Edge-Aware 3D Instance Segmentation Network with Intelligent Semantic Prior

Wonseok Roh[1] Hwanhee Jung[1] Giljoo Nam[2] Jinseop Yeom[1]
Hyunje Park[1] Sang Ho Yoon[3] Sangpil Kim[1]*

[1]Korea University [2]Meta Reality Labs [3]KAIST

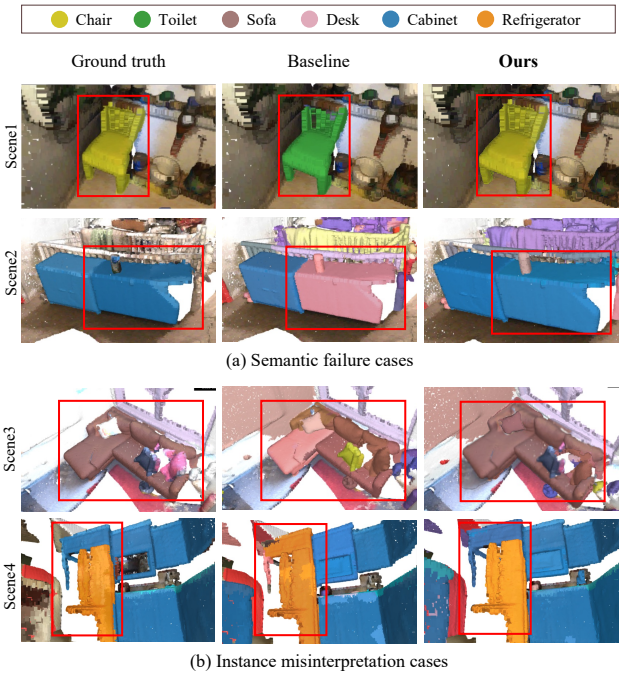
Abstract

*While recent 3D instance segmentation approaches show promising results based on transformer architectures, they often fail to correctly identify instances with similar appearances. They also ambiguously determine edges, leading to multiple misclassifications of adjacent edge points. In this work, we introduce a novel framework, called **EASE**, to overcome these challenges and improve the perception of complex 3D instances. We first propose a semantic guidance network to leverage rich semantic knowledge from a language model as intelligent priors, enhancing the functional understanding of real-world instances beyond relying solely on geometrical information. We explicitly instruct the basic instance queries using text embeddings of each instance to learn deep semantic details. Further, we utilize the edge prediction module, encouraging the segmentation network to be edge-aware. We extract voxel-wise edge maps from point features and use them as auxiliary information for learning edge cues. In our extensive experiments on large-scale benchmarks, ScanNetV2, ScanNet200, S3DIS, and STPLS3D, our EASE outperforms existing state-of-the-art models, demonstrating its superior performance.*

1. Introduction

Understanding 3D scenes is a fundamental task within 3D computer vision. Given 3D point cloud scenes, identifying and perceiving instances on sparse points, along with assigning semantic class labels, play a crucial role in a comprehensive understanding of the entire spatial environment.

In real-world 3D scenarios, significant occlusion and truncation often occur, especially when objects are overlapped or hidden by others. To address these issues, conventional works [3, 7, 8, 16, 18, 40, 43, 44] in 3D instance segmentation (3DIS) primarily focus on accurately generating region proposals (top-down) [16, 43, 44] or effectively grouping points with clustering algorithms (bottom-up) [3, 7, 8, 18, 40]. Inspired by the remarkable sensation of Mask-RCNN [14], the former (proposal-based) methods initially detect instances as bounding boxes and predict masks within each proposed region. However, these strategies are susceptible to the quality of the detection results. On the other hand, the latter (grouping-based) methods use clustering algorithms to group closely related points and aggregate point-wise class labels and instance features. While they show great advancements in 3DIS, these methods require additional manually tuned processing, such as point grouping [18] or voting mechanisms [31], to determine specific geometric properties (*e.g.*, centers, occupancy).



(a) Semantic failure cases

(b) Instance misinterpretation cases

Figure 1. Examples of 3DIS for two challenging cases. (a) Semantic failure cases: the baseline [35] confuses seat-shaped objects as *chair* or *toilet* (Scene1) and also misclassifies large cuboid blocks as *desk* (Scene2). And (b) Instance misinterpretation cases: baseline erroneously segments a single object into multiple parts (Scene3) or merges multiple objects into one instance (Scene4).

Recently, transformer-based 3DIS frameworks [22, 26, 35, 36] have tackled several limitations of traditional approaches by introducing a fully end-to-end pipeline. They directly predict masks from 3D points without resource-

*Corresponding author.
Project Page: https://kuai-lab.github.io/ease2024.

exhaustive processing. These methods have achieved high performance and quickly become predominant in the 3DIS domain. However, despite these breakthroughs, they still face challenges in accurately masking everyday 3D instances relying solely on geometrical information. Specifically, they frequently suffer from cases where each instance appears similarly, but their semantic roles differ, as illustrated in Fig. 1 (a). Thus, gaining insight into the semantic properties of instances beyond mere appearance or position is crucial for effectively addressing complex real-world 3D scenarios. From this observation, we consider leveraging semantic knowledge along with the geometry information.

Built upon the classic structure of the transformer-based framework [22, 35, 36], which trains queries that encode instance-specific knowledge for direct mask prediction, we introduce a carefully designed semantic guidance network. One of the most promising ways to utilize semantic information is through visual language models [17, 32]. They learn generous visual-linguistic knowledge from large-scale image-text pairs, which empowers them to achieve outstanding progress in various 2D or 3D vision tasks [9, 11, 23, 29]. In this work, motivated by the success of these works, we explicitly instruct the transformer-based network to learn contextual variations among instances using text embeddings. Specifically, we strengthen the basic instance queries with profound semantic clues of text embeddings representing each instance. Here, the context details can serve as discriminative semantic priors for perceiving complex 3D instances adequately. Through our extensive experiments, we found that our semantic-guided approach effectively overcomes the challenges of existing methods.

