Zero-Shot Dual-Path Integration Framework for Open-Vocabulary 3D Instance Segmentation

Tri Ton[1*], Ji Woo Hong[1*], SooHwan Eom[1], Jun Yeop Shim[1], Junyeong Kim[2], Chang D. Yoo[1]
[1]Korea Advanced Institute of Science and Technology (KAIST) [2]Chung-Ang University
{tritth, jiwoohong93, sean1105, shimjay17, cd_yoo}@kaist.ac.kr, junyeongkim@cau.ac.kr

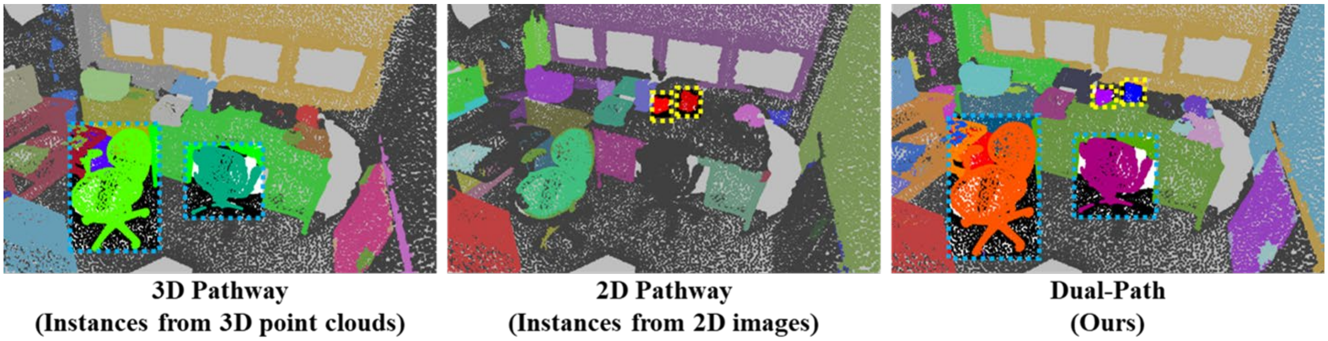

Figure 1. Instance segmentation results from different modality of 2D and 3D. Our Zero-Shot Dual-Path Integration Framework complementarily integrates outputs from two modalities.

Abstract

Open-vocabulary 3D instance segmentation transcends traditional closed-vocabulary methods by enabling the identification of both previously seen and unseen objects in real-world scenarios. It leverages a dual-modality approach, utilizing both 3D point clouds and 2D multi-view images to generate class-agnostic object mask proposals. Previous efforts predominantly focused on enhancing 3D mask proposal models; consequently, the information that could come from 2D association to 3D was not fully exploited. This bias towards 3D data, while effective for familiar indoor objects, limits the system's adaptability to new and varied object types, where 2D models offer greater utility. Addressing this gap, we introduce Zero-Shot Dual-Path Integration Framework that equally values the contributions of both 3D and 2D modalities. Our framework comprises three components: 3D pathway, 2D pathway, and Dual-Path Integration. 3D pathway generates spatially accurate class-agnostic mask proposals of common indoor objects from 3D point cloud data using a pre-trained 3D model, while 2D pathway utilizes pre-trained open-vocabulary instance segmentation model to identify a diverse array of object proposals from multi-view RGB-D images. In Dual-Path Integration, our Conditional Integration process, which operates in two stages, filters and merges the proposals from both pathways adaptively. This process harmonizes output proposals to enhance segmentation capabilities. Our framework, utilizing pre-trained models in a zero-shot manner, is model-agnostic and demonstrates superior performance on both seen and unseen data, as evidenced by comprehensive evaluations on the ScanNet200 and qualitative results on ARKitScenes datasets.

1. Introduction

The advent of 3D instance segmentation task, which predicts 3D object instances and their corresponding categories from 3D point cloud data, has marked a significant milestone for its applications in autonomous driving, robotics, augmented reality, and more. Traditional 3D instance segmentation methods [2, 6, 13, 15, 21, 25, 30, 34, 35, 38, 40, 42, 44, 44, 46] have largely operated within a

[*]equal contribution.

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and (No.2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

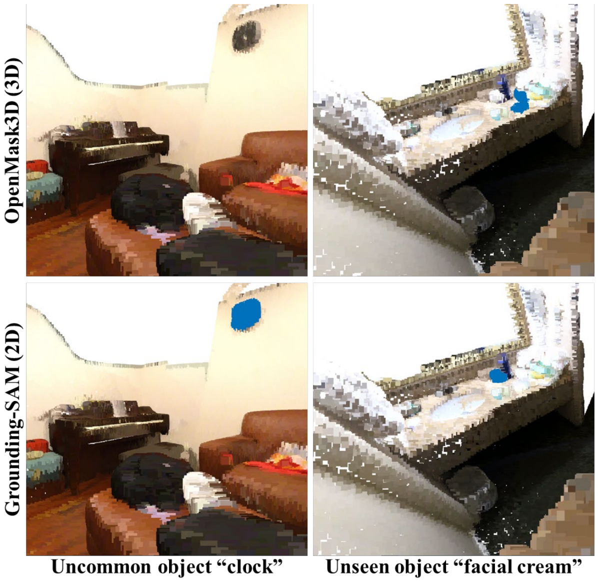

Figure 2. Capability of pre-trained open-vocabulary 2D instance segmentation in detecting uncommon and unseen object classes that remain undetected by pre-trained 3D instance segmentation models.

closed-vocabulary paradigm, where the classes of objects to be segmented are predetermined and known during the training phase. However, real-world scenarios frequently present objects that fall outside of these predefined classes, making closed-vocabulary segmentation inadequate. Open-vocabulary 3D instance segmentation emerged as a necessary evolution to address these real-world complexities. The challenge lies in the system’s ability to generalize beyond the training dataset and to adapt to the unpredictability inherent in real-world environments.

