Open Master Thesis Topic: Large Vision-Language Networks for Open-Vocabulary Robotic Manipulation

TECHNISCHE UNIVERSITÄT DARMSTADT

iROSA

June 22, 2023

Supervisors: Snehal Jauhri (Email: snehal.jauhri@tu-darmstadt.de), Ali Younes, Georgia Chalvatzaki

1 Project Description

Robots are expected to soon leave their factory/laboratory enclosures and operate autonomously in everyday unstructured environments such as households. For this, robotic grasping and manipulation are important problems to solve. Significant progress has been made in recent years due to advancements in 3D Computer Vision and Deep Reinforcement Learning. Recent work at our iROSA lab [1] has successfully utilized learned neural geometric reconstruction of scenes for performing robotic grasping. However, most grasping methods only consider the problem of 'de-cluttering' a scene i.e. picking and removing objects placed in a scene. These methods do not consider the *semantics* of the objects. They also do not consider the problem of scene understanding and finding placing locations for the objects on tables/shelves etc. Semantic information is especially important when considering real-world robotic applications where the robot needs to re-arrange objects as per a set of language instructions or human inputs.

When it comes to semantic segmentation of scenes, many sophisticated neural networks exist [2]. However, a challenge when using such methods in the real world is that the semantic classes rarely align perfectly with the language input received by the robot. For instance, a human language instruction might request a 'glass' or 'water', but the semantic classes detected might be 'cup' or 'drink'. Nevertheless, with the rise of large language and vision-language models, we now have capable segmentation models that do not directly predict semantic classes but use learned association between language queries and classes to give us 'open-vocabulary' segmentation [3]. Some models are especially powerful since they can be used with any arbitrary language query, as shown in Fig 1.

In this thesis, we aim to build on advances in 3D vision-based robot manipulation and large open-vocabulary vision models [3] to build a full pick-and-place pipeline for real-world manipulation. We also aim to find synergies between scene reconstruction and semantic segmentation to determine if knowing the object semantics can aid the reconstruction of the objects and, in turn, aid manipulation.

Very relevant to this thesis is the Open Vocabulary Mobile Manipulation (OVMM) Challenge [4], which was announced recently. The iROSA lab currently plans to participate in the challenge, and if the Master student is willing, this thesis could be a key contribution towards it. This could be a unique opportunity to directly apply the thesis work to a challenge at a top machine learning conference (NeurIPS) and compete against the best machine/robot-learning labs.

Figure 1: Recent work on open-vocabulary segmentation [3] can not only provide semantic masks for any classes but can also use any language query such as 'I am hungry' (left) or 'Sports' (right).

Figure 2: Common robotic pick-and-place scenarios in the Open-Vocabulary Mobile Manipulation Challenge [4].

2 Outline of Work Packages

Note: The following outline is solely intended to give an idea and will be adjusted depending on the project's progress and insights.
WP 1
Duration: 1 month
The student has to perform a literature review of closely related methods in vision-language models and develop a good understanding of open-vocabulary methods. The student has to start getting familiar with the code and simulation environment for picking and placing objects in a scene (sufficient existing code will be provided to help kick-start the project).
WP 2
Duration: 2-3 months
The student has to integrate semantic information into pick-and-place pipelines. The NeurIPS Open Vocabulary Challenge simulation environment `https://github.com/facebookresearch/home-robot` can be used as a testbed. Together with the iROSA team, the student will evaluate the performance of open vocabulary vision-language networks for mobile pick-and-place tasks.
WP 3
Duration: 3-4 months
The student has to explore how semantic information can aid scene reconstruction and, in turn, aid grasping. The student will develop their own 3D-vision neural network and pipeline. Furthermore, the final pipeline will be tested against comparable baselines in simulation and on a real-robot setup at the iROSA lab. The final research will be submitted to a top robotics/computer vision conference.

3 Requirements

Enthusiasm, ambition, and a curious mind go a long way. There will be ample supervision provided to help the student understand basic and advanced concepts. However, prior knowledge of deep learning, vision, and Python programming would be a big plus.

4 Contact

Snehal Jauhri (Email: snehal.jauhri@tu-darmstadt.de), Ali Younes (Email: ali.younes@tu-darmstadt.de)

References

[1] S. Jauhri, I. Lunawat, and G. Chalvatzaki, "Learning any-view 6dof robotic grasping in cluttered scenes via neural surface rendering," 2023. [Online]. Available: https://sites.google.com/view/neugraspnet

[2] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[3] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *CVPR*, 2023, pp. 7061–7070. [Online]. Available: https://github.com/facebookresearch/ov-seg

[4] S. Yenamandra, A. Ramachandran, M. Khanna, K. Yadav, D. S. Chaplot, G. Chhablani, A. Clegg, T. Gervet, V. Jain, R. Partsey, R. Ramrakhya, A. Szot, T.-Y. Yang, A. Edsinger, C. Kemp, B. Shah, Z. Kira, D. Batra, R. Mottaghi, Y. Bisk, and C. Paxton, "The homerobot open vocab mobile manipulation challenge," in *Thirty-seventh Conference on Neural Information Processing Systems: Competition Track*, 2023. [Online]. Available: https://aihabitat.org/challenge/2023_homerobot_ovmm/