Different from the human vision system, which easily distinguishes boundaries between instances, we empirically observe that most 3DIS models [21, 22, 35, 36, 42] struggle with precisely capturing the boundaries in the 3D space. These methods often misinterpret the spatial range of instances and ambiguously determine their edges, leading to numerous misclassifications of adjacent edge points, as shown in Fig. 1 (b). We consider that these issues arise due to the absence of detailed guidance on inter-object boundaries, resulting in fuzzy edges. To tackle these challenges, we advocate for leveraging the edge prediction module to pilot the whole network to exploit edge-advanced features. We predict edge points across point cloud scenes and operate them as auxiliary information for learning practical edge cues, which encourages accurate recognition of the 3D spatial scope. Note that we supervise this module using dynamically generated point-wise pseudo edge labels.

Given landmark datasets for 3DIS, ScanNetV2 [4], ScanNet200 [33], S3DIS [1], and STPLS3D [2], we validate the effectiveness and robustness of our framework. Our method outperforms the existing state-of-the-art methods. To summarize, our main contributions are listed as follows:

- We propose **EASE**, a novel 3D instance segmentation framework that utilizes the rich semantic clues from the language model. Specifically, context details of text embeddings serve as smart semantic priors, effectively enhancing the functional comprehension of 3D instances.
- We introduce the edge prediction module guiding the entire network to take advantage of edge-aware features. Also, we utilize the module's output as auxiliary knowledge for learning edge cues, boosting 3DIS performance.
- We analyze the effectiveness of our proposed method on multiple challenging benchmarks, including ScanNetV2, ScanNet200, S3DIS, and STPLS3D. Extensive experiments on various 3D scenarios validate that our method achieves new State-of-the-Art performance on 3DIS.

2. Related Work

3D Instance Segmentation. The 3D Instance Segmentation (3DIS) task aims to identify individual instances and assign semantic classes to each point within a 3D point scene. Traditional approaches in 3DIS are primarily categorized into three groups: proposal-based [16, 43, 44], grouping-based [3, 7, 8, 18, 40], and transformer-based methods [22, 35, 36]. The proposal-based approaches initially detect object proposals, such as 3D bounding boxes, and then estimate corresponding per-point semantic classes and instance masks. One of the grouping-based approaches, PointGroup [18], conducts point-wise clustering to group points into instances using dual point sets, which comprise original and shifted coordinates. Recently, transformer-based approaches predict instance queries that represent individual objects and decode them into per-point categories and instance labels directly. Mask3D [35] refines queries using masked cross-attention with hierarchical point features, leading the queries to focus on particular instances. SPFormer [36] updates instance queries via transformer decoder utilizing pre-computed superpoints, which can reduce the whole computational cost of the network. MAFT [22], on the other hand, recognizes representational limitations in the initial mask and replaces the masking process with object-localizing position queries. Based on the core idea of transformer-based methods, our novel approach incorporates a semantic guidance network and an auxiliary edge network to improve instance mask segmentation.

Semantic Guided 3D Scene Understanding. Recent research in 3D scene understanding focuses on comprehending large-scale, real-world 3D environments. Conventional studies in this field are broadly categorized into two strategies: closed-set 3D scene understanding, a prevalent approach in diverse tasks [19, 27, 33, 34], and open-vocabulary methods [12, 29, 37], which focus on recognizing unseen categories within datasets. More recently, there has been significant progress in integrating pre-trained visual-language models like CLIP [32] or ALIGN [17] into

