LR-AdaInSeg: Adaptive Instance Segmentation of Incomplete 3D Scenes Driven by Low-Rank Networks

Qin Zhang[1*], **Kun Zhou**[2*], **Xulun Ye**[1†]

[1]Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China
[2]School of Architecture & Urban Planning, Shenzhen University, Shenzhen, China
2411100314@nbu.edu.cn, zhoukun@szu.edu.cn, yexulun@nbu.edu.cn

Abstract

3D full-scene segmentation technology has demonstrated great potential driven by large models, but it often faces challenges of incomplete scenes and identification of invisible classes in practical applications. To address this, we propose the **LR-AdaInSeg** method, which significantly enhances the model's generalization ability in incomplete scenes through two key innovations: First, we design a Bayesian Low-Rank Module, which effectively solves the problem of feature space redundancy through dynamic optimization of the network structure, improving adaptability to incomplete scenes. Second, we combine graph contrastive clustering with the Low-Rank module, leveraging its robust feature representation capability to achieve accurate differentiation of invisible classes. In terms of implementation, we build a multi-scale feature extraction framework based on the 3D U-Net and utilize the 3D prompt points and their 2D masks as supervisory signals to achieve effective fusion of geometric and semantic information. Experiments show that our method achieves advanced performance on multiple benchmarks such as ScanNet, particularly excelling in handling incomplete scenes and invisible class objects.

Figure 1: Illustration of an incomplete scene with some objects not captured that leads to poor segmentation quality.

1. Introduction

3D full-scene point cloud segmentation plays a pivotal role in autonomous driving, machine vision, and related fields. Notably, breakthroughs like the Segment Anything Model (SAM)(Kirillov et al. 2023) have demonstrated remarkable zero-shot segmentation capabilities in 2D image processing. Building upon this success, recent advances—such as SAM3D (Yang et al. 2023), SAMPro3D (Xu et al. 2023b), and SAM2Object (Zhao et al. 2025)—have extended this zero-shot segmentation paradigm to 3D point clouds. Specifically, these methods achieve this by back-projecting 2D segmentation results into 3D space, thereby bridging the gap between 2D and 3D vision tasks. However, such methods are highly dependent on the completeness of the camera viewpoints. For instance, if furniture in a corner of a room remains uncaptured from any angle due to occlusions or perspective limitations (i.e., it is in a partially observed state), as illustrated in Fig. 1, the corresponding 2D image information will be absent. This absence creates inherent gaps in the 3D segmentation masks generated via back-projection, rendering it impossible to effectively identify and segment these unobserved objects or their constituent parts. Therefore, a critical challenge remains: how to achieve accurate and robust zero-shot segmentation of incomplete 3D scenes, particularly those characterized by severe occlusions, limited perspectives, and the potential presence of unknown object categories.

To address the aforementioned challenges, this paper proposes LR-AdaInSeg, an adaptive instance segmentation framework for challenging zero-shot and incomplete 3D scenes. Our framework employs 2D segmentation mask projection to achieve zero-shot segmentation and further introduces the Contrastive Graph Class Similarity (CGCS) module, specifically designed to tackle the segmentation of unknown categories (i.e., categories not seen during training) in incomplete scenes. CGCS leverages semantic similarity among 3D point features to construct graph structures and drive contrastive clustering learning, thereby effectively mining and aggregating potential unknown cate-

*Co-first author.
†Corresponding author.

gory features. To extract more discriminative geometric features, we design a backbone network based on a 3D U-Net architecture. Furthermore, to address the common issue of redundancy in high-dimensional feature spaces, we propose a Bayesian Low-Rank Module. This module enhances the compactness of similar feature representations by imposing structured Low-Rank constraints on the feature matrix and incorporating an uncertainty-aware dynamic weight adjustment mechanism. As a result, it improves the model's robustness and generalization ability in extracting features of unknown categories. Extensive experiments demonstrate that our method achieves robust generalization in 3D instance segmentation tasks involving unseen categories and partially observed scenes, and it has achieved superior performance on datasets such as ScanNet.

Our contributions are summarized as follows:

- We propose a Bayesian Low-Rank Module that automatically identifies important features in the feature space, reduces it to a Low-Rank representation, and extracts salient features even for unseen classes, thereby enhancing the robustness and generalization of the model.
- We integrate supervised and unsupervised learning to optimize a novel Contrastive Graph Class Similarity (CGCS) loss, which jointly accounts for the segmentation loss of both visible and invisible classes.
- Extensive experiments demonstrate that our method achieves improved segmentation performance under both complete and incomplete scene views.

2. Related Work

Open-set 3D scene segmentation: Most research on 3D scene understanding traditionally relies on fully supervised and semi-supervised methods (Graham, Engelcke, and Van Der Maaten 2018; Hackel et al. 2017; Çiçek et al. 2016; Song and Xiao 2016; Wang et al. 2019). Fully supervised methods require large-scale manual annotations, while (Chen, Nießner, and Dai 2022; Chibane et al. 2022; Xu et al. 2023a; Xie et al. 2020; Chen et al. 2020; Huang et al. 2023; Kohli, Sitzmann, and Wetzstein 2020) semi-supervised ones depend on limited labels to guide unlabeled segmentation. Both approaches fundamentally rely on labeled data.

