3D point cloud understanding via Multi-modal Large Language Models

PhD topic proposal - Campagne EDITE 2025

General information

- **Doctoral school:** EDITE (ED130 Informatique Telecommunications et Électronique de Paris)
- **Duration:** 3 years
- **Director:** Laurent Wendling [1] (laurent.wendling@u-paris.fr)
- **Co-supervsion:** Ayoub Karine [2] (ayoub.karine@u-paris.fr)
- **Institute:** Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE), team Systèmes Intelligents de Perception
- **Location:** 45 rue des Saints-Pères, 75006, Paris
- **Keywords:** 3D point cloud, Large Language Model (LLM), foundation models, multi-modal learning, scene understanding
- **Application process:** To apply, follow the instructions provided in this link: https://www.edite-de-paris.fr/ouverture-des-candidatures-campagne-edite-2025/. Student applications are open from April 15th to May 15th.

Proposed topic

Context

With the rapid advancement of 3D acquisition systems such as LIDAR and RGB-D, 3D point cloud (3DPC) are becoming more popular in several applications including autonomous driving, robotic, agriculture and so on [1]. Such data offers highly realistic representations of large variety of 3D objects and scenes. 3DPC is composed by a set of data points in space characterized by its position (x, y, z). Other information can also be stored with each point like color, normal and curvature.

The interpretation of these 3DPC leads to understand 3D object and scene. With the promising results of deep learning in 2D computer vision tasks, this approach is extended to 3DPC for classification, segmentation and object detection. Unlike grid-based 2D image data, 3D point clouds are characterized by spatial sparsity and irregular distribution, making it challenging to directly adapt methods from the 2D domain.

On the other hand, interacting with these 3DPC is of paramount importance for many real-world applications that involve perception and navigation. This interaction can be done through text query. Consequently, the input of the adopted deep learning architectures should handle 2 modalities : text and images. It can be realized by multi-modal learning. It is in this context that the present doctoral subject is situated.

Work to be done

Recent works have demonstrated the potential of handling jointly the 3D and text modalities via multi-modal Large Language Models (LLMs) [2]. The current 3D vision-language tasks include 3D captioning (3D $\rightarrow$ Text), 3D grounding (3D + Text $\rightarrow$ 3D Position), 3D conversation (3D + Text $\rightarrow$ Text), 3D embodied agents (3D + Text $\rightarrow$ Action) and text-to-3D generation (Text $\rightarrow$ 3D). Recently, a new vision-language task is emerged that produce a point-wise segmentation masks given a 3DPC and text query. Some examples of this task are depicted in Figure 1. We will focus on this task in this thesis. The challenge is to develop new methods that provide intelligent interaction ways between text query and 3DPC while harnessing the Large Language Model's reasoning capabilities. TGNN [3] is the first work to solve this challenging problem by proposing an architecture composed by two parts: instance segmentation using 3D-UNet and instance referring using Text-Guided Graph Neural Network. He *et al.* [4] propose

[1] https://helios2.mi.parisdescartes.fr/ lwendlin/
[2] https://www.ayoub-karine.com/