Current methodologies [17, 36] independently process 3D and 2D data of the same scene for discrete sub-tasks without harnessing the inherent synergistic potential between two modalities and heavily depend on the initial 3D instance masks for subsequent classification. From these works, it has been observed that while a class-agnostic 3D instance segmentation model using point cloud data excels at segmenting common objects such as ‘TV’, it often struggles with uncommon classes like ‘paper towel roll’ and unseen classes like ‘facial cream’. Conversely, open-vocabulary 2D instance segmentation excels at detecting unfamiliar objects within 3D data, benefiting from the robust generalization afforded by the vision-language understanding capabilities of pre-trained image classifiers, as illustrated in Fig. 2. Thus, we argue that leveraging both modalities of the same scene offers distinct advantages.

Bridging the gap between two modalities, we propose a Zero-shot Dual-Path Integration Framework designed to synergistically merge the instance proposals derived from the 3D point cloud and the 2D multi-view images, which excel in spatial precision on common objects and diverse recognition capabilities respectively. Our framework comprises three components: 3D pathway, 2D pathway, and Dual-Path Integration.

In the 3D pathway, 3D instance masks are generated from a pre-trained class-agnostic mask proposal generator using a 3D point cloud of spatial information to generate as many proposals as possible. In the 2D pathway, 2D instance masks are generated from an pre-trained open vocabulary 2D instance segmentation using RGB-D multi-view image sequence of 2D visual details, which are then projected into the 3D scan and refined through Instance Fusion Module.

Employing a straightforward integration approach of using all proposals from both the 3D and 2D pathways, as we call Simple Integration, will enhance the performance of instance segmentation to a certain extent as the diversity of proposals enhances the recall rate. Yet, the quality of the proposals is also critical, as integrating numerous but low-quality proposals can lead to increased false positive, which may adversely affect the overall precision. Consequently, while the diversity of proposals can boost the recall, it is essential to maintain a balance between quantity and quality. To this end, in the final Dual-Path Integration phase, our meticulously designed Conditional Integration evaluates proposals from both the 3D and 2D modalities with impartial consideration. This module unfolds in two pivotal stages: (1) Dual-modality Proposal Matching and (2) Adaptive Integration. During the Dual-modality Proposal Matching stage, we compute Intersection-of-Union (IoU) across the proposals from 3D and 2D pathways. This aims to identify pairs of proposals with overlapping regions, suggesting they may represent parts of identical objects. The unique proposals with no overlap with any other proposals of other modality are filtered to be added to the final proposal outputs. Subsequently, Adaptive Integration establishes a balance between segmentation precision and the identification of diverse instances based on the IoU assessment. The samples of the 3D pathway, 2D pathway, and Dual-path instances can be seen in Fig. 1.

Our contributions are as follows:

- **Zero-shot 3D instance segmentation via comprehensive exploitation of pre-trained models of both 3D and 2D modalities**: We introduce zero-shot framework that judiciously leverages the strengths of pre-trained models of both 3D point cloud and 2D image data for 3D instance segmentation. This approach embraces a model-agnostic strategy that avoids the traditional dependency on a pre-trained model of single modality.
- **Dual-Path Integration of Conditional Integration**: We propose a Dual-Path Integration with Conditional Integration process that effectively combines instance mask proposals from both 3D and 2D pathways to reconcile and enhance mask proposals.

- **Enhanced Overall Performance**: Our framework's efficacy is validated through evaluations on the ScanNet200 and qualitative results in ARKitScenes. The results demonstrate an uplift in the overall performance for open-vocabulary 3D instance segmentation, underscoring the potency of our proposed approach.

2. Related Work

Closed Vocabulary 3D Instance Segmentation. 3D instance segmentation techniques are categorized into proposal-based, clustering-based, and transformer-based approaches. Proposal-based methods [15, 42, 44] identify 3D bounding boxes to distinguish instances, yet face challenges with varying point cloud distributions. Clustering-based approaches [2, 6, 13, 21, 25, 30, 38, 40, 44, 46] predict semantic categories and use geometric offsets for instance grouping, though they require extensive manual tuning and may struggle with novel test objects. Transformer-based methods [34, 35], leveraging the Mask2Former framework [3], achieve state-of-the-art results by representing instances as queries within a transformer decoder, effectively utilizing global features for mask prediction, demonstrating robust performance, particularly on the ScanNet benchmark [4].

Open Vocabulary 2D Instance Segmentation. Open-vocabulary 2D segmentation, empowered by large-scale vision-language models such as CLIP [32], LiT [45], and ALIGN [20], has emerged as a significant advancement in instance segmentation. This development has fostered novel open-vocabulary and zero-shot segmentation methodologies. Pixel-level embedding techniques [8, 18, 24, 37, 39, 41] have shown promising results, although their success is contingent on mask precision and requires specific training. To overcome these challenges, integrating robust zero-shot detection and segmentation models such as ViLD [11], OWL-ViT [29], Detic [47], and Grounding-DINO [26] with the Segment Anything Model (SAM) [23] facilitates open-vocabulary 2D instance segmentation. This approach, especially when combining Grounding-DINO with SAM (Grounded-SAM), leverages CLIP features for classification, enabling precise segmentation and classification without the necessity for additional fine-tuning or training. In our work, we employ Grounded-SAM for open-vocabulary 2D instance segmentation.