Recently, zero-shot 3D scene understanding methods still rely on supervised pretraining with predefined datasets. In contrast, 2D visual foundation models have demonstrated strong zero-shot recognition, inspiring their extension to 3D. Yet these approaches still require model adaptation (Rozenberszki, Litany, and Dai 2024), 3D-2D distillation (Chen et al. 2023; Ding et al. 2023, 2024; Liu et al. 2023; Peng et al. 2023), or pre-trained region proposal networks (Huang et al. 2024; Lu et al. 2023; Nguyen et al. 2024; Schult et al. 2023). Our method instead leverages SAM's inherent zero-shot ability, eliminating the need for labeled data or retraining, and directly performing open-world 3D scene segmentation.

Incomplete scene segmentation: Despite the continuous progress in deep learning-based 3D scene understanding, existing methods still encounter significant challenges in scenarios with missing or incomplete input. (Garbade et al. 2019) approaches that integrate RGB semantic flow and deep geometric flow offer improved occlusion handling, yet their heavy dependence on 2D segmentation introduces inherent limitations: semantic labels from the RGB stream are entirely lost in occluded regions (e.g., objects behind cabinet doors), and the early-stage feature concatenation strategy struggles to adaptively fuse multimodal information, resulting in semantic blind spots within the 3D space. Recent advancements further reveal that although (Fan et al. 2021), a distance image-based method, alleviates missing data via KNN interpolation, it faces two core limitations: (1) its closed-set semantic design restricts generalization to novel object categories; and (2) local interpolation in the projection space fails to reconstruct large-scale occluded structures. Simultaneously, (Chen, Gong, and Röning 2024) enhances feature association through instance-level knowledge aggregation, but still suffers from topological distortions (e.g., fragmentation of thin-walled objects) caused by point cloud noise in severely occluded scenes, and lacks explicit modeling of component-level geometric relationships.

However, there is no good method for 3D segmentation of incomplete scenes. To solve this problem, we propose an adaptive network LR-AdaInSeg based on 3D U-Net and identify invisible classes for segmentation by comparing image class similarities. When the input scene is sparse, our method LR-AdaInSeg can still effectively segment the 3D scene, with good adaptability and robustness.

3. Method

To address the challenge of segmentation in incomplete scenes in 3D point clouds, we propose a method composed of four modules: 3D Prompt Proposal, 3D U-Net Network, Bayesian Low-Rank Module, and Detection of Invisible Classes. Due to the difficulty of accurate 3D segmentation, the 3D Prompt Proposal module projects the point cloud onto 2D images, applies SAM for segmentation, and maps the masks back to 3D, providing strong supervision for visible classes. Since invisible classes lack corresponding 2D masks, the 3D U-Net Network is introduced to extract 3D features under the guidance of visible class supervision. Subsequently, the Detection of Invisible Classes module performs contrastive clustering to group features in a self-supervised manner, enabling invisible classes to emerge naturally in the feature space. Finally, to enhance the separability of invisible classes, the Bayesian Low-Rank Module mitigates feature redundancy through adaptive Low-Rank constraints. An overview of our framework is shown in Fig. 2.

3.1. 3D Prompt Proposal

3D-to-2D projection: To reduce the density of the point cloud while preserving as much spatial information as possible, this study employs the Farthest Point Sampling (FPS) algorithm to sparsify the original 3D point cloud. The method constructs a representative subset of points with maximal spatial coverage by iteratively selecting new samples that are farthest in spatial distance from the already selected set. This sampling strategy, based on geometric measures, effectively

