 61.4%. We also present OV-3DIS results using novel text queries in Fig. 1. Our framework effectively retrieves 3D instances based on functional descriptions (*e.g.*, drink water) and object attributes (*e.g.*, red chair).

4.4. Ablation Study

Class-agnostic Evaluation. To evaluate the quality of generated proposals, we report class-agnostic AP and AR on

Methods	AP	AP_{50}	AP_{25}	AR	AR_{50}	AR_{25}
ISBNet (fully-sup.)† [38]	40.2	50.0	54.6	66.8	80.4	87.4
Mask3D (fully-sup.)† [38]	50.6	68.0	76.9	65.3	81.0	88.4
Superpoints† [13]	5.0	12.7	38.9	-	-	-
DBSCAN† [12]	1.6	5.5	32.1	-	-	-
OVIR-3D [37]	14.4	27.5	38.8	-	-	-
Mask Clustering [57]	19.2	36.6	51.7	-	-	-
Open3DIS (2D Only) [39]	29.7	45.2	56.8	49.0	70.0	83.2
Ours (2D Only)	**33.3**	**51.9**	**66.1**	**50.2**	**72.2**	**85.5**
Open3DIS (2D + 3D) [39]	34.6	43.1	48.5	66.2	81.6	91.4
Ours (2D + 3D)	**46.6**	**59.0**	**64.4**	**74.0**	**89.8**	**95.9**

Table 4. **Class-agnostic evaluation on the ScanNet200 [7].**

the ScanNet200 dataset. As shown in Table 4, our 2D-only method outperforms all image-based approaches, surpassing the previous SoTA by 3.6%. Both our method and OVIR-3D [37] share the intuition of sequentially tracking 3D proposals to progressively grow regions. However, the performance gap between our method and OVIR-3D is substantial, highlighting the effectiveness of our frame-wise tracklet matching algorithm and iterative merging/removal with refinements. Furthermore, by leveraging both image-based and point cloud-based 3D proposals, our 2D+3D method achieves the highest ARs across all methods, including fully-supervised approaches.

Method	AP	AP_{50}	AP_{25}
Tracklet-wise sIOU for Tracking	34.7	54.3	69.6
Frame-wise sIOU for Tracking	**35.1** (+0.4)	**56.1** (+1.8)	**70.5** (+0.9)

Table 5. **Impact of different tracklet matching strategies for aggregation on the subset of the ScanNet200 validation set.** Class-agnostic APs are reported.

Different Tracklet Matching Strategies. While we adopt frame-wise sIOU for tracklet matching, some approaches [37] leverage the aggregated 3D mask of each tracklet for matching (i.e., tracklet-wise sIOU for aggregation in Table 5). Specifically, we maintain aggregated 3D superpoint masks of tracked 2D instances for each tracklet and measure sIOU with new observation by only using co-visible superpoints. As reported in Table 5, frame-wise sIOU for tracking brings meaningful performance gain for AP_{50} and AP_{25} over the tracklet-wise sIOU while both maintain reasonably good APs. We conjecture that this is because wrong predictions are always accounted for obtaining aggregated 3D masks in the case of tracklet-wise matching. However, wrong predictions may not have any impact at all during the frame-wise matching if wrong predictions are distinctive from new observations and other predictions have higher sIOU then them, preventing the wrong predictions from being used for matching.

Method	AP	AP_{50}	AP_{25}
Agg. Only	31.4	49.9	63.5
+ Iter. Merging/Removal	31.8 (+0.4)	51.8 (+1.9)	68.5 (+5.0)
+ Overlap Removal	33.5 (+2.1)	54.4 (+4.5)	69.8 (+6.3)
+ Iter. Refine	**35.1** (+3.7)	**56.1** (+6.2)	**70.5** (+7.0)

Table 6. **Impact of iterative merging/removal, overlap removal, and iterative refinement on the subset of ScanNet200 validation set.** Class-agnostic APs are reported.

Impact of Iterative Merging/Removal Unlike existing methods [37, 39, 63], our method has an additional false positive suppression step by merging/removing duplicated proposals. As shown in Table 6, applying iterative merging and removing improves AP_{25} by 5.0%. More importantly, if we apply iterative merging with overlap removal in the 2D grounding step, it further brings significant gains in all AP metrics. This is because overlap removal effectively separates masks spanning multiple instances into each instance or partial masks, which later can be merged/removed. At last, applying iterative refinement further improves the quality, especially in AP and AP_{50} metrics.

