End-of-study engineering internship proposal – 2025

Open-vocabulary object detection for enhanced object-aided visual SLAM

Contact: Dr. Vincent Gaudillière (vincent.gaudilliere@loria.fr)

1 General information

Position: M2 internship

Duration: 3 months, starting in November 2025

Location: Loria, Nancy, France

Affiliation: TANGRAM team (Inria-Loria)

Supervisors: Vincent Gaudillière, Marie-Odile Berger and Gilles Simon

2 Context, description and objectives

This internship will deal with the problem of relocalization in visual SLAM, which involves determining a camera's viewpoint by automatically matching features in an image with elements from a known 3D model of the environment. These features are referred to as landmarks.

Object-based relocalization [2, 9, 5] uses "high-level" landmarks, such as objects (*e.g.*, chairs, tables, cupboards), as opposed to the more commonly used "low-level" keypoints (*e.g.*, SIFT [4], ORB [6]). This approach offers the advantage of relying on more robust and discriminative landmarks but is currently limited to environments that are rich in common objects, or requires fine-tuning the object detector. However, fine-tuning the object detector to handle specific objects often involves a prior tedious data collection and annotation process. The recent emergence of *open-vocabulary* object detectors [8, 1, 7, 3] represents a promising alternative, but their unconstrained label predictions (*i.e.*, object categories) represent a challenge for object re-identification across viewpoints.

In this internship, we will study the integration of an *open-vocabulary* object detector (*e.g.*, YOLO-World [1]) into an object-aided visual SLAM pipeline (OA-SLAM [9]). The first part of the internship will consist in understanding (i) the OA-SLAM software[1] developed in the team, and (ii) one open-vocabulary object detection sofware[2]. If needed, OA-SLAM installation instructions will be completed or updated. The second part will involve implementing the chosen open-vocabulary detector into OA-SLAM, as an alternative to the close-vocabulary object detector currently implemented. For that, special attention will be given to object re-identification based on unconstrained object categories. Optimal strategies will be assessed based on thorough experimental evaluation. The final part of the internship will consist in documenting the newly-developed software to facilitate its reuse by TANGRAM researchers, writing a scientific report and presenting the internship work to the TANGRAM team.

References

[1] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-Time Open-Vocabulary Object Detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 16 901–16 911, iSSN: 2575-7075. [Online]. Available: https://ieeexplore.ieee.org/document/10657649

[2] 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)*, Oct. 2019, pp. 8–18, iSSN: 1554-7868. [Online]. Available: https://ieeexplore.ieee.org/document/8943719

[3] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, vol. 15105, pp. 38–55, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-72970-6_3

[4] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[5] 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)*, May 2024, pp. 13 673–13 679. [Online]. Available: https://ieeexplore.ieee.org/document/10611393

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer*

[1] https://gitlab.inria.fr/tangram/oa-slam

[2] https://github.com/AILab-CVC/YOLO-World

Vision, Nov. 2011, pp. 2564–2571, iSSN: 2380-7504. [Online]. Available: https://ieeexplore.ieee.org/document/6126544

[7] L. Yao, R. Pi, J. Han, X. Liang, H. Xu, W. Zhang, Z. Li, and D. Xu, "DetCLIPv3: Towards Versatile Generative Open-Vocabulary Object Detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 5610–5619, iSSN: 2575-7075. [Online]. Available: https://ieeexplore.ieee.org/document/10658129

[8] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. H. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "GLIPv2: unifying localization and vision-language understanding," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., Nov. 2022, pp. 36 067–36 080.

[9] 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)*, Oct. 2022, pp. 720–728, iSSN: 1554-7868. [Online]. Available: https://ieeexplore.ieee.org/document/9995573