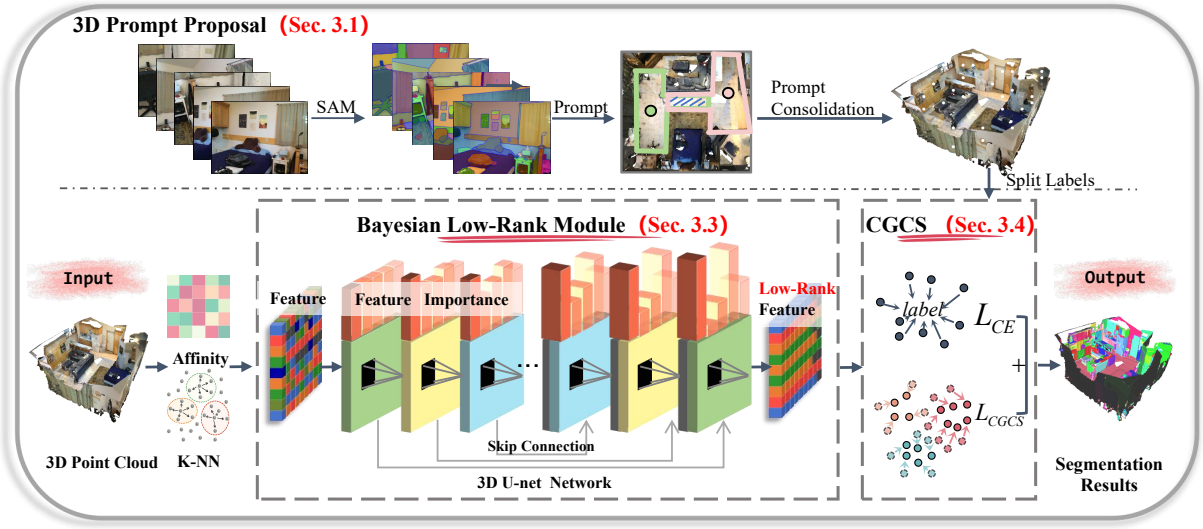

Figure 2: Method overview. We utilize prompt points and the corresponding SAM segmentation masks obtained from the 3D Prompt Proposal module to provide supervisory signals for our network. To address the segmentation of unknown classes in incomplete scenes, our method proceeds as follows: First, geometric features are extracted using a 3D U-Net backbone. Second, the CGCS module constructs a graph-structured contrastive clustering process to mine features of unknown classes. Finally, the Bayesian Low-Rank Module assigns a learnable parameter to each feature dimension, enabling dynamic control over Feature Importance. By discarding irrelevant or uninformative features, it effectively reduces feature redundancy within the network, thereby significantly enhancing the model's robustness and generalization ability in extracting features of unseen categories.

reduces the data size while maintaining the macro-geometric features and semantic information of the scene.

Following (Peng et al. 2023), the corresponding pixel projection x of a prompt point p is computed by the following equation: we consider pinhole camera configurations.

$$\tilde{x} = K \cdot [R \mid \mathrm{t}] \cdot \tilde{p}, \tag{1}$$

where $\tilde{p}$,$\tilde{x}$ represent the homogeneous coordinates of the pixel projection and the prompt point p, firstly, the external parameter matrix $[R \mid \mathrm{t}]$ is used to turn $\tilde{p}$ in the world coordinate system into the coordinates of the camera coordinate system, and then projected to the normalization plane, and then the internal parameter matrix K is mapped to the pixel coordinate system to obtain the homogeneous coordinate $\tilde{x}$ of the final pixel projection.

2D Image Segmentation with SAM: In our framework, we input all the pixel coordinates calculated previously to prompt SAM, thereby obtaining all corresponding 2D segmentation masks. Since the same prompt point projects to the same coordinates on different frames, it ensures inter-frame consistency. Subsequently, we use the View-Guided Prompt Selection and Surface-Based Prompt Consolidation modules (Xu et al. 2023b)to eliminate redundant 3D prompt points, ultimately obtaining the most concise 3D prompt points. Finally, we use the previous 3D-2D correspondence relationship to map high-confidence segmentation labels back to 3D point clouds for subsequent network training.

3.2. 3D U-Net Network

Our framework is built upon a 3D U-Net network, which is well-suited for 3D feature modeling and multi-scale information fusion in volumetric segmentation tasks. The use of $3 \times 3 \times 3$ convolutional kernels allows for effective extraction of local spatial features from voxelized point clouds, capturing geometric correlations between neighboring voxels—such as surface curvature and edge structures—and thus supports fine-grained segmentation. The encoder-decoder architecture enables the extraction of global semantic context through progressive downsampling, followed by gradual resolution recovery via deconvolution. This hierarchical design facilitates coarse-to-fine feature fusion and exhibits robustness in extracting representative features even in the presence of previously unseen objects, contributing to improved segmentation performance and preservation of structural details.

We train the 3D U-Net network using the previously obtained 3D prompt points along with their corresponding segmentation labels, and optimize the network parameters with a cross-entropy loss function.

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log \left(\hat{y}_i \right) + \left(1 - y_i \right) \log \left(1 - \hat{y}_i \right) \right), \tag{2}$$

where N denotes the total number of training points, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability for each training point after network inference.

3.3. Bayesian Low-Rank Module

In high-dimensional feature spaces, redundant information and noise can weaken the model's generalization ability—particularly when processing unseen objects or regions with incomplete details. To mitigate this issue, we introduce a Bayesian Low-Rank Module into the feature

representation layer of the network. This module aims to suppress noise and redundancy while maximizing the model's ability to discriminate different feature classes. Assuming the total number of classes is a, we posit that, after removing redundant information, the rank of features belonging to the same class should be 1. Therefore, the overall rank of all features is a. Our optimization objective is as follows:

$$\mathrm{Rank}(X) = a, \tag{3}$$

where X represents the features of the hidden layer obtained from the neural network.