Open Vocabulary 3D Scene Understanding. Recent advancements in open-vocabulary 3D scene understanding [5, 10, 12, 16, 17, 19, 22, 27, 31, 36] have focused on leveraging 2D vision-language model features for 3D reconstruction due to the absence of large-scale 3D datasets. Techniques like OpenScene [31] and ConceptFusion [19] utilize pixel-wise CLIP features [32] for text-aligned 3D feature extraction. LERF [22] employs similar CLIP-based [32] semantic fields within NeRF [28] frameworks, providing query-specific scene heatmaps but limited object instance understanding. Instance-based methodologies [10, 17, 27, 36, 43] like SAM3D [43] and OpenMask3D [36], project 2D masks onto 3D point clouds for enhanced instance recognition, though challenges remain in instance merging and quality of 3D mask proposals. OVIR-3D [27] attempts to improve instance merging but struggles with large instance backgrounds. Our method overcomes these obstacles by integrating high-performing pre-trained models from both 2D and 3D domains, reducing reliance on single-modality insights.

3. Method

The Zero-Shot Dual-Path Integration Framework is designed to predict class-agnostic 3D instance masks and their corresponding CLIP-based features for open-vocabulary instance classification, as illustrated in Fig. 3. This framework operates on posed RGB-D images and reconstructed 3D point clouds, leveraging queries to isolate class-specific instance masks through a dual-pathway approach: a 3D pathway for mask proposal generation with point clouds and a 2D pathway for RGB-D based mask proposal generation.

In our approach, the 3D pathway employs a mask proposal network to generate 3D instance masks, integrating 2D bounding boxes and CLIP-based features to bridge open-vocabulary concepts with these masks, leveraging architectures adept at detecting sizeable objects [34]. Simultaneously, the 2D pathway applies an open-vocabulary 2D instance segmentation network to RGB-D images, creating 2D mask proposals. These are then projected into the 3D point cloud, refined into finalized instance masks by an Instance Fusion Module.

Integration of these pathways is achieved through our Conditional Integration in the Dual-Path Integration phase, which encompasses Dual-modality Proposal Matching and Adaptive Integration. This process begins with the matching of overlapping proposals from both pathways using Intersection-of-Union (IoU) metrics, followed by the conditional merging of these proposals based on the assessment of their IoU. This integration technique enhances the framework's accuracy in object recognition and segmentation by leveraging the complementary strengths of both 2D and 3D data.

3.1. 3D Pathway: 3D Mask Proposals from Point Cloud

Given a 3D point cloud, denoted as $\mathbf{P} \in \mathbb{R}^{N\times 3}$ where N represents the total number of points, each point is represented by a 3D position. The 3D pathway's objective is

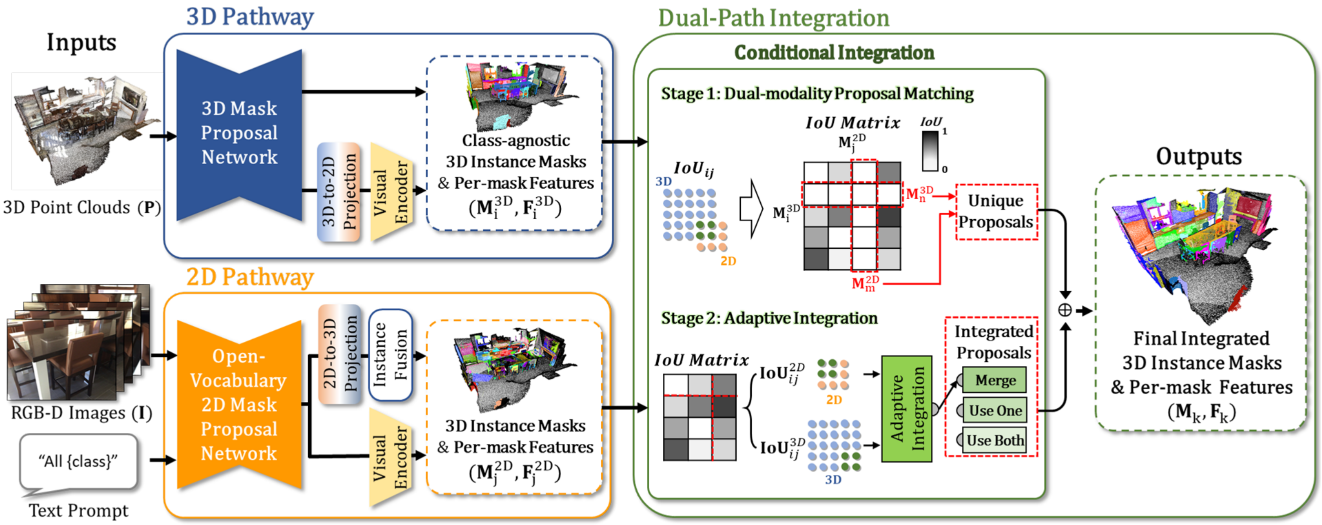

Figure 3. Overview of our Zero-Shot Dual-Path Integration Framework. The 3D pathway takes 3D point cloud $\mathbf{P}$ as input to generated class-agnostic 3D instance masks $\mathbf{M}_\mathrm{i}^{\mathrm{3D}}$ with pre-trained 3D Mask Proposal Network and the per-mask visual features $\mathbf{F}_\mathrm{i}^{\mathrm{3D}}$ are extracted with CLIP visual encoder [32]. The 2D pathway also generates its own 3D instance masks $\mathbf{M}_\mathrm{j}^{\mathrm{2D}}$ using RGB-D Image $\mathbf{I}$ input with Open-vocabulary 2D Mask Proposal Network and 2D-to-3D Projection module, along with the per-mask visual features $\mathbf{F}_\mathrm{i}^{\mathrm{2D}}$ of each mask. The outputs of two pathways are integrated through the Conditional Integration which utilizes Intersection-of-Union (IoU) for Dual-modality Proposal Matching and Adaptive Integration, having final 3D instance results $\mathbf{M}_\mathrm{k}$ and their visual features $\mathbf{F}_\mathrm{k}$ as outputs.

