Leveraging Vision Foundation Models for Zero-Shot Open Vocabulary 3D Scene Understanding

Supasun Khumpraphan

August 15, 2024

Abstract

3D scene understanding aims to extract and interpret semantic information of objects within their contextual environments, using diverse fields such as robotics, augmented/virtual reality, remote sensing, and 3D medical imaging. While recent advancements in large vision language models have demonstrated remarkable performance in 2D domains, their application in 3D perception tasks remains largely unexplored. This research proposal seeks to bridge this gap by leveraging Vision Foundation Models (VFMs) for zero-shot open vocabulary 3D scene understanding. The study aims to adapt existing 2D vision models, such as CLIP, DALL-E, and Vision Transformer (ViT), to effectively process 3D data, including point clouds and 3D meshes. Key objectives include implementing zero-shot learning techniques to enable the model to classify objects in 3D scenes without labeled datasets and developing algorithms for open vocabulary scene understanding to recognize a wide range of objects and environments. The research will explore and integrate multimodal data sources from image-3D point cloud datasets, enhancing the model's contextual interpretation capabilities. Practical applications include improving autonomous navigation, terrestrial mapping, and assistive technologies in robotics, enhancing realism in AR/VR environments, advancing urban modeling and vegetation monitoring in remote sensing, and improving diagnostic capabilities in 3D medical imaging. This project promises to extend the capabilities of VFMs, facilitating significant advancements in various applied domains through robust and comprehensive 3D scene understanding.

1 Introduction

Understanding 3D scenes involves extracting and interpreting the semantic information of objects within their contextual environments, a capability essential to a broad spectrum of applications [4]. From robotics, which relies on accurate scene understanding for tasks like autonomous navigation and assistive technologies, to augmented and virtual reality (AR/VR), which demands high-fidelity environmental interactions, the ability to decode 3D scenes underpins many technological advancements [2]. Additionally, remote sensing applications such as urban modeling, vegetation monitoring, and surveying, as well as 3D medical imaging, benefit significantly from enhanced 3D scene understanding [1]. Despite the impressive strides made in the 2D domain through the utilization of large vision language models, their adaptation for 3D perception tasks remains an underexplored territory [5]. Current methods leveraging Vision Foundation Models (VFMs) such as CLIP, DALL-E, and Vision Transformers (ViT) have shown promising results in understanding and classifying 2D images [3]. However, the transition of these models to the 3D domain presents both a challenge and an opportunity to extend their capabilities to more complex and dimensional data forms, including point clouds and 3D meshes. The goal of this research proposal is to bridge this gap by adapting these powerful 2D vision models to 3D data, thereby enabling zero-shot learning for 3D scene understanding in an open vocabulary setting. Zero-shot learning facilitates the classification and recognition of objects without the need for labeled datasets, leveraging the extensive pretrained knowledge of VFMs. This approach allows for generalization across a wide range of objects and scenes, making the models highly versatile and applicable to numerous practical scenarios. To achieve this, the research will involve a systematic adaptation of existing vision models to process and interpret 3D data. It will explore and adapt image-3D point cloud datasets, integrating multimodal data sources to enhance the models' understanding and contextual interpretation. The developed algorithms will enable open vocabulary scene understanding, allowing the models to recognize and interpret a vast array of objects and environments beyond their training sets. This study will have significant implications across various fields. In robotics, it will improve autonomous navigation, terrestrial mapping, and assistive technologies. AR/VR will enhance the realism and interaction within virtual environments. Remote sensing will advance urban modeling, vegetation monitoring, and surveying. Finally, medical imaging will contribute to better diagnostic and analytical capabilities. Through this comprehensive approach, the research aims to revolutionize 3D scene understanding, paving the way for advanced applications and innovations in numerous domains.

2 Research Questions

The main research questions are as follows:

- How can existing 2D Vision Foundation Models (e.g., CLIP, DALL-E, Vision Transformer) be effectively adapted to process and interpret 3D data, including point clouds and 3D meshes?
- What methodologies can be implemented to enable zero-shot learning in 3D scene understanding, allowing models to classify objects and scenes without relying on labeled datasets?
- How can algorithms be developed to support open vocabulary scene understanding, enabling models to recognize and interpret a wide range of objects and environments beyond their initial training sets?
- What are the most effective approaches to integrating multimodal data sources from existing image-3D point cloud datasets to enhance the contextual interpretation capabilities of Vision Foundation Models?
- In what ways can the adapted 3D scene understanding models be applied to specific use cases such as robotics, augmented/virtual reality, remote sensing, and medical imaging to improve their practical effectiveness and impact?