Based on this, the Low-Rank regularization objective can be formulated as:

$$\min \mathrm{Rank}(X') + \lambda \|X' - X\|_2^2, \tag{4}$$

where $X^{'}$ is the expected Low-Rank approximation for X and λ is a scaling factor.

This problem can be formulated as a rank minimization task, where the goal is to find a Low-Rank approximation of the matrix X' with rank at most k. Since rank minimization is NP-hard, it is typically addressed via nuclear norm regularization and solved using Singular Value Decomposition (SVD), which decomposes X' into components ordered by significance. However, in practical scenarios, two critical challenges arise: (1) for large-scale data, the number of semantic categories k within each mini-batch is unknown and dynamically varies; and (2) SVD is inherently nondifferentiable, which complicates its integration into end-to-end trainable systems. To address the aforementioned two issues, we introduce the Bayesian Low-Rank Module. To transform the non-differentiable SVD method into a differentiable approach, we introduce auxiliary functions that yield the same solutions as SVD but are themselves differentiable. Here we primarily employ:

$$\min_{\mathbf{X}} \mathrm{trace}(X^\top W X) \quad \text{s.t. } X^\top X = I, \tag{5}$$

This further introduce the output of a single-layer network with orthogonal layer used in (Shaham et al. 2018) as:

$$X^{'} = XW, \tag{6}$$

where $X \in \mathbb{R}^{n \times d}$ and $W \in \mathbb{R}^{d \times m}$. In order to automatically select the important feature dimensions, we first introduce independent hyperparameters α_i for each input feature dimension (i.e., each column of W), so that the weight matrix satisfies the prior distribution:

$$p(W|\alpha) = \prod_{i=1}^{d} \mathcal{N}(W_{i:}|0, \alpha_i^{-1} I_m), \tag{7}$$

where the initial α_i is equal, indicating that each dimension of the initial feature is equally important. After subsequent training and optimization, the smaller α_i means the more important the dimension of its constraint, and the larger α_i means the less important the dimension of its constraint.

According to Bayes' theorem, we need to maximize the edge likelihood, so we add an Automatic Relevance Determination(ARD)(Mackay 1992) regularization term to the loss function:

$$\mathcal{L}(W, \alpha) = \frac{1}{2\sigma^2} \|X^{'} - XW\|_F^2 + \frac{1}{2} \sum_{i=1}^{d} \alpha_i \|W_{i:}\|_2^2 + \mathrm{c}, \tag{8}$$

where the first term is the standard mean-square error loss function, σ^2 represents the noise variance, the second term is the ARD regularization term, and the third term is a constant term.

During training, we adopt an alternating optimization strategy to update W and α. Specifically, W is updated using the Adam optimizer, while α is updated based on the current value of W according to the following update rule:

$$\alpha_i = \frac{m}{\|W_{i:}\|_2^2 + \epsilon}, \tag{9}$$

Where ϵ is a small constant added for numerical stability, and m is the output dimension. After training, we obtain W and the corresponding importance of each α_i, and then we use the pruning threshold τ to set the less important dimensions to zero (i.e., the corresponding columns in X and rows in W) and remove them, ultimately obtaining two Low-Rank matrices X_{red} and W_{red}.

Therefore, we use the Low-Rank approximation $X^{'}$:

$$X^{'} \approx X_{\mathrm{red}} W_{\mathrm{red}}, \tag{10}$$

By gradually compressing the weights of redundant features during training, we finally obtained a Low-Rank approximation matrix. However, the reduction of redundant information in a single feature representation layer is not obvious, so we use Low-Rank approximation modules in each network depth, so that the feature representation layers can remain consistent.

3.4. Detection of Invisible Classes

Existing 3D segmentation methods typically rely on full-scene capture. However, not all objects can be captured in real-world scenarios, and existing methods often fail when some objects are missing. Thus, the current challenge lies in segmenting invisible classes in incomplete scenes. Previously, we employed the ARD method to reduce the rank of the feature matrix and capture the most informative features, which laid the foundation for the subsequent division of invisible classes. To discover new categories while maintaining the classification of known categories, we combine supervised classification methods with clustering methods. Here, the classification of known categories is achieved through supervised learning, while the discovery of new categories is accomplished using clustering techniques.

Challenges in joint supervised and unsupervised optimization: Supervised learning constructs a quantifiable loss function based on a clear input-output mapping, such as cross-entropy loss or mean squared error, which accurately reflects the deviation between model predictions and ground truth labels, and cooperates with gradient descent to form a closed-loop optimization process. In contrast, unsupervised learning lacks such direct feedback signals, and its objective functions are often designed around the intrinsic data structure, such as the reconstruction loss of an autoencoder or the

compactness measure of clustering algorithms. These objectives are typically multimodal, non-convex, and difficult to interpret and optimize. When attempting to integrate the two, the scale differences between heterogeneous loss functions can cause conflicts in gradient directions, and competition among different tasks can destabilize parameter updates. Therefore, the main challenge lies in simultaneously optimizing the network with both supervised and unsupervised penalty terms in high-dimensional space, where the unsupervised penalty must be compatible with the overall learning objective and enable the network to infer the distribution of unseen classes.

