Spatio-Temporal AI for Long-term Robot Autonomy

Lukas Schmid

Autonomous robots are becoming a pervasive technology that has the potential to transform our everyday life and will be critical to address major societal challenges over the next decades, such as assistive, medical, home, service, and industrial robotics. To achieve this, *spatial AI* refers to a vision in the community to move toward the *holistic* and *abstract* scene understanding imposed by these tasks. Crucially, the above applications require *long-term human-centric* autonomy where robots operate efficiently and safely in environments shared with humans over extend periods of time. A central unsolved challenge is that human-centric scenes are geometrically *complex*, *semantically rich*, and *highly dynamic*. This necessitates building a *consistent understanding* of a scene through *space and time* during *real-time* robot operation, using only the *limited sensing* and *computation* available.

Research Question. How can a robot build an understanding of a dynamic and changing scene that facilitates future interactions? To address this question, I have crystallized three main themes for my research of I.) **4D Perception**; robustly building dense representations of highly dynamic and changing scenes or "*reasoning about what the robot sees*" (Fig. 1), II.) **Inference**; using these representations to predict probable future states for efficient interaction or "*reasoning about what the robot didn't see*" (Fig. 2), and III.) **Active Perception**; leveraging the embodiment of autonomous robots to gather the data most useful for perception, inference, and learning or "*reasoning about what the robot should see*" (Fig. 3). As an overarching theme, all these directions fruitfully interact to create highly adaptive autonomy that specializes and self-improves over time, which I call *Spatio-Temporal AI*.

I. 4D Perception

In dynamic scenes, it is essential to detect and represent both *short-term* dynamics, *i.e.,* motion within view of the sensor, and *long-term* dynamics, *i.e.,* changes outside the view of the robot. To detect short-term dynamic objects, prominent approaches leverage learned appearance features [27, 2, 18, 16], or map-based post-sequence processing [11, 18, 3]. However, appearance-based methods often struggle in unstructured and out-of-distribution scenes, whereas offline methods are not applicable during robot operation. Similarly, the problem of handling long-term scene changes is commonly addressed via multi-session change detection [5, 10, 37]. Only recently, first online methods have emerged [31, 6, 19, 20].

During my PhD, I developed the first of these online dense perception methods [31] consistent w.r.t. *long-term* dynamics. Further, I invented a novel algorithm to reconstruct *short-term* dynamic scenes [33] using the incrementally built map as motion cue, outperforming appearance-based methods *trained on the target domain* [18], generalizing better than methods trained on similar domains [2, 27, 16], and even approaching the performance of offline methods with complete hindsight [18]. These methodologies and software are widely used in both academia and industry, for example, [33] has recently been integrated by NVIDIA into their spatial AI stack. However, I realized that addressing each kind of dynamics separately is a notably easier problem. During my postdoc, I have developed a probabilistic framework that for the first time unifies short and long-term dynamics (Fig. 1), laying the foundations for *spatio-temporal metric-semantic* robot perception. The resulting framework is the first of its kind and has already been well adopted by the community. Finally, to extend semantic reasoning with the advent of *foundation models*, I have developed an information-theoretic foundation of how to compactly extract the *useful*, *i.e., task-relevant*, information out of the virtually infinite data captured by a vision-language model [13], for the first time enabling the construction of open-set 3D scene graphs in *real-time* at the same or higher fidelity than existing methods that took 6h to process the same scene [8]. My line of work on 4D perception has been recognized with an *Outstanding Systems Paper Award*, featured as a *spotlight article* on the landing page of MIT, and [31, 34, 13] were listed as pioneering works in a recent survey [15].

However, representing scenes through *space and time* has the central limitation of accumulating an ever-growing map and poor scaling. To achieve truly *life-long robot operation*, I currently research new marginalization strategies based on hierarchical optimization to keep scaling bounded. Further work will develop novel approaches for object instance re-localization combining techniques from language and descriptor learning [35, 9] with experience stored in the 4D map. This will enable capturing *instance histories* for detailed object and agent-centric reasoning. These works will be the basis for consecutive projects on *multi-session* and *multi-robot 4D perception*, leveraging the additional temporal information for map matching and optimization. As a result, this will enable the deployment of spatio-temporal AI systems in dynamic real-world scenarios and will for the first time allow for a detailed understanding of the evolution of a scene in real-time.