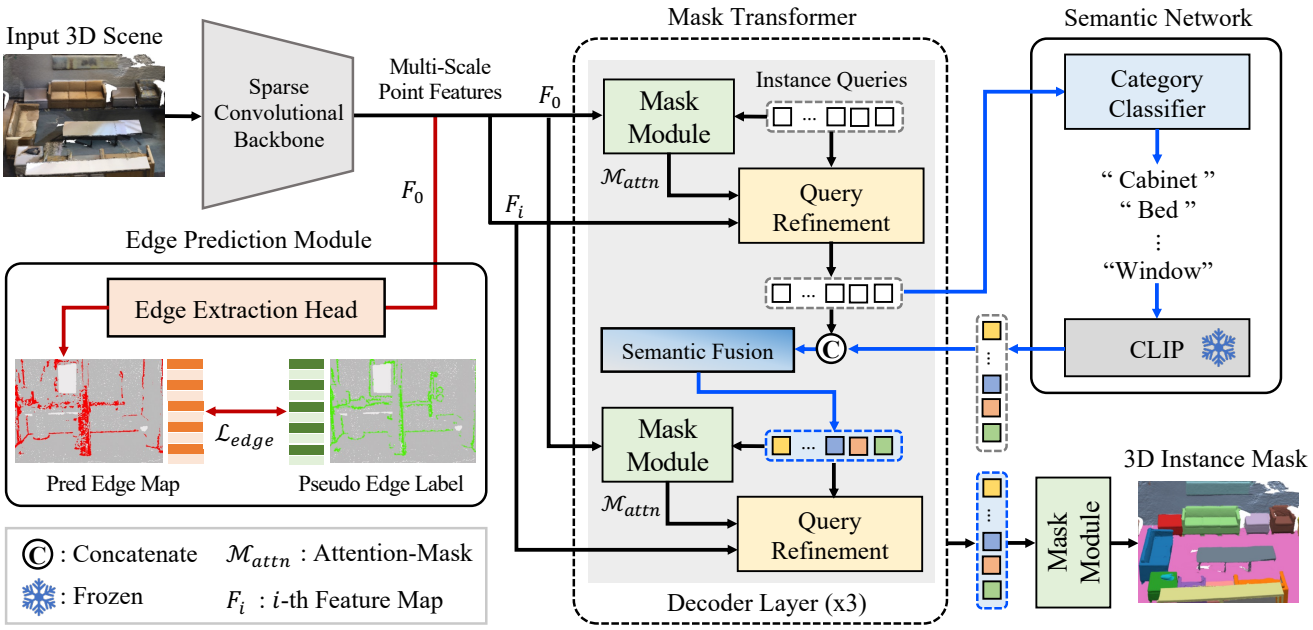


Figure 2. An overview of our framework EASE. Built upon the classic architecture of the transformer-based 3DIS model, our model takes 3D scenes and directly infers instance masks. Our model consists of four main modules: (1) Sparse Convolutional Backbone, (2) Semantic Network, which strengthen the queries to learn contextual variations of instances with rich semantic embeddings from the pre-trained language model CLIP, (3) Mask Transformer Decoder with Mask Module and Query Refinement blocks, and (4) Edge Prediction Module, where our edge extraction head is trained to predict a voxel-wise edge map, encouraging the network to utilize edge-advanced features.

these methodologies to enhance semantic understanding. Building on this idea, for instance, [33] utilizes a closed-set framework and enhances 3D feature learning by aligning it with CLIP text embedding space for semantic classification. In this work, we employ contextual semantic priors to improve the comprehension of complex 3D instances, enhancing the overall 3D instance segmentation performance.

3. Method

In this section, we introduce a novel 3D Instance Segmentation (3DIS) framework EASE leveraging the rich semantic guidance from the language model. Furthermore, we propose the edge prediction module, which encourages the pipeline to use edge-aware features for effectively perceiving instance boundaries. We provide an overview of the entire pipeline (Sec. 3.1, Fig. 2, and Alg. 1) and then elaborate on the details of our method: (i) Intelligent Semantic Prior (Sec. 3.2) and (ii) Edge-Aware 3DIS Framework (Sec. 3.3).

3.1. Overview

Our end-to-end 3DIS framework, as illustrated in Fig. 2 and Alg. 1, aims to enhance the comprehension of various instances across 3D point cloud scenes. Based on the classic transformer-based architecture [22, 35], which operates standard transformer building blocks, we directly infer instance masks from 3D point clouds. Our model consists of four main modules: (1) Sparse Convolutional Backbone, (2) Semantic Network, which provide the deep semantic cues of diverse real-world instances, (3) Mask Transformer Decoder with Mask Module and Query Refinement blocks, and (4) Edge Prediction Module, which benefits the whole network to take advantage of edge-aware features.

Mask Transformer. First, the Sparse Convolutional U-Net Backbone takes a colored point cloud $P \in \mathbb{R}^{N_p \times 6}$ as input and voxelizes P into X_0 voxels $V \in \mathbb{R}^{X_0 \times 3}$ to extract a multi-resolution hierarchical feature maps $F_i \in \mathbb{R}^{X_i \times D}$, where $i \in \{0, 1, 2, 3, 4\}$. Following [28, 35], we set zero-initialized non-parametric instance queries $Q \in \mathbb{R}^{N_q \times D}$, referring to point positions sampled with *furthest point sampling* (FPS) [30]. Given the F_i and Q, the mask transformer decoder layer iteratively enhances the queries using the Mask Module (*MM*) and Query Refinement (*QR*) blocks. In the *MM*, as shown in Fig. 3 (a), we classify the category c_i for $i = \{1, 2, \ldots, N_c\}$ of each query via linear classification head f_{class} using following cross-entropy loss:

$$\mathcal{L}_{cls} = -\mathbb{E}_{c, w_c \sim \mathbb{D}} \left[\sum_{r \in N_c} w_c[r] \log f_{\text{class}}(Q)[r] \right] \quad (1)$$

where w_c denotes one-hot encoded category labels and $\mathbb{D}$ represents (input) data distribution. Also, the Q are fed through the query head f_{query} to project them to the same feature space as F_0 for mask prediction. Afterwards, we

Algorithm 1: Overview of our framework EASE.

Input: Colored point cloud $P \in \mathbb{R}^{N_p \times 6}$.
Result: Binary 3DIS mask $\hat{M}_f \in \{0,1\}^{X \times N_q}$.
Procedure:

```
# Sparse Convolutional Backbone.
Extract multi-resolution feature maps F_i from P.
# Edge Prediction Module (EPM).
Predict edge map Ê using full-resolution feature F_0.
Supervise Ê with pseudo edge label E.
Encourage the network to learn edge-aware features.
# Mask Transformer Decoder.
for l ∈ N_l do
    for i ∈ {1, 2, 3, 4} do
        Update queries Q using M_attn and F_i.
        # Semantic Network.
        Match the set of classes with text labels.
        Encode text into text embeds T using CLIP.
        Reinforce Q to contain semantic clues of T.
    end
end
```

compute the similarity between the projected query features Q' and F_0 using the dot product operation, then calculate the probability of the instance mask employing the sigmoid function as follows.

$$M_{attn} = \{m_{i,j} = [\sigma(F_0 \cdot f_{\text{query}}(Q)^T)_{i,j} > 0.5]\} \quad (2)$$

where the threshold value is 0.5 for binary mask. Further, to refine the query representation, we leverage the *QR* block, including masked cross-attention and self-attention mechanisms. Here, we utilize M_{attn} as the foreground mask for the masked cross-attention layer, where Q iteratively attend to each multi-scale feature (F_1-F_4, $i \geqslant 1$) as follows:

$$Q = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{D} + M_{attn})\mathbf{V} \quad (3)$$

where $\mathbf{K}$ and $\mathbf{V}$ are linearly projected from each point feature, and $\mathbf{Q}$ are from Q. Hence, the transformer focuses more on valuable instances instead of the unessential background. Subsequently, in the self-attention layer, the keys, values, and queries are all linear projections of Q from the cross-attention layer. Finally, the transformer decoder layer is recurrently applied for multiple iterations N_l and feature scales i, ultimately producing the final set of refined queries.