Contrastive Graph Class Similarity Loss: We solve this problem by implementing contrastive constraints between samples in the learned feature space. Specifically, we construct the class affinity matrix G, where if samples x_n and x_i belong to the same class, then $G = 0$, and if they belong to different classes, then $G = \eta(\eta > 0)$. Our loss function is defined as:

$$\Phi = \min_{\{s_n\}} \sum_{n=1}^{N} \| \|y_n - y_i\|_2^2 - G_{n,i}\|_2^2, \quad (11)$$

where y_n and y_i represent class designators (or embeddings) for x_n and x_i, respectively. If $G = 0$, then $y_n = y_i$; If $G = \eta$, then $y_n \neq y_i$.

Lemma 1. Given the affinity matrix G, if the set of demonstrators y_n satisfies the equation (11), then for any pair (n, i): $G_{n,i} = 0 \implies y_n = y_i, \quad G_{n,i} = \eta \implies y_n \neq y_i$.

Proof: Assume G = 0 (indicating identity), but we have $y_n = y_i$. Then the term

$$\left(\|y_n - y_i\|_2^2 - G_{n,i}\right)^2 = \left(\|g_n - g_i\|_2^2 - 0\right)^2,$$

would be strictly positive, contradictory assumptions that equation (11)has been minimally reached. If $G = \eta$ but $y_n = y_i$, then a similar argument applies. Therefore, in order to achieve the minimum possible sum, it must be satisfied $G_{n,i} = 0 \implies y_n = y_i, \quad G_{n,i} = \eta \implies y_n \neq y_i$.

In practice, the corresponding G is agnostic to the invisible class, so we use the KNN method to approximate the unknown class. At the same time, we introduce a regularized penalty term on the network output $f_\theta(x_n)$ to promote near-discrete embedding. Therefore, the improved CGCS loss is given by:

$$\mathcal{L}_{\mathrm{CGCS}} = \min_{\theta} \Bigg(\sum_{n=1}^{N} \sum_{i=1}^{N} \left\| \left\| f_\theta(x_n) - f_\theta(x_i) \right\|_2^2 - G_{n,i} \right\|_2^2 + \varphi \sum_{n=1}^{N} \|f_\theta(x_n)\|_1 \Bigg), \quad (12)$$

where θ represents the model parameters,and φ balances the sparsity term.

Model Mixed Loss: After introducing the unsupervised CGCS loss function, we combine it with the supervised cross-entropy loss of Eq.(2) to obtain the final loss function of the model:

$$\mathcal{L} = \delta \mathcal{L}_{\mathrm{CE}} + \mu \mathcal{L}_{\mathrm{CGCS}}, \quad (13)$$

where δ and μ are the corresponding weights of cross-entropy and CGCS terms. This hybrid loss function allows the model to maintain high segmentation accuracy on visible classes as well as effectiveness on invisible classes. As a result, our framework can achieve a comprehensive segmentation of the entire scene, even in the absence of scenes.

4. Experiments

4.1. Experimental Setting

Datasets: To verify the superiority of the proposed method, we have adopted three progressive datasets built based on the ScanNet framework to meet the needs of scene understanding at different levels: ScanNetv2 is the base version of this series, containing RGB-D sequences of 1,513 real indoor scenes, totaling approximately 2.5 million image frames. ScanNet200 expands the semantic category system on the basis of ScanNetv2, increasing the labeled objects from the base 21 categories to 200 categories, with a focus on enhancing coverage of long-tail distributions and fine-grained objects (such as chairs of different shapes, containers, etc.). ScanNet++ focuses on high-fidelity 3D reconstruction and refined annotation, constructing more challenging scenes like shelves and dense desks through high-precision mesh structures, optimized texture mapping, and dense instance annotation for small-scale objects and complex structures.

Evaluation indicators: AP (Average Precision) is a commonly used evaluation indicator in tasks such as instance detection and object detection. It is used to measure the overall performance of the model on all prediction results. It is obtained by calculating the area under the precision-recall curve, which reflects the balance between the prediction accuracy and completeness of the model under different confidence thresholds. Specifically, AP divides the recall rate from 0 to 1 into multiple intervals, calculates the maximum precision value in each interval, and then averages these precision values. The higher the AP value, the higher the model has while maintaining a high recall rate.

4.2. Complete Scene Experimental Comparison

Our method LR-AdaInSeg aims to solve the semantic segmentation problem in incomplete scenes. Nevertheless, we still conducted comparative experiments in complete scenes to verify the effectiveness and generalization ability of our method. The experiments were conducted on three datasets: ScanNetv2, ScanNet200 and ScanNet++. We compared with other methods, including: UnScene3D, Open3DIS, SAMgraph, SAM3D, SAI3D, Segment3D, SAM2Object, SAMPro3D. The experimental results are shown in Table 2.