to segment the given point cloud into class-agnostic 3D instance mask proposals, represented through a collection of binary masks, where each mask is denoted as $\mathbf{M}_\mathrm{i}^{\mathrm{3D}} = (M_{\mathrm{i},1}^{\mathrm{3D}}, ..., M_{\mathrm{i,N}}^{\mathrm{3D}})$ where $M_{\mathrm{i,n}}^{\mathrm{3D}} \in \{0, 1\}$, meaning n-th point belongs to i-th object instance. For generating the 3D instance masks, Mask3D [34] is used as our 3D instance segmentation network, which utilizes U-Net style sparse convolutional backbone [9] as a feature extractor. A fixed number of object queries go through the transformer decoder layers to attend to global features iteratively, directly outputting instance predictions. This generates a binary mask for each instance with predicted class labels and their confidence scores. However, since we want a class-agnostic mask proposal network, Mask3D [34] is modified to omit the predicted class labels and confidence scores to only concentrate on generating binary instance mask proposals. This way, we obtain open-vocabulary representations instead of semantic class predictions confined to closed-vocabulary.

Then, we derive per-mask text-aligned visual features using pre-trained CLIP [14] for querying open-vocabulary concepts associated with predicted instance masks. Inspired by [36], we first select the top k RGB-D images with the highest visibility of each instance mask. In prior work [36], points projected onto the images are utilized as the prompts to guide the SAM [23] in generating the 2D bounding boxes, which are utilized for cropping the images for CLIP feature extraction. However, this projection can cause errors due to approximation in the occlusion test, and the random selection of the k projected points might erroneously include points lying outside the actual instances, leading to the generation of poor-quality bounding boxes. Consequently, the extracted CLIP features from those poor-quality bounding boxes suffer from inadequacies. To bypass this issue, we directly use the projected 3D bounding box into 2D images to get CLIP-based features, eliminating potential errors in random point selection. We employ multi-level crops of specific regions for feature enrichment to encapsulate extensive contextual details from the surrounding environment. Leveraging the CLIP visual encoder [32], the CLIP feature vectors $\mathbf{F}_\mathrm{i}^{\mathrm{3D}} \in \mathbb{R}^d$ of the cropped object images are extracted and average-pooled to generate final mask-feature representations for each object.

3.2. 2D Pathway: 3D Mask Proposals from Multi-view RGB-D Images

The objective of 2D pathway is to generate 3D instance masks, where each mask is denoted as $\mathbf{M}_\mathrm{j}^{\mathrm{2D}} = (M_{\mathrm{j},1}^{\mathrm{2D}}, ..., M_{\mathrm{j,N}}^{\mathrm{2D}})$, where $M_{\mathrm{j,n}}^{\mathrm{2D}} \in \{0, 1\}$, meaning n-th point belongs to j-th object instance, from RBG-D image $\mathbf{I}_\mathrm{t}$ where t is the image frame at time t with known camera intrinsic matrix C and world-to-camera extrinsic matrix (pose) E_t. For this process, we first utilize Grounded-SAM, a fusion of Grounding DINO [26] and SAM [23], as pre-trained 2D open-vocabulary instance segmentation network to obtain 2D mask proposals $\mathbf{m}_{\mathrm{t,j}}^{\mathrm{2D}}$. Grounding DINO takes the text prompt as an input to produce the 2D bounding boxes, which SAM subsequently uses to obtain 2D mask proposals. These 2D mask proposals $\mathbf{m}_{t,j}^{2D}$ undergo subsequent projection into the 3D point cloud with known camera intrinsic, pose, and depth. Additionally, we extract CLIP-based features $\mathbf{F}_{\mathrm{j}}^{\mathrm{2D}} \in \mathbb{R}^d$ from the corresponding cropped image for each proposal.

Given that mask proposals sourced from 2D images may be fragmented due to occlusion, these proposals from the 2D pathway are passed to an Instance Fusion process for complete proposals. Drawing upon methodologies from [27], the Instance Fusion Module accumulates 3D projected instances within a memory bank. It then periodically executes a filtering and merging process, utilizing the 3D Intersection-of-Union metric in conjunction with feature similarity analysis.

3.3. Dual-Path Integration

Our Dual-path Integration framework incorporates Conditional Integration that operates through a meticulously structured two-stage process. First, Dual-modality Proposal Matching stage utilizes the Intersection-of-Union (IoU) metrics to effectively identify unique proposals from each respective modality. Given 3D instances from 3D point cloud $\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}$ and 2D multi-view images $\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$ where i and j are the number of generated instances for each pathway, the module systematically calculates the Intersection-of-Union (IoU), denoted as $\mathbf{IoU}_{ij}$ as Equation 1:

$$\mathbf{IoU}_{ij} = \frac{|\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}} \cap \mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}|}{|\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}} \cup \mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}|}. \tag{1}$$

This computation is conducted for each possible pair of instances across the modalities. As a result, IoU matrix providing a comprehensive representation of the spatial relationships between all instances across the 3D and 2D modalities is created.