3.2. Intelligent Semantic Prior

It is important to highlight that precise geometric properties (*e.g.*, locations, distances and orientations) from 3D points significantly enhance 3DIS accuracy. While previous studies [22, 35, 36] benefit from the generous geometric potential of physical information, they are still limited in perceiving intricate real-world 3D instances. Hence, we propose to utilize semantic knowledge of individual instances as smart priors, effectively addressing the challenges.

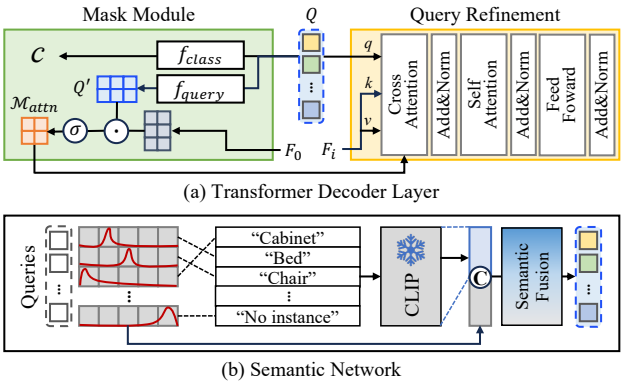


Figure 3. Detailed architecture of (a) Transformer Decoder Layer: consists of Mask Module and Query Refinement blocks. And (b) Semantic Network: we first map the categories of queries Q with the corresponding text description for text embeddings T from CLIP [32]. Then, we explicitly strengthen the Q with T using the Semantic Fusion network to learn semantic details.

Deep Semantic Knowledge for 3D Instances. Recently, visual language models [17, 32] have highlighted the benefits of their capability to provide general embedded knowledge from large-scale image-text multimodal data. In this work, we explicitly instruct the transformer-based 3DIS network to learn contextual variations among instances by utilizing the power of the prevalent language model CLIP [32]. To achieve this goal, we utilize a carefully designed Semantic Network, as shown in Fig. 3 (b). Specifically, with the refined query Q from each *QR* block, we first categorize each query following the linear classification head f_{class} in the *MM*. Given the set of category candidates, we map the category number c_i for $i = \{1, 2, \ldots, N_c\}$ with the corresponding text descriptions (*e.g.*, *"Cabinet"*, *"Bookshelf"*). Next, these text labels go through the large-scale pre-trained language model, which outputs text embedding vectors T. We then concatenate Q with T and strengthen the basic instance queries with deep semantic clues from text embeddings using the Semantic Fusion network as follows:

$$Q = \text{Sigmoid}(\mathbf{W}_2 \cdot \phi(\mathbf{W}_1 \cdot (Q \oplus T))^T) \quad (4)$$

where $\oplus$ indicates channel-wise concatenation, $\mathbf{W}_i$ denotes the learnable parameters of the i-th linear layer, and ϕ is the non-linear activation function. We apply this process for each iteration of the *QR*. Ultimately, we consider these explicitly guided queries with contextual variations as discriminative semantic priors for the practical perception of complex 3D instances, enhancing 3DIS accuracy.

3.3. Edge-Aware 3DIS Framework

In real-world 3D scenarios, instances are commonly positioned in diverse arrangements without following standard rules or patterns. Especially when instances are closely ad-

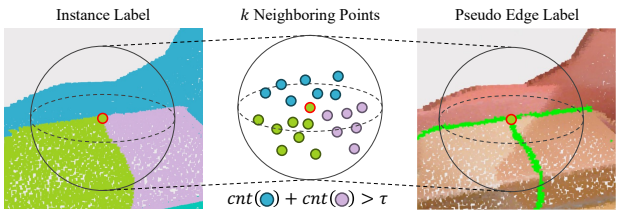


Figure 4. Pseudo edge label calculation using KNN algorithm. We collect k neighboring points of each point (red border) and compare their instance labels. If the count of different label points (blue and purple) among its surroundings surpasses a predefined threshold τ, we annotate the central point as an edge point (green).

jacent or overlapping, it becomes increasingly challenging for the model to define the edges precisely. Therefore, even with significant progress in technical strategies, most conventional methods [22, 35, 36] often misinterpret the spatial extent of each instance and ambiguously estimate their edges as illustrated in Fig. 1. To this end, we introduce an edge-aware 3DIS framework with an edge prediction module, enabling the network to utilize edge-advanced features.

Pseudo Edge Label Calculation. To optimize the edge prediction module, we first compute point-wise pseudo edge labels for all 3D point scenes. We operate a traditional k-Nearest Neighbors (KNN) classification algorithm based on the KD-tree data structure to explore the k nearest points. Here, we utilize point-wise instance labels from datasets. For all points in a 3D scene, we set each point as a central point and compare their instance labels with those of k neighbor points. As shown in Fig. 4, if the count of distinct label points among its surroundings exceeds a predefined threshold τ, we identify the central point as a boundary point. We finally generate pseudo-binary edge maps $E' \in \{0,1\}^{N_P \times 1}$, where points on the edges are 1 and internal points are 0 for the entire set of points P. Then, we voxelize the point-wise edge labels into the voxel-wise edge labels $E \in \{0,1\}^{X_0 \times 1}$ to supervise the edge extraction head. Note that we precompute the pseudo edge labels for every 3D scene in the datasets before training.

