PhD proposal – 2026

Geometric-Semantic Adaptation of Multimodal LLMs for High-level Landmark Detection in Complex Environments

Contacts: Prof. Gilles Simon (gilles.simon@loria.fr)
Dr. Vincent Gaudillière (vincent.gaudilliere@loria.fr)

1 General information

Position: PhD

Duration: 36 months, starting in October 2026

Location: Loria, Nancy, France

Affiliation: TANGRAM team (Inria-Loria)

Supervision: Gilles Simon (supervisor), Vincent Gaudillière (co-advisor) and Marie-Odile Berger

2 Description

2.1 Context

Landmark detection, description and matching is the cornerstone of autonomous visual localization systems deployed in unknown environments. While most widely-adopted and accurate solutions exploit low-level landmarks (i.e., points or lines), dealing with large-scale and visually ambiguous environments remains highly challenging due to the inherent multiplicity, ambiguity and sensitivity of local primitives. In the perspective of localization systems with larger scope of application, high-level landmarks such as objects present in the scene have proven to offer key advantages like lower multiplicity, higher detection repeatability across viewpoints, and lower ambiguity compared to their local counterparts [1, 2, 8]. However, existing solutions require the prior intervention of an expert to identify the object landmarks that can be used for localization in a given environment. Moreover, object detectors used in these methods must be finetuned to recognize objects beyond common categories. The recent emergence of Multimodal Large Language Models (MLLMs) represents

a promise of lower human intervention and easier deployment, but the consistency of their predictions under camera movements and their geometric accuracy are still to demonstrate. Moreover, challenges posed by complex man-made environments such as museums or factories, often featuring intra-class variations of uncommon objects, are to be addressed.

2.2 Objectives

The research of this PhD will be articulated around the concept of useful landmark for localization in complex environments. Indeed, unlike cases where object detection or segmentation methods are used with no other objective than their own, using objects as landmarks for localization introduces specific constraints in terms of repeatability across different viewpoints, distinctiveness with respect to other landmarks, geometric accuracy and adequate distribution within the environment.

To address these challenges, we propose to exploit the possibilities offered by MLLMs (*e.g.*, BLIP-2 [3], LLaVA [4], MiniGPT-4 [7]), able to follow instructions or answer questions about an image, to extract localization information from images. More precisely, we want to examine how their general-purpose detection and segmentation abilities can be redirected towards automatically identifying high-level localization landmarks in specialized environments. For that, we first propose to assess both geometric and semantic sensitivity of different MLLMs to different combinations of visual and textual prompts [5, 6], in order to derive automated prompting strategies. In particular, we want to study integration of 3D geometric and fine-grained semantic information within the prompts, and assess geometric accuracy of corresponding models' answers. If necessary, we will then propose dedicated learning strategies for inducing the desired geometric capabilities within the model. In a second phase, we want to examine potential complementarity between MLLMs and scene graphs built from images to combine localization methods with adequate scene modeling.

Discussions with local LLMs experts will be held throughout the project to help the PhD student deal with the specific characteristics of the language modality.

3 Profile

- The candidate is completing a Master's or engineering's degree in Computer Vision, Electrical Engineering, Computer Science, Applied Mathematics or a related field.
- A strong background in image processing or/and in computer vision is required.
- Strong programming skills in Python.
- Strong mathematical background.
- Familiarity with deep learning frameworks such as PyTorch.
- Commitment, team working and a critical mind.
- Fluent verbal and written communication skills in English.

4 How to apply

Interested candidates are encouraged to send their applications (detailed CV, transcripts and a brief motivation letter) as soon as possible to the following addresses: gilles.simon@loria.fr and vincent.gaudilliere@loria.fr. Applications will be processed upon reception.

References

[1] V. Gaudillière, G. Simon, and M.-O. Berger. Camera Relocalization with Ellipsoidal Abstraction of Objects. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 8–18, Oct. 2019. ISSN: 1554-7868.

[2] V. Gaudillière, G. Simon, and M.-O. Berger. Perspective-2-Ellipsoid: Bridging the Gap Between Object Detections and 6-DoF Camera Pose. *IEEE Robotics and Automation Letters*, 5(4):5189–5196, Oct. 2020.

[3] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 19730–19742, Honolulu, Hawaii, USA, July 2023. JMLR.org.

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual Instruction Tuning. Nov. 2023.

[5] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V, Nov. 2023. arXiv:2310.11441 [cs].

[6] L. Yang, X. Li, Y. Wang, X. Wang, and J. Yang. Fine-Grained Visual Text Prompting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1594–1609, Mar. 2025.

[7] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. Oct. 2023.

[8] M. Zins, G. Simon, and M.-O. Berger. OA-SLAM: Leveraging Objects for Camera Relocalization in Visual SLAM. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 720–728, Oct. 2022. ISSN: 1554-7868.