Method	mAP	mAP_{50}	mAP_{25}
Ours w/ CLIP	27.5	34.7	38.2
+ Alpha-CLIP	30.5 (+3.0)	37.6 (+2.9)	41.1 (+2.9)
+ SMS-based Filtering	**32.7** (+5.2)	**41.4** (+6.7)	**45.3** (+7.1)

Table 7. **Impact of Alpha-CLIP and SMS-based filtering in instance classification on the ScanNet200 dataset.**

Impact of Alpha-CLIP and SMS Filtering. Table 7 demonstrates the effectiveness of using Alpha-CLIP and SMS-based filtering. As shown, using Alpha-CLIP improves the performance from 27.5 to 30.5 mAP, proving the importance of considering object-centric representation in instance classification. However, we note that our method with CLIP still surpasses existing baselines by a large margin (i.e., 3.8% over Open3DIS and 2.8% over OpenYOLO3D). Using SMS-based filtering also brings gains in AP metrics by effectively removing unconfident instances from both image-based and point cloud-based proposals. Full ablation study on all three datasets can be found in the supplementary materials.

5. Conclusion

In this paper, we carefully combine existing concepts and devise each stage to achieve precise 3D proposal generation and accurate instance classification. Our robust 3D tracking allows for more precise 3D proposal aggregation. Also, overlap removal in 2D predictions accompanied with iterative merging/removal enables much fewer false positive 3D proposals, such as overlapped or partial masks. At last, we adopt Alpha-CLIP to obtain object-centric CLIP representation and remove unconfident 3D proposals by filtering with a standardized maximum similarity score. Although our method achieves SoTA precision and recall across datasets, our method is computationally intense because heavy 2D foundation models [28, 43, 48] are adopted in our pipeline. Also, we found that our method fails to improve performance on small objects (e.g., ScanNet++ in the supplementary) but rather remain similar to existing approaches. This is because iterative merging and removal is more effective for medium or large objects. Improving such limitations remains our future work.

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. 6, 7

[2] Mohamed El Amine Boudjoghra, Angela Dai, Jean Lahoud, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Open-yolo 3d: Towards fast and accurate open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2406.02548*, 2024. 1, 3, 6, 7

[3] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[4] 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

[5] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2, 3

[6] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11580–11590, 2021. 6

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 6, 8

[8] 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*, pages 7010–7019, 2023. 1

[9] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[10] 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

[11] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020. 2

[12] 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. 8

[13] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 2, 8

[14] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[16] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 2

[17] 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

[18] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2028–2038, 2023. 3

[19] 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

[20] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2994–3003, 2024. 6

[21] Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 791–800, 2018. 6

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

[23] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 3

[24] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 2

[25] 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

[26] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, pages 15946–15969. PMLR, 2023. 2

[27] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 3

[28] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 8

[29] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20943–20953, 2024. 2

[30] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2019. 2

[31] Seungjun Lee, Yuyang Zhao, and Gim Hee Lee. Segment any 3d object with language. *arXiv preprint arXiv:2404.02157*, 2024. 3

[32] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[33] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1, 2

[34] 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

[35] 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, 3

[36] Shih-Hung Liu, Shang-Yi Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning gaussian instance segmentation in point clouds. *arXiv preprint arXiv:2007.09860*, 2020. 2

[37] 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 *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 1, 2, 3, 4, 6, 7, 8

[38] 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, 2, 3, 6, 7, 8

[39] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4018–4028, 2024. 1, 2, 3, 5, 6, 7, 8

[40] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 3, 6

[41] Chau Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 779–788, 2024. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 7

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1, 2, 3, 8

[44] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9471–9480, 2019. 6

[45] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 1, 2, 3, 6

[46] Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening and coloring batch transform for gans. *arXiv preprint arXiv:1806.00420*, 2018. 6

[47] 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. 6, 7

[48] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 2, 5, 8

[49] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 1, 2, 3, 5, 6, 7

[50] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020. 2

[51] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[52] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 2

[53] 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

[54] 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

[55] 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

[56] 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. 2

[57] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28274–28284, 2024. 1, 2, 3, 4, 8

[58] 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. *Advances in neural information processing systems*, 32, 2019. 2

[59] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19823–19832, 2024. 1

[60] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2, 6

[61] 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

[62] 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

[63] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3292–3302, 2024. 1, 2, 3, 6, 8

[64] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 2

[65] 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.

[66] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 1, 2

[67] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 2