Experimental results show that our proposed Bayesian Low-Rank mechanism provides a new idea for semantic decoupling of complex three-dimensional scenes by optimizing the distribution of structured features.

4.3. Incomplete Scene Experimental Comparison

In order to verify the segmentation performance of the model in the case of incomplete scenes, we conduct comparative experiments with other methods in incomplete scenar-

Method	Conference	ScanNetv2			ScanNet200			ScanNet++		
		AP	AP50	AP25	AP	AP50	AP25	AP	AP50	AP25
UnScene3D	CVPR_2024	12.1	25.0	50.9	-	-	-	-	-	-
Open3DIS	CVPR_2024	-	-	-	15.2	25.8	29.9	-	-	-
SAM-graph	ECCV_2024	10.5	29.6	53.3	19.8	36.4	60.2	8.9	22.0	40.2
SAM3D	ISBI_2024	13.9	28.3	49.7	7.2	13.3	18.7	6.5	11.6	27.7
SAI3D	CVPR_2024	20.3	43.8	66.7	10.7	15.1	20.4	13.0	25.7	46.6
Segment3D	ECCV_2024	-	-	-	**24.3**	33.7	-	9.8	18.4	35.5
SAM2Object	CVPR_2025	**25.1**	47.2	67.7	10.1	12.8	20.2	16.5	30.7	42.8
SAMPro3D	3DV_2025	19.2	40.9	64.3	20.5	43.8	63.7	17.4	32.2	48.9
LR-AdaInSeg	Ours	23.7	**47.5**	**68.4**	22.6	**45.8**	**66.5**	**18.0**	**33.5**	**50.0**

Table 1: This table presents the comparative experimental results on the ScanNetv2, ScanNet200, and ScanNet++ datasets under *incomplete scenes* (λ=0.7). The best results are highlighted in bold.

Method	Conference	ScanNetv2			ScanNet200			ScanNet++		
		AP	AP50	AP25	AP	AP50	AP25	AP	AP50	AP25
UnScene3D	CVPR_2024	15.9	32.2	58.5	-	-	-	-	-	-
Open3DIS	CVPR_2024	-	-	-	23.7	29.4	32.8	-	-	-
SAM-graph	ECCV_2024	15.1	33.3	59.1	22.1	41.7	62.8	12.9	25.3	43.6
SAM3D	ISBI_2024	20.2	34.0	53.3	9.8	15.2	20.7	7.2	14.2	29.4
SAI3D	CVPR_2024	30.8	50.5	70.6	12.7	18.8	24.1	17.1	31.1	49.5
Segment3D	ECCV_2024	-	-	-	**27.7**	39.8	-	12.0	22.7	37.8
SAM2Object	CVPR_2025	**34.0**	**52.7**	70.3	13.3	19.0	23.8	20.2	34.1	48.7
SAMPro3D	3DV_2025	24.3	45.7	67.7	26.3	47.2	68.6	**20.3**	35.6	53.2
LR-AdaInSeg	Ours	28.7	50.3	**70.9**	24.7	**48.6**	**68.9**	19.6	**35.7**	**53.6**

Table 2: This table presents the comparative experimental results on the ScanNetv2, ScanNet200, and ScanNet++ datasets under *complete scenes*. The best results are highlighted in bold.

BLRM	CGCS	ScanNetv2		ScanNet200		ScanNet++	
		AP($\lambda = 1$)	AP50($\lambda = 0.8$)	AP($\lambda = 1$)	AP50($\lambda = 0.8$)	AP($\lambda = 1$)	AP50($\lambda = 0.8$)
×	✓	45.4	43.2	42.8	40.9	31.6	27.4
✓	×	49.1	45.7	46.5	44.3	34.0	31.5
✓	✓	**50.3**	**48.6**	**48.6**	**47.5**	**35.7**	**34.1**

Table 3: This table verifies the effect of Bayesian Low-Rank Module(BLRM) and CGCS module on the performance improvement of 3D instance segmentation. Ablation experiments were performed in complete and incomplete scenes respectively.

ios (λ=0.7). The quantitative comparison results are shown in Table 1. As shown in the results, our proposed LR-AdaInSeg achieves higher accuracy in most scenarios. Notably, it outperforms SAM2Object and SAMPro3D by margins of up to **2.8** and **1.3**, respectively, on ScanNet++—a more demanding benchmark under incomplete-view conditions. This improvement highlights the effectiveness of LR-AdaInSeg in uncovering semantic similarities among unknown category features and aggregating fragmented features. This is also confirmed in the results in Fig.3, which visually shows that the model effectively solves the problem of unknown category segmentation in incomplete scenes through the collaboration of CGCS feature aggregation, 3D U-Net geometry extraction, and Bayesian Low-Rank redundancy processing. Even if some scenes are incomplete, it still maintains good accuracy and robustness.