To identify unique proposals from each modality for inclusion in the final instance proposals $\mathbf{M}_{\mathrm{k}}$, we systematically evaluate instances from $\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$ and $\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}$ against the IoU matrix. This process aims to detect instances without overlap across the entirety of instances from the alternate modality. This evaluation can be articulated as follows:

$$\forall j, \text{if } (\forall i,\ \mathbf{IoU}_{ij} = 0), \text{ then add } \mathbf{M}_{\mathrm{j}}^{\mathrm{2D}} \text{ to } \mathbf{M}_{\mathrm{k}}$$
$$\forall i, \text{if } (\forall j,\ \mathbf{IoU}_{ij} = 0), \text{ then add } \mathbf{M}_{\mathrm{i}}^{\mathrm{3D}} \text{ to } \mathbf{M}_{\mathrm{k}}.$$

This procedure ensures that any instance $\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$ lacking overlap with all $\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}$ instances (indicated by an $\mathbf{IoU}_{ij}$ value of 0) is directly incorporated into $\mathbf{M}_{\mathrm{k}}$. Conversely, it also guarantees the inclusion of any $\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}$ instance that does not overlap with all $\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$ instances into $\mathbf{M}_{\mathrm{k}}$. Additionally, proposal pairs exhibiting the smallest $\mathbf{IoU}_{ij}$ for all instances $\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$ are added to $\mathbf{M}_{\mathrm{k}}$, aiming to enrich the segmentation with a broader array of detected objects.

Excluding the unique instances identified in the first stage above, our Conditional Integration progresses to the Adaptive Integration stage, where it performs additional computation of IoUs bifurcated into two distinct perspectives, as in Equations 2:

$$\mathbf{IoU}_{ij}^{3D} = \frac{|\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}} \cap \mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}|}{|\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}|}, \quad \mathbf{IoU}_{ij}^{2D} = \frac{|\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}} \cap \mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}|}{|\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}|}. \tag{2}$$

Equation 2 (left) introduces $\mathbf{IoU}_{ij}^{3D}$, which quantifies the proportion of overlap between an instance from 3D pathway $\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}$ and a instance from 3D pathway $\mathbf{M}_{\mathrm{i}}^{\mathrm{3D}}$ relative to the entire instance from 3D pathway. Conversely, Equation 2 (right) defines $\mathbf{IoU}_{ij}^{2D}$ as the ratio of their intersection to the total area of the instance from 2D pathway$\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$. Together, these IoU metrics offer a dual perspective on the relations between pairs of instances from the 3D pathway and the 2D pathway. By analyzing the extent to which an instance from one pathway encompasses the spatial domain of the instance from another pathway, it offers insights into the priority between two proposals.

Subsequently, we select the proposal pair with highest $\mathbf{IoU}_{ij}$ for each proposals of $\mathbf{M}_{\mathrm{j}}^{\mathrm{2D}}$ and assess them into four scenarios for Adaptive Integration: (1) high IoU for both $\mathbf{IoU}_{ij}^{2D}$ and $\mathbf{IoU}_{ij}^{3D}$, (2) low IoU for both, (3) high IoU for the $\mathbf{IoU}_{ij}^{2D}$ but low for $\mathbf{IoU}_{ij}^{3D}$ (e.g. proposal from the 2D pathway is subgroup of that from the 3D pathway), and (4) the vice-versa. By selecting the proposal pair exhibiting the highest $\mathbf{IoU}_{ij}$, we select the proposal from the alternate pathway that demonstrates the most significant relationship, encapsulating both concordance and discordance, for further evaluation.

The four distinct scenarios of Adaptive Integration are elaborated below:

1. **Significant overlap**: For proposal pairs exhibiting extensive overlap, it is inferred that they likely depict the same object. Thus, these proposals are merged into a singular, comprehensive proposal for inclusion in $\mathbf{M}_{\mathrm{k}}$, ensuring a unified representation.
2. **Slight overlap**: When a pair demonstrates only slight overlap, yet selected beforehand for having maximum overlap among all considered pairs, it is surmised that the proposals likely denote two distinct objects in close proximity. Accordingly, both proposals are maintained separately in $\mathbf{M}_{\mathrm{k}}$, preserving the individuality of each detected object.
3. **Proposal from 2D is subgroup of proposal from 3D**: In instances where a proposal from 2D pathway is almost entirely encompassed by a proposal from 3D pathway, it is treated as a unique finding exclusive to the 2D pathway and thus given precedence for inclusion into $\mathbf{M}_{\mathrm{k}}$. This decision is based on the assumption that the proposal from 2D modality may highlight a detail or aspect not captured from the 3D modality. Although the

larger overlapping proposal is neglected in this instance, its potential value is recognized. We anticipate that it will align with nearby 2D pathway proposals, thereby being considered under different scenarios.

4. **Proposal from 3D is subgroup of proposal from 2D**: Analogous to scenario (3) with reversed roles, the proposal from 3D pathway is prioritized for addition to $\mathbf{M}_\mathrm{k}$. This reflects the broader spatial coverage and potentially significant detection afforded by the 3D pathway.

