PhD proposal – 2025

# Language-aided Detection and Matching of Semantic Landmarks for Visual Localization 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 September/October 2025

**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 such as points or lines, dealing with large-scale and/or visually ambiguous environments remains highly challenging due to the inherent multiplicity, ambiguity and sensitivity of such local primitives. In the perspective of visual localization systems with broader scope of application, high-level landmarks such as objects present in the scene have proven to offer key advantages such as lower multiplicity, higher detection repeatability across viewpoints and sensors, and potentially lower ambiguity compared to their local counterparts [1, 2, 8]. However, current solutions are limited to pre-defined categories of objects and detectors need to be fine-tuned to handle novel uncommon categories. The recent emergence of *zero-shot* or *open-vocabulary* object detectors based on vision-only and vision-language foundation models represents a promising

alternative, but their exploitability for solving precise visual localization task (i.e., pose estimation) is still to demonstrate. Moreover, the challenges posed by complex man-made environments such as factories, often featuring intra-class variations of specialized equipment rather than common distinctive objects, are to be addressed. Ultimately, the question of environments that do not contain objects per se, such as natural terrains, remains largely unexplored.

## 2.2 Objectives

The research of this PhD will be articulated around the concept of useful landmark for localization, that can fit different environments and application scenarios. Indeed, unlike cases where object detection or segmentation methods are used with no objective than their own, using objects as landmarks for localization introduces specific constraints. Notably, landmarks must be consistently perceived from a wide range of viewpoints and reliably re-identified when they reappear in new images. Such requirements might be more or less stringent depending on the type of environment within which the system is deployed. In other words, perceiving common objects in moderately complex scenes is less demanding than perceiving uncommon objects in real-life specialized environments. To understand the complexity of landmark selection and derive automated processes, we are targeting challenging application scenarios within complex unknown environments, such as autonomous computer vision systems operating in a factory or on an extraterrestrial planet.

To address these challenges, we propose to exploit the possibilities offered by pre-trained foundation models (e.g., [3, 6, 5]) and we are particularly interested in the possible contributions of vision-language alignment models such as CLIP [6]. More precisely, we want to first examine how general-purpose unsupervised detection and segmentation models can be guided towards extracting Potential Objectness Landmark (POL) in specialized environments in a zero-shot manner, by leveraging adequate visual and text prompting strategies [7]. We then want to study how language-based description of POL can encapsulate geometric and semantic properties relevant for POL re-identification across viewpoints, according to the way these descriptions are extracted from images. Finally, we want to combine the proposed landmark detection and description approaches with off-the-shelf object-based localization methods [2, 8, 4], in order to be tested in two complementary types of environments: industrial settings (e.g., factories, plants, ships) and extraterrestrial terrains (i.e., Moon or Mars surface).

# 3 Profile

- The candidate holds 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] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, Oct. 2023. ISSN: 2380-7504.

[4] S. Matsuzaki, T. Sugino, K. Tanaka, Z. Sha, S. Nakaoka, S. Yoshizawa, and K. Shintani. CLIP-Loc: Multi-modal Landmark Association for Global Localization in Object-based Maps. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13673–13679, May 2024.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. ISSN: 2640-3498.

[7] 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.

[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.