4.4. Ablation Studies and Analysis

We conducted ablation experiments on the ScanNetv2, ScanNet200, and ScanNet++ datasets under both complete and incomplete scenarios to evaluate the effectiveness of the module designs in LR-AdaInSeg, including the 3D U-Net, Bayesian Low-Rank Module, and CGCS. Among these, the 3D U-Net module is indispensable, as it is necessary for learning features of invisible classes.

Efficacy of Bayesian Low-Rank Module: We study the contribution of the Bayesian Low-Rank Module by removing it from the entire framework (first row vs third row).

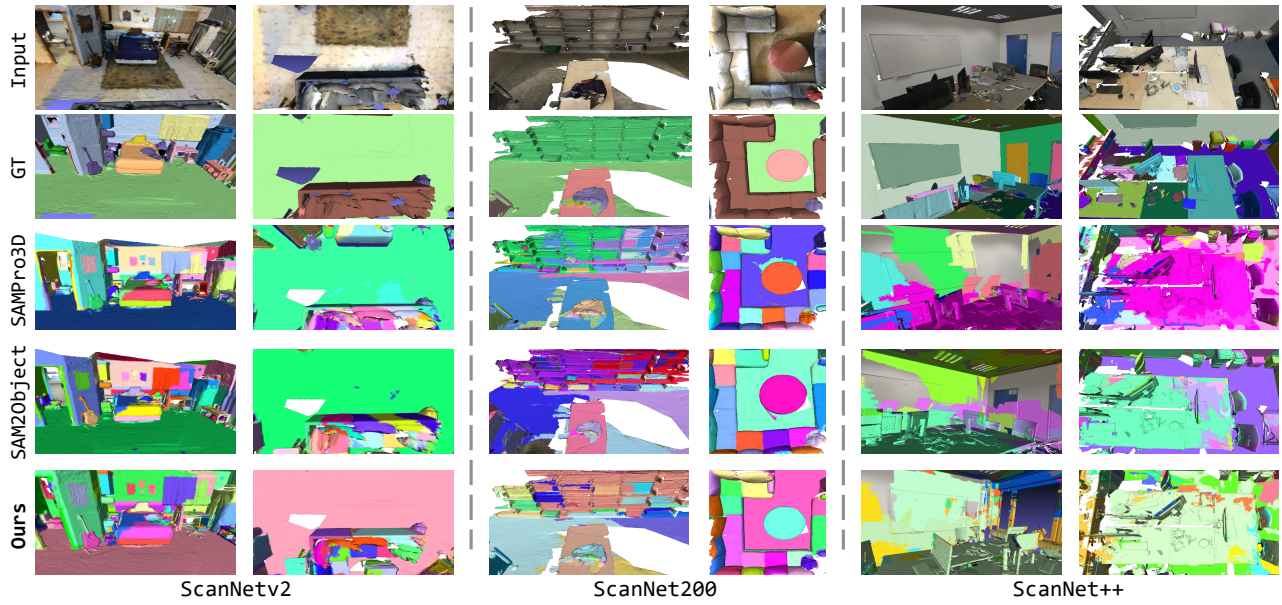

Figure 3: Comparison of 3D instance segmentation performance with scene completeness threshold λ=0.7 on ScanNetv2, ScanNet200, and ScanNet++ datasets. The figure shows the segmentation results of different methods in representative scenes. The bottom row shows the method proposed in this paper. As the main comparison baseline, the results of the SAMPro3D and SAM2Object method are also included in the figure.

As shown in Table 3, the absence of this module leads to a general drop in performance on all datasets and evaluation metrics. For example, on ScanNetv2, the performance of AP50@λ=0.8 drops from 48.6 to 43.2. This significant performance drop highlights the key role of the Bayesian Low-Rank Module in enhancing the model to handle Low-Rank approximations and improving the overall segmentation performance, especially for unseen categories.

Efficacy of CGCS module: We further evaluate the CGCS module by removing it from the full model (second row vs. third row). As shown in Table 3, removing the CGCS module has a negative impact on the performance of all datasets. For example, on ScanNetv2, the performance of AP50(λ=0.8) drops from 48.6 to 45.7. This shows that the CGCS module plays a vital role in effectively capturing global context information, which is essential for achieving robust segmentation performance in both complete and incomplete scenes. Ablation studies confirm that each module in LR-AdaInSeg makes a unique contribution to the effectiveness of the model, and their combination is the key to addressing the challenge of unseen category segmentation in sparse point cloud data.

4.5. Hyperparameter Analysis

λ and ϕ (threshold for intersection and parallel ratio)(Section 3.4): As shown in Fig.4, we explore the impact of scene completeness on segmentation performance (AP) metrics under different IoU. The results show that with the increase of scene integrity, the corresponding AP value also increases, but it can be seen that the model performs well under low scene integrity, especially benefiting from the effective mining of unknown class features by the CGCS module and the powerful geometric feature extraction ability of the 3D U-Net backbone network.