3 Expected Outcomes

The expected outcomes of the research are:

- Successful Adaptation of Vision Foundation Models for 3D Data: The research is expected to demonstrate effective methodologies for adapting existing 2D Vision Foundation Models, such as CLIP, DALL-E, and Vision Transformer (ViT), to process and interpret 3D data, including point clouds and 3D meshes.

- Implementation of Zero-Shot Learning Techniques: The study aims to develop robust zero-shot learning techniques that enable the adapted models to classify and understand objects and scenes in 3D environments without the need for labeled datasets, leveraging pre-trained knowledge for generalization.

- Development of Open Vocabulary Scene Understanding Algorithms:Algorithms will be created to support open vocabulary scene understanding, allowing the models to recognize and interpret a broad range of objects and environments beyond their initial training sets, guided by natural language descriptions.

- Enhanced Multimodal Data Integration: The research will lead to the successful integration of multimodal data sources, enhancing the models' ability to understand and interpret 3D scenes contextually. This includes adapting and utilizing existing image-3D point cloud datasets.

- Application Across Diverse Domains: The adapted models and developed techniques will have significant practical applications, enhancing autonomous navigation and assistive technologies in robotics, improving realism and interaction in AR/VR environments, advancing urban modeling and vegetation monitoring in remote sensing, and enhancing diagnostic and analytical capabilities in 3D medical imaging.

4 Methodology

4.1 Data Source

The primary data sources for this research will include existing image-3D point cloud datasets such as PointNet, PointNet++, and VoxelNet. These datasets offer a diverse range of 3D scenes and objects, providing a rich foundation for model training and evaluation. Additional multimodal datasets integrating image, point cloud, and other sensor data will also be utilized to enhance the model's contextual understanding capabilities.

4.2 Data Cleaning and Preprocessing

Data cleaning and preprocessing are crucial steps to ensure the quality and consistency of the datasets. The process will involve filtering out irrelevant data points and reducing noise in point cloud datasets to improve data quality. Standardizing data formats and scales across different datasets will ensure uniformity. Data augmentation techniques will generate a more diverse dataset, aiding the model's ability to generalize across various scenarios. Segmenting the 3D point clouds into meaningful components will facilitate easier interpretation and model training.

4.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis will virtualize to gain insights into the datasets and identify key patterns and relationships. This will include calculating summary statistics to understand the distribution and central tendencies of the data, creating 2D and 3D visualizations to explore the spatial relationships and structures within the data, and identifying correlations between different features in the datasets to inform model design and feature selection.

4.4 Model Implementations

The core of this research involves adapting existing 2D Vision Foundation Models (VFMs) such as CLIP, DALL-E, and Vision Transformer (ViT) to process and interpret 3D data. The model implementation process will include extending the architecture of 2D VFMs to handle 3D data inputs, including point clouds and 3D meshes. Zero-shot learning algorithms will be implemented to enable the models to classify objects in 3D scenes without labeled datasets, leveraging pre-trained knowledge. Additionally, algorithms will be developed to allow the models to recognize and interpret a wide range of objects and environments beyond the training set, guided by natural language descriptions.

4.5 Model Evaluation

Model evaluation is very important in assessing the performance and effectiveness of the adapted models. The evaluation process will involve utilizing metrics such as accuracy, precision, recall, and F1-score to measure model performance. Comparing the performance of the adapted models against existing state-of-the-art methods in 3D scene understanding will provide a benchmark. Ablation studies will be conducted to understand the impact of different components and techniques on model performance.

4.6 Model Deployment

The final step of the research will focus on deploying the developed models for practical applications. This will involve integrating the models into various application domains such as robotics, AR/VR, remote sensing, and medical imaging. Ensuring the models can process data in real time will be particularly important for applications requiring immediate feedback, such as autonomous navigation and augmented reality. Developing user-friendly interfaces for interacting with the models will facilitate ease of use and accessibility for end-users.