Through these scenario-specific strategies, our integration process adeptly balances the quantity and quality of proposals offered by both pathways, enhancing the overall accuracy and completeness of instance segmentation. To differentiate between 'high' and 'low' IoU values, thresholds for each pathway are designated as θ_{2D} and θ_{3D}, respectively, with their optimal values determined through empirical experimentation. Leveraging visual features derived from 3D point clouds and 2D multi-view images, $\mathbf{F}^{\mathrm{3D}}_{\mathrm{i}}$ and $\mathbf{F}^{\mathrm{2D}}_{\mathrm{j}}$, we employ an averaging process when merging two masks. After the Adaptive Integration, we obtain the final instance proposals $\mathbf{M}_\mathrm{k}$ and the corresponding CLIP-based features $\mathbf{F}_\mathrm{k}$ ready to perform a semantic label assignment.

During the inference phase, a textual query denoted as q correlates with a repository of representative features linked to individual 3D instances. The text feature $\mathbf{F}_\mathrm{q}$ will be extracted from the CLIP encoder to compare with instance features $\mathbf{F}_\mathrm{k}$ in the scene. Subsequently, these instances denoted as $\mathbf{M}_\mathrm{k}$, undergo a ranking process based on their resemblance to the query. The top-ranked instances, thus determined, are retrieved and subsequently returned.

4. Experiments

In this section, we present both quantitative and qualitative outcomes of our Zero-shot Dual-Path Integration Framework. We conduct a quantitative assessment, comparing our open-vocabulary 3D instance segmentation method against existing approaches within a closed-vocabulary setting. Ablation studies further dissect the impact of baseline architectures for each pathway, alongside the examination of metrics and thresholds pivotal to our Dual-path Integration process. Additionally, we showcase qualitative results from the ScanNet200 and ARKitScenes datasets to underscore our method's effectiveness in open-vocabulary 3D instance segmentation, demonstrating its proficiency in accurately identifying a wide spectrum of objects.

4.1. Dataset and Metric

Dataset. Our evaluation framework employs the ScanNet200 benchmark dataset [4], utilizing its validation set of 312 unique scenes for 3D instance segmentation performance assessment across a closed vocabulary of 200 categories. Further, we adopt the categorization scheme by Rozenberszki et al. [33], dividing the ScanNet200 object classes into "head", "common", and "tail" subsets, with 66, 68, and 66 categories respectively, to analyze model performance across varying object occurrence frequencies. Additionally, we leverage the ARKitScenes dataset [1], which consists of over 5K scans from approximately 1.6K diverse indoor settings, offering 3D mesh reconstructions, RGB and depth images, and ARKitSLAM-estimated camera poses. This dataset aids in simulating realistic indoor scanning trajectories. Performance analysis is further enhanced by employing queries from OpenSUN3D [7] within the Challenge development set, demonstrating the effectiveness of our approach in advanced indoor scene understanding tasks.

Metric. In our evaluation, we adopt the widely recognized metric for 3D instance segmentation: average precision (AP). The AP scores are calculated at mask overlap thresholds of 50%, 25%, and average over the overlap range of [0.5 : 0.95 : 0.05], in line with the ScanNet [4] evaluation protocol. Furthermore, we analyze the AP scores across the "head", "common", and "tail" subsets of ScanNet200. This allows us to gain deeper insights into the performance of our method across different frequency categories.

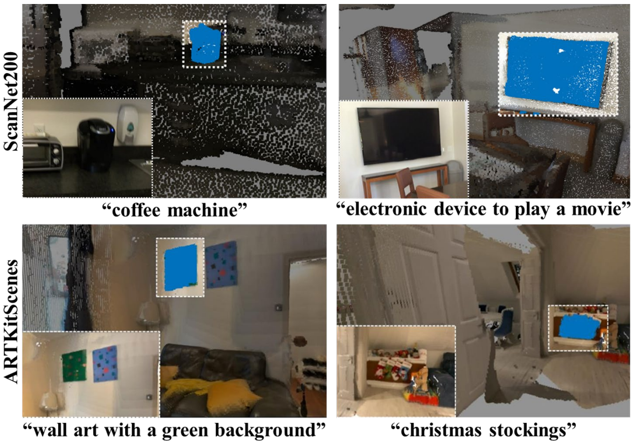
ScanNet200
"coffee machine"
"electronic device to play a movie"
ARTKitScenes
"wall art with a green background"
"christmas stockings"

Figure 4. Qualitative results showcasing the proficiency of our framework in performing open-vocabulary 3D instance segmentation. The displayed results include objects from two distinct datasets: the upper two objects are from ScanNet200 scenes, while the lower two are from ARKitScenes, demonstrating our framework's adaptability and effectiveness across diverse environments.

4.2. Experimental Details

In our experiments, we conducted computations using a single RTX 8000 GPU. We utilized posed RGB-D pairs from the ScanNet200 dataset, processing one frame out of every ten frames within the RGB-D sequences. To extract image features from mask crops, we employed the CLIP visual encoder [32] from the ViT-L/14 model, known for its feature dimensionality of 768. For Adaptive Integration, the

Model	AP ↑	AP_{50} ↑	AP_{25} ↑	head (AP) ↑	common (AP) ↑	tail (AP) ↑
SAM3D [43]	6.1	14.2	21.3	7.0	6.2	4.6
OpenScene [31]	11.7	15.2	17.8	13.4	11.6	9.9
OVIR-3D [27]	13.0	24.9	**32.3**	14.4	12.7	11.7
OpenMask3D [36]	15.4	19.9	23.1	17.1	14.1	**14.9**
Dual-Path (Ours)	**19.9**	**25.0**	27.1	**27.9**	**18.9**	11.5

Table 1. Evaluation of Open-vocabulary 3D instance segmentation on the 312 scenes of ScanNet200 [4] validation set. Average Precision (AP) measured over a range of overlap thresholds, 50% overlaps, and 25% overlaps. AP of “head”, “common”, and “tail” subsets [33] of ScanNet200 are also reported.