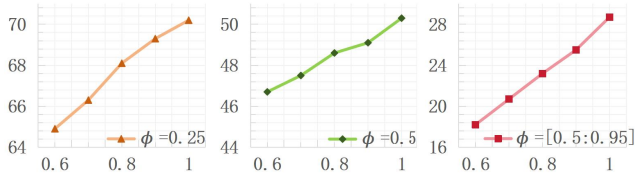

Figure 4: AP under IoU thresholds 0.25, 0.5, and [0.5:0.95] across scene completeness $\lambda \in [0.6, 1.0]$. Each plot shows AP (y-axis) versus λ (x-axis).

5. Conclusion

This paper introduces LR-AdaInSeg, a novel 3D instance segmentation framework designed to detect invisible classes in incomplete views without compromising segmentation accuracy. Our approach includes a Contrastive Graph Class Similarity (CGCS) module, which leverages semantic similarity among 3D point features to build graph structures and perform contrastive clustering, effectively revealing and aggregating potential unknown category features. Additionally, we propose a Bayesian Low-Rank module to efficiently reduce feature redundancy in high-dimensional space within the original 3D U-Net architecture. Extensive experiments demonstrate the effectiveness of LR-AdaInSeg.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62471266, 62006131,

62071260, and the Ningbo Major Research and Development Plan Project (No.2023Z225).

References

Chen, B.; Gong, C.; and Röning, J. 2024. Filling missing values matters for range image-based point cloud segmentation. *IEEE Transactions on Intelligent Vehicles*.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, Y.; Hu, V. T.; Gavves, E.; Mensink, T.; Mettes, P.; Yang, P.; and Snoek, C. G. 2020. Pointmixup: Augmentation for point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 330–345. Springer.

Chen, Y.; Nießner, M.; and Dai, A. 2022. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *European Conference on Computer Vision*, 543–560. Springer.

Chibane, J.; Engelmann, F.; Anh Tran, T.; and Pons-Moll, G. 2022. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *European conference on computer vision*, 681–699. Springer.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Ding, R.; Yang, J.; Xue, C.; Zhang, W.; Bai, S.; and Qi, X. 2023. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7010–7019.

Ding, R.; Yang, J.; Xue, C.; Zhang, W.; Bai, S.; and Qi, X. 2024. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fan, Z.; Liu, H.; He, J.; Zhang, M.; and Du, X. 2021. MPDNet: A 3D missing part detection network based on point cloud segmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1810–1814. IEEE.

Garbade, M.; Chen, Y.-T.; Sawatzky, J.; and Gall, J. 2019. Two stream 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9224–9232.

Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J. D.; Schindler, K.; and Pollefeys, M. 2017. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847*.

Huang, D.; Peng, S.; He, T.; Yang, H.; Zhou, X.; and Ouyang, W. 2023. Ponder: Point cloud pre-training via neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16089–16098.

Huang, Z.; Wu, X.; Chen, X.; Zhao, H.; Zhu, L.; and Lasenby, J. 2024. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision*, 169–185. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kohli, A. P. S.; Sitzmann, V.; and Wetzstein, G. 2020. Semantic implicit neural scene representations with semi-supervised training. In *2020 International Conference on 3D Vision (3DV)*, 423–433. IEEE.

Liu, Y.; Kong, L.; Cen, J.; Chen, R.; Zhang, W.; Pan, L.; Chen, K.; and Liu, Z. 2023. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36: 37193–37229.

Lu, S.; Chang, H.; Jing, E. P.; Boularias, A.; and Bekris, K. 2023. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, 1610–1620. PMLR.

Mackay, D. J. C. 1992. *Bayesian methods for adaptive models*. California Institute of Technology.

Nguyen, P.; Ngo, T. D.; Kalogerakis, E.; Gan, C.; Tran, A.; Pham, C.; and Nguyen, K. 2024. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4018–4028.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 815–824.

Rozenberszki, D.; Litany, O.; and Dai, A. 2024. Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19957–19967.

Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2023. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8216–8223. IEEE.

Shaham, U.; Stanton, K.; Li, H.; Nadler, B.; Basri, R.; and Kluger, Y. 2018. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*.

Song, S.; and Xiao, J. 2016. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 808–816.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 574–591. Springer.

Xu, M.; Xu, M.; He, T.; Ouyang, W.; Wang, Y.; Han, X.; and Qiao, Y. 2023a. Mm-3dscene: 3d scene understanding by customizing masked modeling with informative-preserved reconstruction and self-distilled consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4380–4390.

Xu, M.; Yin, X.; Qiu, L.; Liu, Y.; Tong, X.; and Han, X. 2023b. Sampro3d: Locating sam prompts in 3d for zero-shot scene segmentation. *arXiv preprint arXiv:2311.17707*.

Yang, Y.; Wu, X.; He, T.; Zhao, H.; and Liu, X. 2023. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*.

Zhao, J.; Zhuo, J.; Chen, J.; and Ma, H. 2025. SAM2Object: Consolidating View Consistency via SAM2 for Zero-Shot 3D Instance Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19325–19334.