Learning Edge Cues. With dynamically generated pseudo edge labels E, we aim to enrich regular features to recognize the accurate spatial scope of 3D instances with precise boundaries. Specifically, we introduce an Edge Prediction Module, which predicts a voxel-wise dense edge map $\hat{E} \in (0,1)^{X_0 \times 1}$ to interactively guide the network in capturing edges of various instances. Given a full-resolution feature map F_0 from the backbone network, our edge extraction head, which consists of multiple shared MLPs, estimates the probabilities of edges for each voxel as follows:

$$\hat{E} = \text{MLP}_{edge}(F_0) = \text{Sigmoid}(\mathbf{W}_2 \cdot \phi(\mathbf{W}_1 \cdot F_0)^T) \quad (5)$$

where $\mathbf{W}_i$ denotes the learnable parameters of the i-th linear layer, and ϕ indicates the non-linear activation function. Following [6, 10], as edge voxels constitute a small portion of the entire voxel set, we formulate weighted binary cross entropy loss to train the edge extraction head as follows:

$$\mathcal{L}_{edge} = -\sum_{i=1}^{X_0}(w \cdot E_i \log \hat{E}_i + (1 - E_i)\log(1 - \hat{E}_i)) \quad (6)$$

where w represent the weight value used to balance the substantial difference between the numbers of edge points and others. Then, we propagate the edge cues of prediction as auxiliary information for the backbone network, improving 3D instance understanding with edge-aware features. Furthermore, we train our edge network jointly in an end-to-end manner with low computational costs.

Loss Function We finally predict instance mask $\hat{M}_f$ using the final set of queries in the *MM*. For training, we compute the following loss $\mathcal{L}_{mask}$ as a sum of the two losses:

$$\mathcal{L}_{mask} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}}(M_f, \hat{M}_f) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(M_f, \hat{M}_f) \quad (7)$$

where M_f denotes the ground-truth instance mask, and $\mathcal{L}_{dice}$ represents the Dice loss [5]. Ultimately, our model is trained end-to-end by minimizing the following loss $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \lambda_{mask}\mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{edge}\mathcal{L}_{edge} \quad (8)$$

where each λ is a hyperparameter derived from grid searches to handle the strength of respective loss term.

4. Experiments

4.1. Experimental Setup

Datasets. In this study, we train and evaluate the overall performance using four landmark datasets for 3D instance segmentation: ScanNetV2 [4], ScanNet200 [33], S3DIS [1], and STPLS3D [2]. These four datasets provide 3D point cloud scan data collected in various scenarios. Note that detailed descriptions of each dataset and all implementation details are provided in the supplemental material.

Evaluation Metrics. To evaluate the 3D instance segmentation accuracy, we operate Average Precision (AP), a typical metric for various computer vision tasks. We report mean average precision (mAP), which represents an average score with IoU thresholds ranging from 50% to 95% with a step size of 5%. Also, we provide mAP_{50} and mAP_{25}, indicating the scores with IoU thresholds of 50% and 25%, respectively. Further, we estimate mean precision (mPrec) and mean recall (mRec) on the S3DIS [1] dataset.

4.2. Performance Comparison with SOTA Methods.

ScanNetV2. We quantitatively compare our proposed network EASE with existing state-of-the-art methods. These methods demonstrate significant capabilities, but they still rely on geometric information from 3D points and often

Method	ScanNet Val		ScanNet Test	
	mAP	mAP_{50}	mAP	mAP_{50}
GSPN [44]	19.3	37.8	–	30.6
3D-SIS [16]	–	18.7	16.1	38.2
MASC [25]	–	–	25.4	44.7
3D-Bonet [43]	–	–	25.3	48.8
MTML [20]	20.3	40.2	28.2	54.9
3D-MPA [8]	35.5	59.1	35.5	61.1
DyCo3D [15]	35.4	57.6	39.5	64.1
PointGroup [18]	34.8	56.7	40.7	63.6
MaskGroup [45]	42.0	63.3	43.4	66.4
OccuSeg [13]	44.2	60.7	48.6	67.2
SSTNet [24]	49.4	64.3	50.6	69.8
HAIS [3]	43.5	64.1	45.7	69.9
SoftGroup [39]	46.0	67.6	50.4	76.1
Mask3D [35]	55.2	73.7	56.6	78.0
QueryFormer [26]	56.5	74.2	58.3	78.7
MAFT [22]	59.9	76.5	59.6	78.6
EASE (Ours)	**60.2**	**77.2**	**59.8**	**78.9**

Table 1. Comparison of 3D Instance Segmentation performance with state-of-the-art approaches on the ScanNetV2 [4] dataset. We evaluate mean average precision (mAP) with different IoU thresholds over 18 classes on the ScanNetV2 validation / hidden test set.

face challenges in effectively comprehending 3D instances. To address these limitations, we present a novel edge-aware framework featuring a semantic guidance network. As reported in Tab. 1, EASE generally outperforms other methods, achieving new state-of-the-art accuracy in terms of mAP and mAP_{50} on both validation (60.2 / 77.2) and hidden test (59.8 / 78.9) sets. These results confirm that our edge cues and smart semantic prior benefits the entire network, leading to high precision in 3D instance segmentation.

S3DIS. In Tab. 2, we evaluate 3DIS performance on Area 5 and 6-fold cross-validation of the S3DIS [1] dataset. For Area 5 evaluation, we employ data from Area 5 for validation and the other areas for training. Also, for 6-fold cross-validation, we assess validation scores across six other areas and compute the average. In both evaluations, EASE demonstrates significant performance improvements up to +2.4 / 2.5 (Area 5) and +0.8 / 0.9 (6-fold) for mAP / mAP_{50}.

ScanNet200. We also demonstrate considerable performance on the ScanNet200 validation set (see Tab. 3). Our method EASE precisely segments 3D instances for various categories compared to current state-of-the-art approaches.

