Postdoctoral Research Topic

- Title of the proposed topic: Open-World 3D Scene Understanding by Fusion of LiDAR and Vision-Language Models
- Research axis of the 3IA: Axe 4 - AI for Smart and Secure Territories
- Supervisor: Ezio Malis
- Research group: ACENTAURI project-team, Inria Center at Université Côte d'Azur

1 Context

In France, 49% of people living in agglomerations with fewer than 100,000 inhabitants lack access to public transport near their homes [1]. As a result, suburban areas are highly dependent on personal vehicles. This dependency not only places significant financial stress on suburban households but also has a substantial environmental impact. Autonomous transportation is a viable solution as it provides people with reliable alternatives. Despite years of research efforts, the deployment of autonomous transportation remains at pilot projects [2].

Autonomous vehicles heavily rely on LiDARs to perceive their environment because this sensor can produce precise depth measurement at a high density. LiDARs measurements are generally sparse, mainly geometric and lacks semantic information. Therefore, LiDAR-based perception models often fail to detect objects of some classes. This drawback is an important reason for the limited deployment of autonomous transportation. As the diversity of the open environment introduces various types of objects that are hardly covered by any datasets, the risk of detection failure is high and so is the risk of unsafe circulation.

In contrast to the lack of semantic in LiDAR measurement, the semantic information is abundant in vision and language, covering every class of objects that we can name. The recent advancements of language modelling enables embedding natural languages to the latent space of Large Language Models (LLM) (e.g., Llama, or GPT). Leveraging this capacity of LLM, a number of works (e.g.,[3, 4]) grounds the latent space of vision models to theirs, enabling querying vision models using natural languages. This new class of vision models, referred to as Vision-Language Models (VLM), is able to make predictions beyond the classes covered by their training data.

2 Postdoc Subject

The main goal of this postdoc is to develop open-world 3D scene understanding models through the fusion of LiDAR-based models and VLM. This goal can be achieved by solving the following two scientific challenges.

The first scientific challenge to address is how to effectively fuse the latent space of LiDAR-based models with VLM. This is challenging due to the difference between measurements by LiDAR and images. While images are dense 2D grids, point clouds made by LiDARs are sparse and unstructured. This difference gives rise to the central question of LiDAR-camera fusion, that is to find a unified representation for the two modalities that allow an effective fusion. A large body of work simply projects LiDAR points to images using projective geometry (e.g.,[5, 6]) and uses the image plane as the representation for fusion. Arguing that such a projection results in a sparse fusion as point clouds only cover a small fraction of images, modern works lift images to 3D voxel grids using monocular depth prediction [7, 8, 9]. A comprehensive benchmarking of these methods will offer valuable insight. Moreover, model-based approaches for 3D reconstruction should also be explored to enable a mathematically sound lifting of images to 3D.

The second scientific challenge is the adaptation of the obtained LiDAR-VLM model in the range of 3D scene understanding tasks such as semantic segmentation, object detection, scene graph generation. A prospective method is transfer learning which freezes the backbone of the LiDAR-VLM model and fine tunes its output layers to adapt it to a particular task. An important result that we aim to prove is the data efficiency of the LiDAR-VLM model. As this model is packed with semantic information and knowledge of geometric relation among points, they should require less training data compared to models that are trained from scratch while not compromising the performance. Another result that we aim to obtain is the zero-shot or few-shot capacity of the LiDAR-VLM model. In the zero-shot setting, we want to show that the LiDAR-VLM model can reliably detect objects that are entirely absent from its training set. On the other hand, the few-shot setting allows for fine-tuning the model with a small number of examples of the objects that we want to detect.

The obtained LiDAR-VLM model will be benchmarked against state-of-the-art methods on publicly available autonomous driving datasets [10, 11]. It will also be tested on the autonomous vehicles of the ACENTAURI team.

3 Work Plan

The work of this postdoc includes:

- Studying the state-of-the-art of LiDAR-camera fusion and VLM
- Benchmarking model-based and learning-based techniques for LiDAR-camera fusion
- Designing new LiDAR-Camera fusion method that focus on the accurate and dense alignment of the two modalities
- Adapting the obtained LiDAR-VLM model to perception tasks including semantic segmentation, object detection, and scene graph generation
- Writing articles for international journals and conferences

4 Skills

The candidate should preferably have a PhD in Computer Science or Robotics with a solid background on deep learning and 3D scene understanding. Experience with LiDAR

and Computer Vision is a plus. The candidate should be proficient in Python, PyTorch or Scikit-learn. The candidate should also be endowed with a strong passion for multidisciplinary studies and all aspects of research ranging from fundamental work to experimental work. Finally, a good level of English is important.

5 How To Apply

Interested candidates must send to Ezio Malis at `ezio.malis@inria.fr` the following documents:

- Letter of recommendation of the supervisor
- Curriculum vitæ including the list of the scientific publications
- Motivation letter
- Letter of recommendation of the thesis supervisor

All the requested documents must be gathered and concatenated in a single PDF file named in the following format: LAST NAME of the candidate_Last Name of the supervisor_April_2025.pdf.

References

[1] "Les trois quarts des français continuent d'utiliser la voiture pour leurs trajets domicile- travail," https://www.ouest-france.fr/economie/transports/voiture/les-trois-quarts-des-francais-continuent-dutiliser-la-voiture-pour-leurs-trajets-domicile-travail-0d202596-75c4-11ef-8ac3-f7eb9db10673, accessed: 2025-02-11.

[2] "Kelride paves the way for weatherproof autonomous public transport in germany," https://easymile.com/news/kelride-paves-way-weatherproof-autonomous-public-transport-germany, accessed: 2025-02-11.

[3] M. Sodano, F. Magistri, L. Nunes, J. Behley, and C. Stachniss, "Open-world semantic segmentation including class similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3184–3194.

[4] Y. Zeng, X. Zhang, H. Li, J. Wang, J. Zhang, and W. Zhou, "X 2-vlm: All-in-one pre-trained model for vision-language tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[5] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 35–52.

[6] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[7] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[8] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.

[9] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-bev: What really matters for multi-sensor bev perception?" in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2759–2765.

[10] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[11] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 178–20 188.