II. Inference

Beyond understanding the *past and present* of a scene, predicting its future is essential for effective interaction. This has been widely studied in the context of human trajectory prediction [23, 25, 24, 14], where most methods focus on collision avoidance with typical prediction horizons of $\sim$5s [23, 25, 24]. However, when considering longer prediction horizons of up to 60s, people start to interact with their

Lukas Schmid is with the MIT SPARK Lab, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. Email: lschmid@mit.edu

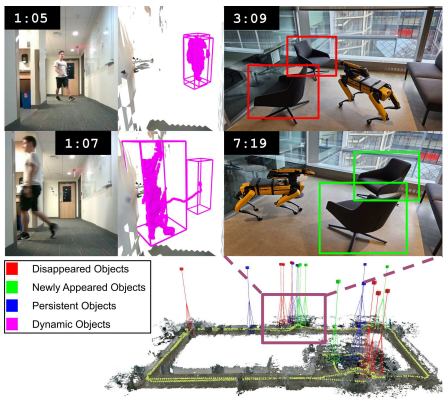

Fig. 1: **4D Perception:** Joint reconstruction of static, moving, and changing objects [34].

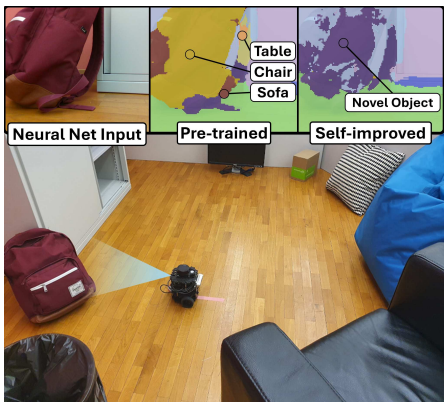

Fig. 2: **Continuous Adaptation:** Active self-improvement of semantic segmentation [39].

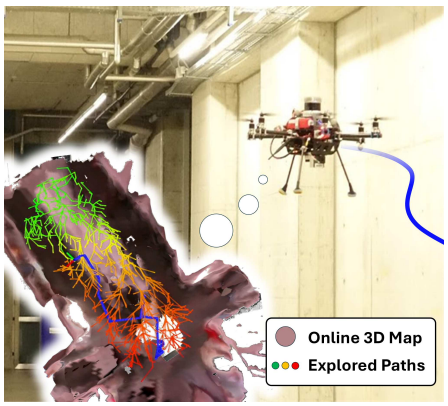

Fig. 3: **Active Sensing:** Mapping of unknown scenes on-board aerial robots [28].

environment, leading to highly complex non-linear trajectories and rendering purely geometric scene representations such as occupancy maps [25, 24, 14] insufficient.

To overcome this, I have developed a novel algorithm leveraging the rich semantic information of the previously introduced 3D scene graphs to reason about multi-modal *sequences of interactions* and then physically ground these in a spatio-temporal distribution over future positions of the person [7]. This enables prediction of future trajectories of up to 60s where people may interact with the scene, and achieved a 54% lower negative log-likelihood (NLL) compared to existing methods [24, 14]. I have further generalized this to objects by formalizing the *long-term semantic scene change prediction* problem [12] and showing that this can be solved (with some tricks) as a supervised learning problem. Although, due to the highly multi-modal nature of long-term predictions, the prediction accuracy is only ∼70%, I could show that the learned priors are still essential for *proactive and efficient* autonomy in dynamic scenes, speeding up an active change detection task by 66% on average [12].

To expand these crucial capabilities of prediction and adaptation in changing conditions, I am currently working on combining semantic priors, *e.g.,* from large language models [7], with *observations* gathered by the robot, *e.g.,* summarized in a map-of-dynamics [38]. This will allow zero-shot generalization to new scenes, but continually specialize as robots gather more data. A second research stream will focus on *inverse prediction*, *i.e.,* causal explanation of past states and what likely happened in-between observations. This information is essential for temporal queries and as uncertainty signal when reconstructing 4D maps, but also lends itself to extend *self-supervised adaptation* techniques by leveraging the developed 4D maps and causal explanations as self-supervision signal, *e.g.,* using approaches similar to [39] (Fig. 2).

III. Active Perception

The goal of active perception is to move the robot in order to gather the *sensor data* most useful to the task at hand, such as exploring unknown scenes. However, since each measurement changes what the robot can do and wants to see, most approaches reason only over short horizons, such as the next (few) view(s) [4, 26]. In contrast, a fundamental contribution was the development of a general algorithm for *informative path planning* (IPP) [28], where I proposed a novel formulation to optimize *any* information gain against *any* cost *globally* in large spaces (Fig. 3). The general nature of this algorithm allowed me to extend it to the first method for globally consistent volumetric exploration [29], active learning [39], and collaborative mapping for space robots [22]. Furthermore, I was able to demonstrate that techniques from *representation learning* and *3D scene completion* can speed up exploration of unknown scenes to close-to-optimal performance as if the environment was known [30], or reduce computation cost by almost an order of magnitude [32], enabling deployment on low-cost mobile hardware. In contrast to end-to-end methods such as imitation (IL) [1, 21] or reinforcement learning (RL) [36, 17], these methods [30, 32] maintain the *safety and interpretability* of classical methods for real-world deployment. Finally, to allow robots to continuously adapt, I developed an approach that autonomously gathers data of uncertain areas and utilizes the resulting map to train its perception network, for the first time demonstrating *fully autonomous self-improvement* of semantic segmentation neural networks on a real robot [39]. In addition to demonstrations on numerous aerial, legged, and wheeled robots, I released all my algorithms open-source[1], collecting thousands of stars and hundreds of forks on github and being implemented by over 50 research groups across the globe.