STPLS3D. Remarkably, our EASE network also achieves higher scores on the STPLS3D [2] dataset, which consists of outdoor 3D point cloud scenes. EASE exceeds the second-best results with +1.1 / 1.6 / 1.3 margins in mAP / mAP_{50} / mAP_{25}. These results highlight the effectiveness of our method in understanding diverse real-world 3D instances, both from indoor and outdoor environments.

Method	S3DIS Area 5				S3DIS 6-fold CV			
	AP	AP_{50}	$Prec_{50}$	Rec_{50}	AP	AP_{50}	$Prec_{50}$	Rec_{50}
ASIS [41]	–	–	55.3	42.4	–	–	63.6	47.5
3D-Bonet [43]	–	–	57.5	40.2	–	–	65.6	47.6
3D-MPA [8]	–	–	63.1	58.0	–	–	66.7	64.1
PointGroup [18]	–	57.8	61.9	62.1	–	64.0	69.6	69.2
DyCo3D [15]	–	–	64.3	64.2	–	–	–	–
MaskGroup [45]	–	65.0	62.9	64.7	–	69.9	66.6	69.6
SSTNet [24]	42.7	59.3	65.5	64.2	54.1	67.8	73.5	73.4
SoftGroup [39]	51.6	66.1	**73.6**	66.6	54.4	68.9	**75.3**	69.8
Mask3D [35]	56.6	68.4	68.7	66.3	64.5	75.5	72.8	74.5
MAFT [22]	–	69.1	–	–	–	–	–	–
QueryFormer [26]	57.7	69.9	70.5	**72.2**	62.0	73.3	72.7	73.4
EASE (Ours)	**59.0**	**71.6**	69.4	68.7	**65.3**	**76.4**	73.6	**74.6**

Table 2. Comparison of 3D Instance Segmentation performance with state-of-the-art approaches on the S3DIS [1] Area 5 and 6-fold cross validation set. We evaluate mAP across different IoU thresholds, along with mean precision (mPrec) and mean recall (mRec) at a 50% IoU threshold over 13 classes of the S3DIS.

Method	mAP	mAP_{50}	mAP_{25}
SPFormer [36]	25.2	33.8	39.6
Mask3D [35]	27.4	37.0	42.3
QueryFormer [26]	28.1	37.1	43.4
MAFT [22]	29.2	38.2	43.3
EASE (Ours)	**29.9**	**38.8**	**44.7**

Table 3. Comparison of 3D Instance Segmentation performance with state-of-the-art approaches on the ScanNet200 [33] val set.

Method	mAP	mAP_{50}	mAP_{25}
PointGroup [18]	23.3	38.5	48.6
HAIS [3]	35.1	46.7	52.8
SoftGroup [39]	47.3	63.1	71.4
Mask3D [35]	63.4	79.2	85.6
EASE (Ours)	**64.5**	**80.8**	**86.9**

Table 4. Comparison of 3D Instance Segmentation performance with state-of-the-art approaches on the STPLS3D [2] test dataset.

4.3. Ablation Studies

Effect of Semantic and Edge Modules. In Tab. 5, we evaluate the variants of our method w/ and w/o the Semantic Network (SN) and Edge Prediction Module (EPM). The addition of two modules improves 3DIS accuracy across all experiments. The SN especially underscores the significance of incorporating rich semantic knowledge for effective segmentation, achieving +1.3 / 0.7 and + 1.5 / 1.7 improvements on mAP / mAP_{50} for ScanNetV2 and S3DIS, respectively. EPM further enhances performance via the auxiliary edge information for edge-aware features, resulting in gains of +0.9 / 0.8 and +1.9 / 0.9. Notably, integrating

Method			ScanNet Val	S3DIS Area 5
Baseline	Semantic	Edge	mAP / mAP$_{50}$	mAP / mAP$_{50}$
✓	-	-	58.4 / 75.9	56.6 / 68.4
✓	✓	-	59.7 / 76.6	58.1 / 70.1
✓	-	✓	59.3 / 76.7	58.5 / 69.3
✓	✓	✓	**60.2 / 77.2**	**59.0 / 71.6**

Table 5. Ablation study to see the effect of our proposed two main modules: Semantic Network and Edge Prediction Module.

Method			ScanNet Val	S3DIS Area 5
Usage	E-wise Sum	Concat	mAP / mAP$_{50}$	mAP / mAP$_{50}$
-	-	-	58.4 / 75.9	56.6 / 68.4
f_{cls}	✓	-	58.9 / 76.0	57.8 / 68.5
f_{cls}	-	✓	59.3 / 76.5	58.3 / 70.1
$S{\cdot}F$	✓	-	59.6 / 76.9	58.6 / 69.6
$S{\cdot}F$	-	✓	**60.2 / 77.2**	**59.0 / 71.6**

Table 6. Ablation study of leveraging text embeddings to provide semantic details. f_{cls} represents the classification head in the Mask Module, and $S{\cdot}F$ stands for the Semantic Fusion network.

SN and EPM leads to considerable advances (+1.8 / 1.3 and +2.4 / 3.2), confirming the effectiveness of each module.

Applications of Text Embeddings. In Tab. 6, we analyze various applications of text embeddings to determine the most effective strategy for the functional understanding of complex 3D instances. We utilize text embeddings for category classification from f_{cls} of the Mask Module to boost classification accuracy. Also, we strengthen basic queries using the Semantic Fusion ($S{\cdot}F$) network as described in Sec. 3.2. For experiments with the E-wise Sum, we map the text embeddings to the same feature space as the basic queries through MLPs and sum together. While incorporating semantic cues via element-wise sum enhances 3DIS accuracy, it exhibits limited improvements compared to concatenation. By comparison, applying the semantic knowledge across the entire transformer decoder through the $S{\cdot}F$ outperforms using them only for the classification head.

