Self-supervised Adaptation for Open Vocabulary Semantic Segmentation via 3D Mapping

Supervisors: Michele Antonazzi, Timon Homberger

May 25, 2026

Introduction

In recent years, open-vocabulary semantic segmentation (OVSS) has become extremely popular as it enables a robot to detect objects from textual prompts without requiring a predefined set of class lables. This task is performed by modern visual foundation models (e.g., SED [6] or CATSeg [2]) based on CLIP [4]. Despite their great potential and generalization capabilities, these models are affected by the well-known *domain shift* when deployed on mobile robots. It occurs when the training and test data distributions differ, causing performance degradation. Sensor noise, varying illumination conditions and small objects with peculiar textures and visual aspects can easily produce ambiguities and misclassifications as they may be underrepresented in common internet-scale datasets used for training contrastive foundational models such as CLIP. Unsupervised domain adaptation [5] is a relevant paradigm as it mitigates domain shift by fine-tuning the model using robot data without relying on ground-truth labels or humans in the loop. Despite promising, these approaches have been rarely applied to OVSS.

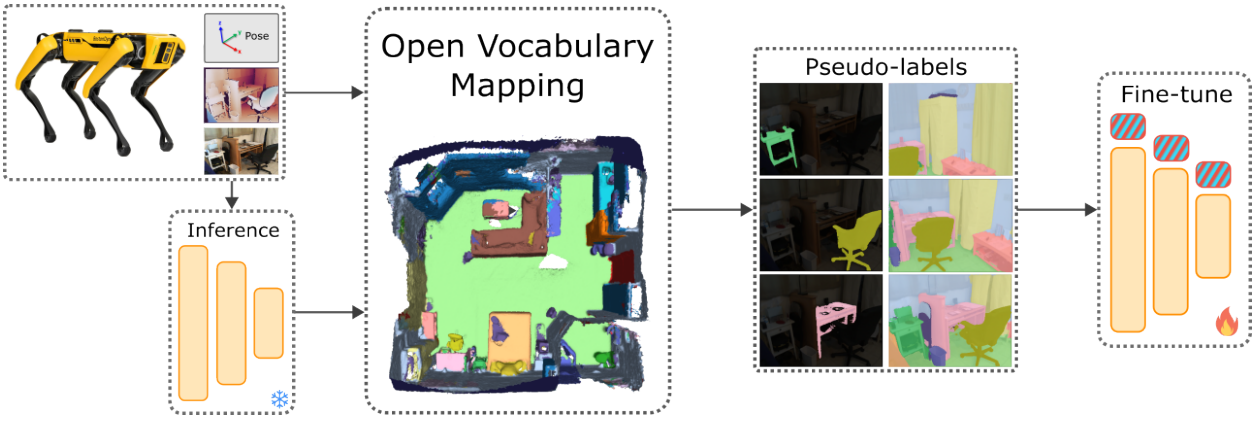

Figure 1: The high-level pipeline to develop during the project.

Goal of the project

The goal of the project is to design, implement, and validate a self-supervised domain adaptation pipeline for OVSS of robotic data (see Figure 1 for a general overview), that can be seen as an extension of [1]. From a stream of robotic perceptions (RGB images and depth data) acquired in an environment, use an open-vocabulary segmentation module to produce self-supervised pseudo-labels. In standard domain adaptation, pseudo-labels refer to labels (i.e., predictions) produced by the model, but in OVSS, they are typically represented as high-dimensional feature vectors. Due to noise and inaccuracies, these pseudo-labels cannot be directly used for self-supervision. To improve their quality, the pseudo-labels can be spatially aggregated using an existing 3D open-vocabulary mapping framework (FUS3DMaps [3]) to obtain an accurate instance-level semantic representation of the environment. The last step consists of querying the map (i.e., back-project 3D views into the image space) to obtain more reliable pseudo-labels to fine-tune the model for performance improvement.

The schedule of the activities can be roughly organized as follows (ongoing adjustments are possible):

- Literature review of UDA for semantic segmentation
- Download and setup the datasets to use for the development (e.g., Scannet++ [7])
- Run an OVSS model (such as SED [6]) to measure performance to have a first baseline
- Identify the main sources of errors and limitations of the baseline to improve during the project development
- Setup FUS3DMaps [3] to obtain 3D semantic maps from robot perceptions and identify what kind of supervision signal we want to obtain from the map (object categories, feature embeddings, dense category masks, etc, ...)
- Modify, improve, and extend the mapping framework to query the map and obtain the supervision signal (ray-tracing algorithm)
- Obtain the pseudo-labels, fine-tune the model, and quantify the performance improvement

Additional information

Starting date: ASAP

What we expect: good programming skills (C++ and Python), proven experience with deep-learning frameworks (PyTorch), good problem-solving skills

What you will gain: deep knowledge of domain adaptation techniques, practical experience in working and extending a state-of-the-art mapping framework for robotics, possibility to use real robotic platforms, opportunity to publish obtained results in a scientific paper

Contacts: Michele Antonazzi micant@kth.se, Timon Homberger timonh@kth.se

References

[1] Michele Antonazzi, Lorenzo Signorelli, Matteo Luperto, and Nicola Basilico. Instance-guided unsupervised domain adaptation for robotic semantic segmentation. In *Proceedings of the 2026 IEEE International Conference on Robotics and Automation (ICRA) (to appear)*, 2026.

[2] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4123, June 2024.

[3] Timon Homberger, Finn Lukas Busch, Jesús Gerardo Ortega Peimbert, Quantao Yang, and Olov Andersson. Fus3dmaps: Scalable and accurate open-vocabulary semantic mapping by 3d fusion of voxel- and instance-level layers, 2026.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 8748–8763, 2021.

[5] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Nico M Schmidt, Hanno Gottschalk, Tim Fingscheidt, et al. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11:54296–54336, 2023.

[6] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3426–3436, June 2024.

[7] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12–22, October 2023.