IoU thresholds were empirically determined to be optimal at $\theta_{2D} = 0.5$ and $\theta_{3D} = 0.9$.

4.3. Quantitative Results

In the comprehensive evaluation presented in Table 1, performance of our approaches in closed-vocabulary instance segmentation tasks within the ScanNet200 [4] benchmark is provided. This distinction in performance is particularly marked within the “head” and “common” categories, while the disparity narrows in the “tail” categories.

For the previous works on open-vocabulary models, OpenScene [31] is constructed based on 2D model OpenSeg [8] trained on labeled datasets for 2D semantic segmentation. OpenMask3D [36], a state-of-the-art open-vocabulary model, is built upon the Mask3D for generating class-agnostic 3D mask proposals. Compared to these previous methods, our Dual-Path Integration Framework has a distinct performance advantage in AP. This outcome substantiates our initial hypothesis about the efficacy of our method.

4.4. Qualitative Results and Comparisons

Fig. 4 presents the qualitative results that underscore the efficacy of our proposed framework within both seen (ScanNet200) and unseen (ARKitScenes) data, thereby affirming the framework’s extensive adaptability and proficiency across diverse environments. Being a zero-shot open-vocabulary framework, it enables the segmentation of objects through free-form text queries, even for objects absent in traditional instance segmentation datasets.

In Figure 5, we provide qualitative comparisons that highlight the distinction between our framework and the OpenMask3D in segmenting uncommon objects from the “tail” category of ScanNet200, alongside unseen objects not present in the dataset’s predefined categories. These comparisons underscore our proposed framework’s enhanced proficiency in accurately segmenting objects that have posed challenges to previous methods, which predominantly leveraged 3D point cloud data for instance segmentation.

4.5. Ablation Studies

We conducted an ablation study to assess the individual contributions of components within our Dual-Path Integration Framework across 312 scenes from the ScanNet200 validation set, as outlined in Table 2. Our 3D pathway yielded superior performance compared to the reported results in OpenMask3D [36], primarily due to our method of directly using the projected 3D bounding box into 2D images to get CLIP-based features for eliminating potential errors in random point selection. For 2D pathway, relying solely on the 2D data by using 3D projected CLIP [32] resulted in suboptimal performance. While the ’Simple Integration’ method of using all proposals from both the 3D and 2D pathways achieves enhanced performance, it falls short of adequately addressing the low quality and redundancy problem of the proposals, failing to leverage the intrinsic strengths unique to each pathway. In contrast, our Dual-path Integration strategy employs a selective and adaptive approach to integrate the two modalities, yielding improvements over Simple Integration.

In Table 3, the recall rate of “tail” category of the ScanNet200 validation set is reported. It highlights the capability of our Dual-Path Integration in identifying uncommon objects. A significant factor contributing to the enhancement of the performance is the robust generalization ability offered by the vision-language understanding capabilities inherent in the pre-trained image classifier of the 2D pathway. The notable improvement in the recall rate when comparing the Dual-path with the 3D pathway indicates the efficacy of the 2D pathway in our method.

Table 4 presents an ablation study on the impact of the IoU thresholds θ_{3D} and θ_{2D} within the Adaptive Integration process. The analysis reveals that the Average Precision at a threshold of 25% (AP_{25}) remains relatively unaffected by variations in these thresholds. In contrast, both AP_{50} and the mean Average Precision (mAP) exhibit sensitivity to changes in thresholds, suggesting that a higher threshold is crucial for refining segmentation precision. This pattern underscores the importance of carefully calibrating the IoU thresholds to optimize the overall segmentation performance and achieve a balance between recall and precision.

3D pathway	2D pathway	Simple Integration	Dual-path Integration	AP ↑	AP_{50} ↑	AP_{25} ↑
✓				16.2	20.7	22.7
	✓			10.9	15.6	20.3
✓	✓	✓		18.2	23.8	25.8
✓	✓		✓	**19.9**	**25.0**	**27.1**

Table 2. Ablation study on contributions of each component within our Zero-shot Dual-Path Integration Framework in 312 scenes of ScanNet200 [4] validation set.