5 Conclusion

This research proposal aims to advance the field of 3D scene understanding by leveraging Vision Foundation Models (VFMs) for zero-shot open vocabulary recognition and interpretation. Recognizing the impressive capabilities of VFMs in the 2D domain, this study seeks to adapt these models to the more complex realm of 3D data. By exploring and integrating existing image-3D point cloud datasets, the proposed research will develop methods to extend VFMs such as CLIP, DALL-E, and Vision Transformer (ViT) to handle 3D data effectively. The methodology includes a comprehensive approach to data sourcing, cleaning, and preprocessing, ensuring high-quality and consistent datasets for model training. Through exploratory data analysis, the research will identify key patterns and relationships within the data, informing the adaptation and extension of VFMs. The implementation of zero-shot learning techniques will enable the models to classify and understand 3D scenes without relying on labeled datasets, leveraging their pre-trained knowledge for generalization. Developing algorithms for open vocabulary scene understanding will allow the models to recognize and interpret a wide range of objects and environments guided by natural language descriptions. The integration of multimodal data sources will enrich the models' contextual understanding, enhancing their performance in various practical applications. The expected outcomes include significant advancements in autonomous navigation, terrestrial mapping, housekeeping robots, and assistive technologies for elderly care and agriculture within robotics. Additionally, the research will improve realism and interaction in augmented and virtual reality environments, advance urban modeling, vegetation monitoring, and surveying in remote sensing, and enhance diagnostic and analytical capabilities in 3D medical imaging. By following this comprehensive and systematic approach, this research aims to bridge the gap between 2D and 3D scene understanding, paving the way for innovative applications and substantial advancements across multiple domains.

6 Candidate profile

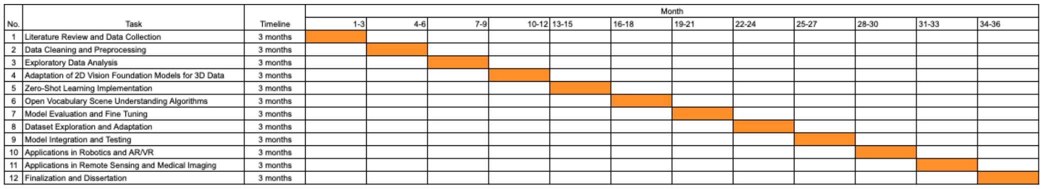
- MSc Artificial Intelligence
  University of Essex (First class honor)
  Publication: Submitted an "Experimental Analysis of Speech Emotion and Gender Recognition" paper to the IEEE Transactions on Affective Computing journal, which is under review.
  January 2022 - January 2023

- BEng Software Engineer
  Thammasat University
  June 2015 - June 2018

7 Project Plan

References

[1] Jibril Muhammad Adam, Weiquan Liu, Yu Zang, Muhammad Kamran Afzal, Saifullahi Aminu Bello, Abdullahi Uwaisu Muhammad, Cheng



			Month											
No.	Task	Timeline	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
1	Literature Review and Data Collection	3 months	🟧											
2	Data Cleaning and Preprocessing	3 months		🟧										
3	Exploratory Data Analysis	3 months			🟧									
4	Adaptation of 2D Vision Foundation Models for 3D Data	3 months				🟧								
5	Zero-Shot Learning Implementation	3 months					🟧							
6	Open Vocabulary Scene Understanding Algorithms	3 months						🟧						
7	Model Evaluation and Fine Tuning	3 months							🟧					
8	Dataset Exploration and Adaptation	3 months								🟧				
9	Model Integration and Testing	3 months									🟧			
10	Applications in Robotics and AR/VR	3 months										🟧		
11	Applications in Remote Sensing and Medical Imaging	3 months											🟧	
12	Finalization and Dissertation	3 months												🟧

Figure 1: Plan projects with a monthly schedule.

Wang, and Jonathan Li. Deep learning-based semantic segmentation of urban-scale 3d meshes in remote sensing: A survey. *International Journal of Applied Earth Observation and Geoinformation*, 121:103365, 2023.

[2] Qiao Lan, Dingzhu Wen, Zezhong Zhang, Qunsong Zeng, Xu Chen, Petar Popovski, and Kaibin Huang. What is semantic communication? a view on conveying meaning in the era of machine intelligence. *Journal of Communications and Information Networks*, 6(4):336–371, 2021.

[3] Xu Yan, Haiming Zhang, Yingjie Cai, Jingming Guo, Weichao Qiu, Bin Gao, Kaiqiang Zhou, Yue Zhao, Huan Jin, Jiantao Gao, et al. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. *arXiv preprint arXiv:2401.08045*, 2024.

[4] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.