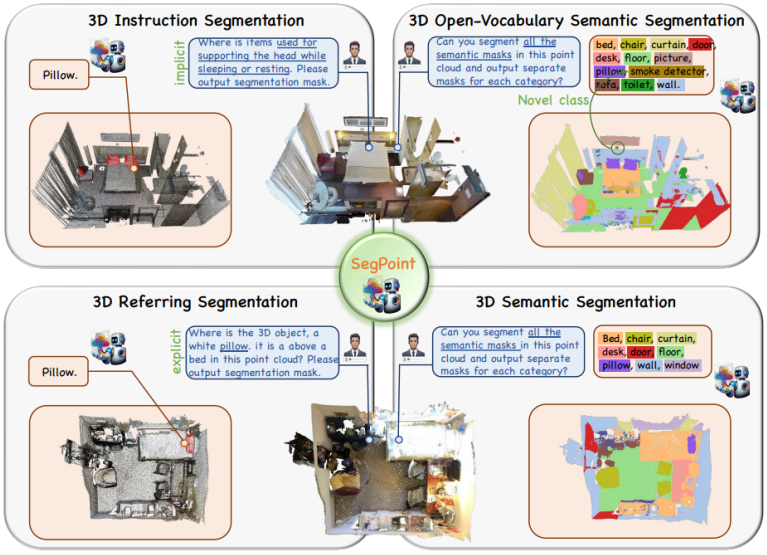

Figure 1: 3D point cloud segmentation guided by text query [5].

to join vision and language features using the contrastive learning. SegPoint [5] developed a unified framework for 3D point cloud segmentation and proposed a new dataset, Instruct3D, for 3D instruction segmentation. More recently, Deng *et al.* [6] develop 3D-LLaVA method that unifies various 3D tasks using an omni superpoint transformer.
The different steps of this PhD thesis are :

▷ **Step 1 – Literature review (T1 → T5):** the candidate will review the different 2D/3D vision-language methods that exploit multi-modal LLMs. An implementation of some baseline methods is to be realized.

▷ **Step 2 – Exploiting 2D vision foundation models and 2D-LLMs (T6→ T13):** to deal with unstructured 3DPC data, the 3DPC will be projected to 2D views. In this way, we will leverage the growth of 2D vision foundation models (e.g. SAM [7]) and 2D-LLMs (e.g. LISA [8]). The 3D-2D alignment will be conducted through the knowledge distillation [9].

▷ **Step 3 – Construction of 3DPC foundation model (T14 → T21):** the literature misses from a vision foundation model for 3DPC. Through a self-supervised learning, a new 3DPC foundation model is to develop. The innovation will be the definition of the pre-text tasks related to the 3DPC.

▷ **Step 4 – Development of new 3D-LLMs for point-level reasoning and segmentation (T22 → T30):** the developed 3DPC foundation model in step 3 will be incorporated in a unified framework for 3D point cloud segmentation via the reasoning ability of LLMs. In addition of text prompts, other non-textual prompts could be investigated such as points and boxes.

▷ **Step 5 – PhD Thesis writing and Defense (T31 → T36)**

The contributions related with each step will be evaluated in public datasets (Instruct3, ScanNet, ...). The results will be published in outstanding computer vision conferences (CVPR, ICCV, ECCV, WACV, ...) and journals (PR, IEEE TM, PRL, ...).

Desired background for the candidate

We are looking for a Master 2 student or final year of MSc, or engineering school in computer science. The ideal candidate should have knowledge in deep learning, computer vision, Python programming and an interest in handling 3D data.

Bibliography

[1] Salima Bourbia, Ayoub Karine, Aladine Chetouani, Mohammed El Hassouni, and Maher Jridi. "No-reference 3d point cloud quality assessment using multi-view projection and deep convolutional neural network". In: *IEEE Access* 11 (2023), pp. 26759–26772.

[2] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. "When LLMs step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models". In: *arXiv preprint arXiv:2405.10255* (2024).

[3] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. "Text-guided graph neural networks for referring 3d instance segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2. 2021, pp. 1610–1618.

[4] Shuting He and Henghui Ding. "RefMask3D: Language-guided transformer for 3D referring segmentation". In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 8316–8325.

[5] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. "Segpoint: Segment any point cloud via large language model". In: *European Conference on Computer Vision*. Springer. 2024, pp. 349–367.

[6] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. "3D-LLaVA: Towards Generalist 3D LMMs with Omni Superpoint Transformer". In: *arXiv preprint arXiv:2501.01163* (2025).

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[8] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. "Lisa: Reasoning segmentation via large language model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9579–9589.

[9] Ayoub Karine, Thibault Napoléon, and Maher Jridi. "Channel-spatial knowledge distillation for efficient semantic segmentation". In: *Pattern Recognition Letters* 180 (2024), pp. 48–54.