Weight Value of Edge Prediction Loss. To encourage the network to learn practical edge cues, we utilize an edge prediction network. In particular, we compute the weighted binary cross-entropy loss to effectively train the edge extraction head from sparse edge points. Here, we explore the impact of the coefficient value w of $\mathcal{L}_{edge}$ on 3DIS accuracy. First, as shown in Fig. 5, we empirically find that w influences the thickness of edges in a 3D scene. In Tab. 7, for the ScanNet [4] consisting of relatively small-scale scenes, a smaller value of w (*e.g.* 2), resulting in thin edges, leads to higher performance. However, for the vast-scale scenes of the S3DIS [1], a higher value of w (*e.g.* 10) leads to better performance with thicker edges. We finally observe the correlation between the scene scale and the weight value w; the optimal value for edge learning is proportional to the scale. Besides, methods using high-resolution feature maps (*e.g.* F_0) for edge prediction generally perform better than those using low-resolution features maps (*e.g.* F_4).

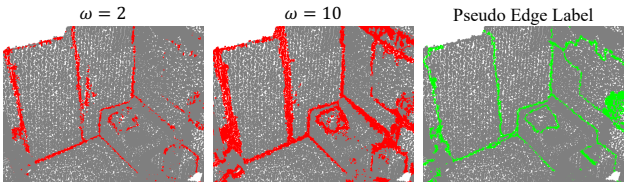


Figure 5. Visualization of edge prediction results (red) for different weight values w in the weighted binary cross-entropy loss $\mathcal{L}_{edge}$, along with corresponding pseudo-edge label (green).

Feature map	w in $\mathcal{L}_{edge}$	ScanNet Val	S3DIS Area 5
		mAP / mAP$_{50}$	mAP / mAP$_{50}$
F_0	2	**59.3 / 76.7**	56.9 / 68.6
F_4	2	59.0 / 76.1	56.7 / 68.5
F_0	6	58.9 / 76.3	58.4 / 68.9
F_4	6	58.5 / 76.0	58.2 / 68.6
F_0	10	58.4 / 75.8	**58.5 / 69.3**
F_4	10	58.5 / 75.7	58.2 / 69.0

Table 7. Ablation study to investigate the impact of varying weight values w in the weighted binary cross-entropy $\mathcal{L}_{edge}$, using high (F_0) and low (F_4) resolution feature maps for edge prediction.

4.4. Qualitative Analyses

Visual Comparison. In this section, we qualitatively confirm the usefulness of our novel framework EASE. We visualize the predicted semantic and instance masks of the baseline model [35] and ours on the ScanNetV2 [4] validation set, in Fig. 6. We emphasize the key differences using green and yellow boxes. As shown in the semantic results (Sem.), ours accurately classifies instances with similar apearances (*e.g.*, *desk* and *bed*, *door* and *cabinet*) instead of the baseline, which incorrectly recognizes them. Also, as illustrated in the instance mask results (Inst.) of Scene1 and Scene2, ours segments a single object as one, unlike the baseline, which splits it into multiple parts. Further, in Scene 3 (Inst.), ours correctly distinguishes multiple objects (*door* and *cabinet*) in contrast to the baseline model. These qualitative results demonstrate that our method provides more detailed and robust segmentation results than the baseline method.

Impact of Utilizing Semantic Priors. To verify the effectiveness of the Semantic Network, we present t-SNE [38] visualizations of query features for instances with similar

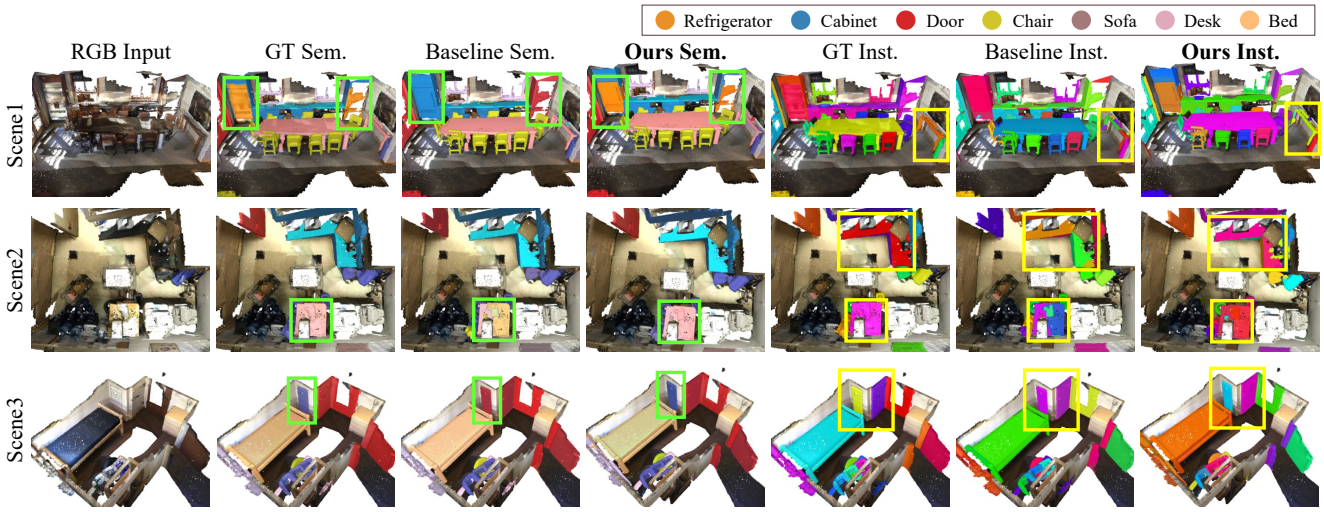


Figure 6. Qualitative comparison of 3D Instance Segmentation performance on the ScanNetV2 [4] validation set. We visualize semantic (Sem.) and instance (Inst.) masks of the baseline model and ours with Ground Truth (GT) masks. The critical differences are emphasized using green and yellow-colored boxes. Note that the color map (top right) represents semantic labels. Best viewed in color.

