Refinement of Pseudo-Labels in Unsupervised Domain Adaptation for Semantic Segmentation Using SAM

Summary

In the field of visual perception for robotics, enabling a robot to understand complex indoor environments remains a critical challenge.

Semantic segmentation, one of the most fundamental tasks in visual perception, aims to assign a class label to each individual pixel of an image. This fine-grained understanding allows the system to not only detect the presence of objects but also localize them with pixel-level precision—crucial for autonomous navigation and manipulation tasks in cluttered scenes.

While modern semantic segmentation models can classify objects in an image, they often fail to generalize when deployed in new environments. This happens because the data distribution in the real world often differs from the one used during training—a problem known as *domain shift*. As a result, segmentation performance drops significantly, limiting the robot's ability perform well in unfamiliar settings.

This thesis proposes a method to improve model performance in new domains by enhancing the quality of *pseudo-labels*—automatically generated labels used for training in the absence of ground truth. The goal is to carry out *unsupervised domain adaptation (UDA)*, where a model trained in one environment can adapt to another without manual labeling. This is especially important for robots that need to operate in diverse environments, where manually annotating new data would be costly and impractical. To evaluate the impact of pseudo-label refinement, we use the pretrained DeepLabV3 weights [1] employed by Liu et al. [4], originally proposed for continual UDA through neural rendering. These weights serve as a consistent baseline for both the DeepLab segmentation predictions and the voxel-based pseudo-labels generated by the Kimera pipeline.

The proposed approach builds on an existing 3D mapping pipeline (*Continual Adaptation of Semantic Segmentation Using Complementary 2D-3D Data Representations* [3]), which integrates semantic predictions into a voxel-based 3D map. This system itself is built atop the Kimera framework [6], an open-source library for real-time metric-semantic SLAM. While voxel mapping helps maintain spatial consistency, it introduces noise and imprecision—especially when using coarse voxel sizes. To address this, a new *refinement stage* is introduced using the *Segment Anything Model (SAM2)* [5], a zero-shot model capable of high-quality image segmentation based on simple prompts.

In this work, two prompting strategies are proposed and tested: a *calculated approach* based on object centroids and label heuristics, and a *grid-based method* that samples the entire image uniformly. These prompts are used to guide SAM in generating sharper segmentation masks, which are then fused with the original pseudo-labels to improve accuracy.

Experiments are conducted on indoor scenes from the ScanNet dataset [2], which were never seen during training. Results show that the SAM-refined pseudo-labels are more accurate—especially in low-voxel mesh where the voxel-based labels are most degraded. The refined masks preserve object boundaries and reduce semantic noise, leading to better model supervision without any manual annotation.

Importantly, the system proves useful even when applied just once (UDA), not in a full continual learning setting. Still, it provides a foundation for future systems that adapt continually over time. The method shows promise for **reducing annotation costs** and enabling smarter, more flexible robots that can adapt to new homes, offices, and indoor spaces without starting from scratch.

Future work will aim to integrate this refinement method into a fully online continual training loop. Preliminary testing of this fine-tuning stage has already started and is showing good results, indicating that continual adaptation with SAM-refined labels can further boost performance in long-term deployment scenarios. Additional improvements may include uncertainty-based prompt filtering and the use of a shared backbone model to support multiple perception tasks with general and reusable features.

Bibliography

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017.

[3] Jonas Frey, Hermann Blum, Francesco Milano, Roland Siegwart, and Cesar Cadena. Continual adaptation of semantic segmentation using complementary 2d-3d data representations. *IEEE Robotics and Automation Letters*, 7(4):11665–11672, 2022.

[4] Zhizheng Liu, Francesco Milano, Jonas Frey, Roland Siegwart, Hermann Blum, and Cesar Cadena. Unsupervised continual semantic adaptation through neural rendering, 2023.

[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[6] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.