Lightweight Open-Vocabulary 3D Object Detection via Semantic-Aware Voxel Aggregation

Syed Murtaza Hussain Abidi, Han Yuxin, *Soo Young Shin.
murtazaabidi@kumoh.ac.kr, ku.98953@kumoh-xin.by-works.net, * wdragon@kumoh.ac.kr

Abstract

Open-vocabulary 3D object detection enables recognition of unseen object categories without predefined label sets, yet existing approaches largely depend on point clouds or heavy supervision. This work presents a multi-view image-based framework for open-vocabulary 3D object detection that operates without human annotations. The method integrates class-agnostic 3D localization with hierarchical semantic alignment using pre-trained vision–language representations. A graph-guided pseudo-object generation strategy aggregates multi-view observations into coherent 3D structures, providing effective supervision during training. At inference, the system requires only posed RGB images and performs semantic matching using precomputed text embeddings. Experiments on large-scale indoor benchmarks demonstrate improved accuracy and efficiency compared to existing open-vocabulary and multi-view baselines.

Keywords: Open-vocabulary, 3D object detection.

I. Introduction

Recent advances in vision–language models (VLMs) have significantly expanded the scope of open-vocabulary visual understanding, enabling models to recognize concepts beyond predefined label sets. While open-vocabulary learning has been extensively explored in 2D image understanding, extending this capability to 3D object detection remains challenging due to the scarcity of annotated 3D data and the high cost of acquiring dense geometric representations at inference time. Existing open-vocabulary 3D detection methods predominantly rely on point clouds or depth sensing, either during training or inference [1]. Although effective, these approaches impose substantial hardware and computational constraints, limiting their applicability in real-world scenarios such as mobile robotics, augmented reality, and large-scale indoor mapping. Multi-view image-based 3D detection offers a promising alternative, yet most prior works remain confined to fixed-vocabulary settings or require extensive supervision [2].

In this work, address the problem of open-vocabulary multi-view 3D object detection without human annotations, using only posed RGB images during inference. Inspired by recent progress in voxel-based multi-view 3D representations and semantic alignment using pre-trained VLMs and noisy data [3].

Unlike prior approaches that rely on frame-wise pseudo labels, our method introduces a hierarchical semantic alignment strategy that integrates visual cues into a coherent 3D representation. We further propose a graph-guided pseudo-object construction mechanism that aggregates partial observations across viewpoints, producing high-quality pseudo 3D supervision without manual annotation. Our framework enables efficient open-vocabulary inference by decoupling semantic reasoning from heavy visual encoders at test time, making it suitable for real-time and resource-constrained environments. Extensive experiments on standard indoor benchmarks demonstrate that our approach achieves superior performance compared to existing open-vocabulary and multi-view baselines, while maintaining significantly lower computational cost.

II. Proposed System and Methodology

a. Dataset Overview:

3D scene understanding benchmarks: ScanNet200 and ARKitScenes. ScanNet200 provides large-scale RGB-D video sequences of indoor environments with over 200 object categories, enabling rigorous evaluation under open-vocabulary settings. ARKitScenes contains diverse real-world indoor scans captured using mobile devices and is primarily used to assess recall and generalization. For both datasets, we follow standard train–test splits and utilize only posed multi-view RGB images during inference. Depth information is used exclusively during training for pseudo-label generation and is not required at test time. This setup reflects realistic deployment conditions for image-based 3D perception systems.

b. Overall System model:

The proposed framework addresses open-vocabulary 3D object detection by integrating multi-view geometric reasoning with semantic representation learning. Given a set of posed RGB images, 2D visual features are extracted and back-projected into a unified 3D voxel space using known camera intrinsics and extrinsics. View-consistent aggregation is applied to enhance geometric stability and reduce noise arising from partial observations. A class-agnostic detection head operates on the resulting voxel features to localize candidate 3D objects, enabling robust object proposal generation without requiring category-specific supervision.

To enable open-vocabulary recognition, a hierarchical semantic alignment strategy is introduced to associate voxel-level and instance-level features with pre-trained vision–language embeddings. During training, class-agnostic 2D segments are lifted into partial 3D fragments and organized into a scene-level graph based on geometric overlap and spatial proximity. Graph-based clustering yields coherent pseudo 3D object instances that provide supervision for localization. Semantic alignment is performed at multiple levels by matching voxel and instance representations to visual-language features, allowing the model to capture diverse object appearances across viewpoints while maintaining semantic consistency.

III. Experimental Analysis:

The proposed approach is evaluated on large-scale indoor benchmarks under open-vocabulary settings using standard precision and recall metrics. Results demonstrate consistent improvements over existing multi-view and open-vocabulary baselines, particularly in challenging long-tail categories. Ablation studies confirm the effectiveness of graph-guided pseudo-object construction and hierarchical semantic alignment, while efficiency analysis shows that the method achieves competitive accuracy with significantly reduced inference cost by relying solely on RGB images at test time.

Experiments were conducted on ScanNet200 and ARKitScenes following standard train–test splits. Multi-view RGB images with known camera poses are used as input, while depth information is utilized only during training for pseudo-label generation. Performance is evaluated using mAP@25 and mAR@25 under open-vocabulary settings. All comparisons are performed under identical inference constraints to ensure fair evaluation.

IV. Conclusion and Future work

In this paper, we proposed framework for open-vocabulary multi-view 3D object detection without human annotations. By combining graph-guided pseudo-object generation with hierarchical semantic alignment, our approach effectively bridges the gap between multi-view geometry and open-vocabulary semantic reasoning. The proposed method achieves strong detection performance while remaining computationally efficient, making it suitable for real-world deployment. This work opens new directions for scalable 3D perception systems that can generalize beyond fixed label spaces using only visual data.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00553810, 50%) This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2025-RS-2022-00156394, 50%) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)

REFERENCES

[1] Hsu, Peng-Hao, Ke Zhang, Fu-En Wang, Tao Tu, Ming-Feng Li, Yu-Lun Liu, Albert YC Chen, Min Sun, and Cheng-Hao Kuo. "Openm3d: Open vocabulary multi-view indoor 3d object detection without human annotations." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8688-8698. 2025.

[2] Baruch, Gilad, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer et al. "Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data." arXiv preprint arXiv:2111.08897 (2021).

[3] Jia, Chao, 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, pp. 4904-4916. PMLR, 2021.