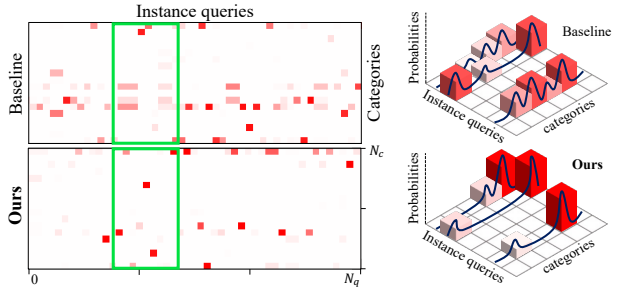


Figure 7. Category probability map (spatial distribution) for queries with high confidence. The x-axis denotes queries, the y-axis represents categories, and the z-axis (right) is probabilities.

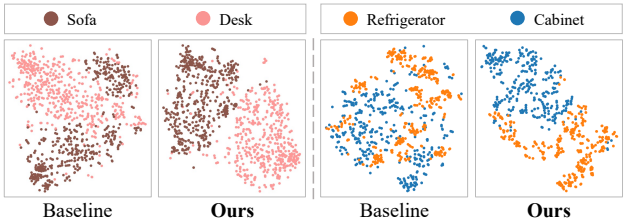


Figure 8. t-SNE [38] visualization of query features representing each instance. Compared to the baseline [22], which produces disorganized clusters, ours creates more distinct clusters.

appearances, which pose challenges for 3DIS models. As shown in Fig. 8, the baseline (MAFT [22]) model exhibits disorganized clusters, whereas ours shows relatively distinct feature space for challenging instances. Further, in Fig. 7, we visualize the spatial distribution of category probability for instance queries. The baseline tends to confuse per-query categories and predicts multiple categories for each query with low probability, causing semantic misclassifications. However, our model EASE overcomes this problem using semantic priors, providing clearer insight. These qualitative findings confirm that our semantic network effectively prompts the network to learn instance-specific semantic knowledge from intelligent text embedding priors.

5. Conclusion

Current 3D instance segmentation approaches often face challenges in understanding real-world 3D instances relying solely on geometrical information. Specifically, (i) these methods suffer from identifying instances with similar appearances, and (ii) they often misinterpret the spatial extent of instances, resulting in unclear edges. To tackle these challenges, EASE, our novel framework, focuses on context details of text embeddings from the language model as smart priors, enhancing practical semantic understanding. Also, we employ the edge prediction module, guiding the network to reduce misclassifications near edges with edge-aware features. Our extensive experiments verify the effectiveness of EASE, achieving new state-of-the-art scores.

Acknowledgments This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023. (RS-2023-00227409, Development of sketch-based semantic 3D modeling technology for creating user-centric Metaverse content spaces for indoor spaces, 80%), (R2022020068, 4D Content Generation and Copyright Protection with Artificial Intelligence, 10%), Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. (24ZC1200, Research on hyper-realistic interaction technology for five senses and emotional experience, 9%), and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00079, Artificial Intelligence Graduate School Program (Korea University), 1%).

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 2, 5, 6, 7

[2] Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*, 2022. 2, 5, 6

[3] 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. 1, 2, 6

[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. 2, 5, 6, 7, 8

[5] Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In *Proceedings of the European conference on computer vision (ECCV)*, pages 562–578, 2018. 5

[6] Shenglan Du, Nail Ibrahimli, Jantien Stoter, Julian Kooij, and Liangliang Nan. Push-the-boundary: Boundary-aware feature propagation for semantic segmentation of 3d point clouds. In *2022 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2022. 5

[7] Cathrin Elich, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. 3d bird's-eye-view instance segmentation. In *German Conference on Pattern Recognition*, pages 48–61. Springer, 2019. 1, 2

[8] 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. 1, 2, 6

[9] 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

[10] Jingyu Gong, Jiachen Xu, Xin Tan, Jie Zhou, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Boundary-aware geometric encoding for semantic segmentation of point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1424–1432, 2021. 5

[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. 2

[12] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *6th Annual Conference on Robot Learning*, 2022. 2

[13] 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. 6

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[15] 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. 6

[16] 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. 1, 2, 6

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2, 4

[18] 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. 1, 2, 6

[19] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 518–535. Springer, 2020. 2

[20] 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. 6

[21] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8500–8509, 2022. 2

[22] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3693–3703, 2023. 1, 2, 3, 4, 5, 6, 8

[23] 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. 2

[24] 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. 6

[25] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity

with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478*, 2019. 6

[26] Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Query refinement transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18516–18526, 2023. 1, 6

[27] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. 2

[28] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 3

[29] 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. 2

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[31] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1

[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. 2, 4

[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. 2, 3, 5, 6

[34] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 2

[35] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 1, 2, 3, 4, 5, 6, 7

[36] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2393–2401, 2023. 1, 2, 4, 5, 6

[37] 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

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7, 8

[39] 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. 6

[40] 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. 1, 2

[41] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2019. 6

[42] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *European Conference on Computer Vision*, pages 235–252. Springer, 2022. 2

[43] 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. 1, 2, 6

[44] 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. 1, 2, 6

[45] Min Zhong, Xinghao Chen, Xiaokang Chen, Gang Zeng, and Yunhe Wang. Maskgroup: Hierarchical point grouping and masking for 3d instance segmentation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 6