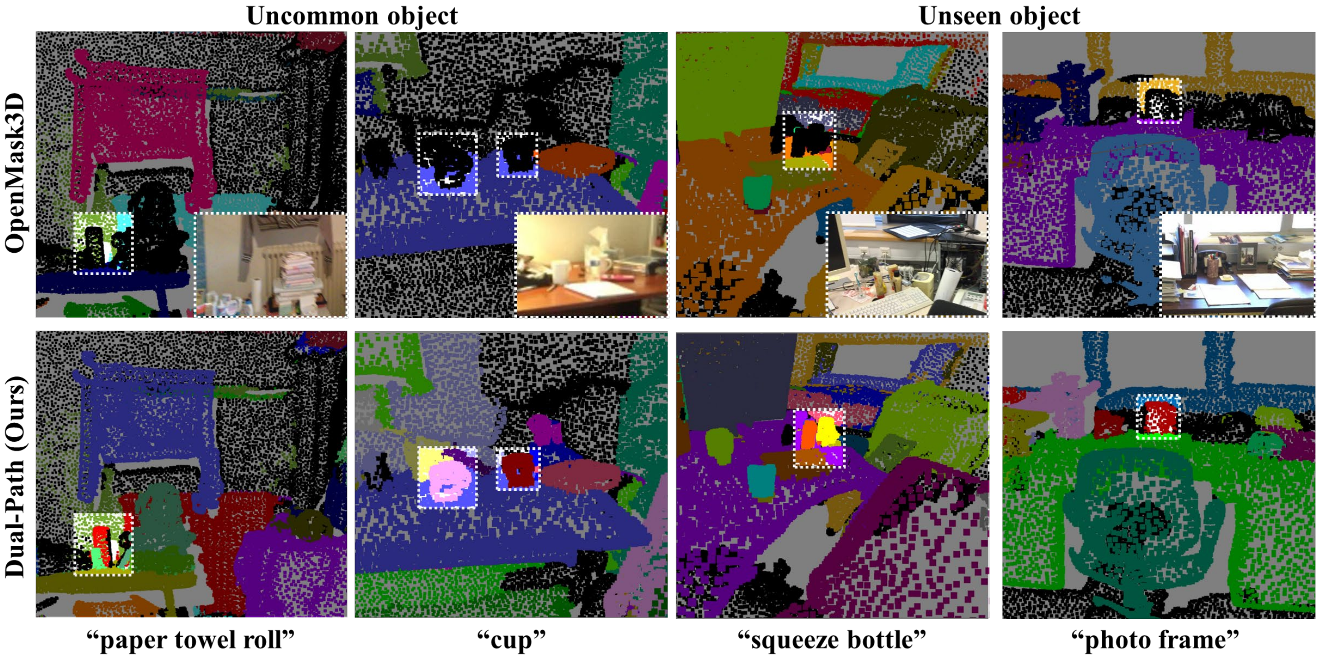

Figure 5. Qualitative comparison between our Dual-Path Integration Framework and OpenMask3D. The black regions indicate no proposals. Our framework, benefiting from the integration of proposals endowed with high visual understanding capabilities from the 2D pathway, excels in identifying and segmenting uncommon and unseen objects.

Model	RC ↑	RC_{50} ↑	RC_{25} ↑
3D pathway	14.2	19.5	22.7
2D pathway	17.8	25.7	**32.9**
Dual-Path	**21.0**	**27.9**	31.0

Table 3. Ablation study on the Recall Rate (RC) measured over a range of overlap thresholds, 50% overlaps, and 25% overlaps of "tail" category in 312 scenes of ScanNet200 [4] validation set.

θ_{3D}	θ_{2D}	mAP ↑	AP_{50} ↑	AP_{25} ↑
0.25	0.25	18.7	24.3	**27.2**
0.50	0.50	19.5	**25.0**	**27.2**
0.90	0.50	19.2	24.8	27.0
0.50	0.90	**19.9**	**25.0**	27.1
0.90	0.90	19.8	25.0	27.0

Table 4. Ablation study on IoU threshold θ_{3D} and θ_{2D} for Adaptive Integration on ScanNet200 [4] validation set.

5. Conclusion

This paper proposes the Zero-Shot Dual-Path Integration Framework, a model-agnostic strategy that harnesses mask proposals from pretrained models across 3D point cloud and 2D multi-view image modalities for open-vocabulary 3D instance segmentation. Our novel Conditional Integration process capitalizes on the strengths of instance segmentation within each modality through a two stage methodology: Dual-modality Proposal Matching and Adaptive Integration, aimed at identifying and categorizing significant proposal pairs into distinct categories for effective integration of results from two different modalities. Evaluations conducted on the ScanNet200 benchmark dataset and ARKitScenes dataset illustrate our framework's substantial improvements over prior methodologies, validating the efficacy of integrating 3D and 2D segmentation techniques.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 6

[2] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *ICCV*, 2021. 1, 3

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 3

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

[5] 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. 3

[6] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[7] Francis Engelmann, Ayca Takmaz, Jonas Schult, Elisabetta Fedele, Johanna Wald, Songyou Peng, Xi Wang, Or Litany, Siyu Tang, Federico Tombari, et al. Opensun3d: 1st workshop challenge on open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2402.15321*, 2024. 6

[8] 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. 3, 7

[9] 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. 4

[10] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv*, 2023. 3

[11] 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. 3

[12] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *Proceedings of the 2022 Conference on Robot Learning*, 2022. 3

[13] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3

[14] 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. 4

[15] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019. 1, 3

[16] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. 3

[17] 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. 2, 3

[18] 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. 3

[19] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *arXiv*, 2023. 3

[20] 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. 3

[21] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[22] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[23] 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. 3, 4

[24] 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. 3

[25] 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. 1, 3

[26] 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. 3, 4

[27] 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. 3, 5, 7

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[29] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*. 3

[30] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 1, 3

[31] 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, 7

[32] 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. 3, 4, 6, 7

[33] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 6, 7

[34] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. 2023. 1, 3, 4

[35] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation, 2022. 1, 3

[36] 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. 2, 3, 4, 7

[37] 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. 3

[38] 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. 1, 3

[39] 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. 3

[40] 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. 1, 3

[41] 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. 3

[42] 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. 1, 3

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

[44] L. Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3942–3951, 2018. 1, 3

[45] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[46] Weiguang Zhao, Yuyao Yan, Chaolong Yang, Jianan Ye, Xi Yang, and Kaizhu Huang. Divide and conquer: 3d point cloud instance segmentation with point-wise binarization. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 562–571, 2023. 1, 3

[47] 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. 3