In the future, I will extend this to address *active monitoring* of dynamic scenes, where my presented [34, 12, 39] and proposed methods will allow for a detailed consideration of scene dynamics in planning. As my work has highlighted the importance of scene understanding for prediction [7, 12], we will further start to close the *perception-adaptation* loop by developing informative path planning algorithms that let robots observe areas of the scene most likely to improve perception and inference performance. While it seems unlikely that a set of pre-programmed capabilities will allow robots to operate in *all* relevant environments (*e.g.,* consider the large variety of homes), the proposed advances will instead equip robots with the ability to autonomously improve and adapt to their specific environment, embodiment, and human preference over time.

[1]All software and data available at schmluk.github.io/code.

References

[1] Shi Bai, Fanfei Chen, and Brendan Englot. Toward autonomous mapping and exploration for mobile robots through deep supervised learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2379–2384. IEEE, 2017.

[2] Xieyuanli Chen, Shijie Li, Benedikt Mersch, Louis Wiesmann, Jürgen Gall, Jens Behley, and Cyrill Stachniss. Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data. *IEEE Robotics and Automation Letters (RA-L)*, 6(4):6529–6536, 2021.

[3] Xieyuanli Chen, Benedikt Mersch, Lucas Nunes, Rodrigo Marcuzzi, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Automatic labeling to generate training data for online lidar-based moving object segmentation. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):6107–6114, 2022.

[4] Tung Dang, Frank Mascarich, Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Graph-based path planning for autonomous robotic exploration in subterranean environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3105–3112.

[5] Marius Fehr, Fadri Furrer, Ivan Dryanovski, Jürgen Sturm, Igor Gilitschenski, Roland Siegwart, and Cesar Cadena. TSDF-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5237–5244.

[6] Jiahui Fu, Chengyuan Lin, Yuichi Taguchi, Andrea Cohen, Yifu Zhang, Stephen Mylabathula, and John J. Leonard. PlaneSDF-based change detection for long-term dense mapping. *IEEE Robotics and Automation Letters*, 7(4):9667–9674.

[7] Nicolas Gorlo, **L. Schmid**, and Luca Carlone. Long-term human trajectory prediction using 3d dynamic scene graphs. *IEEE Robotics and Automation Letters*, 9(12):10978–10985, December 2024.

[8] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

[9] Chenguang Huang, Shengchao Yan, and Wolfram Burgard. Bye: Build your encoder with one sequence of exploration data for long-term dynamic scene understanding. *arXiv preprint arXiv:2412.02449*, 2024.

[10] Edith Langer, Timothy Patten, and Markus Vincze. Robust and efficient object change detection by combining global semantic information and local geometric verification. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8453–8460.

[11] Hyungtae Lim, Sungwon Hwang, and Hyun Myung. Erasor: Egocentric ratio of pseudo occupancy-based dynamic object removal for static 3d point cloud map building. *IEEE Robotics and Automation Letters (RA-L)*, 6(2):2272–2279, 2021.

[12] Samuel Looper, Javier Rodriguez-Puigvert, Roland Siegwart, Cesar Cadena, and **L. Schmid**. 3D VSG: Long-term semantic scene change prediction through 3d variable scene graphs. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 8179–8186, 2023.

[13] Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, **L. Schmid**, and Luca Carlone. Clio: Real-time task-driven open-set 3d scene graphs. *IEEE Robotics and Automation Letters*, 9(10):8921–8928, October 2024.

[14] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Intl. Conf. on Computer Vision (ICCV)*, pages 15233–15242, 2021.

[15] Ruben Mascaro and Margarita Chli. Scene representations for robotic spatial perception. *Annual Review of Control, Robotics, and Autonomous Systems*, 8, 2024.

[16] Benedikt Mersch, Xieyuanli Chen, Ignacio Vizzo, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Receding moving object segmentation in 3d lidar data using sparse 4d convolutions. *IEEE Robotics and Automation Letters*, 7(3):7503–7510, 2022.

[17] Farzad Niroui, Kaicheng Zhang, Zendai Kashino, and Goldie Nejat. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. *IEEE Robotics and Automation Letters (RA-L)*, 4(2):610–617, 2019.

[18] Patrick Pfreundschuh, Hubertus FC Hendrikx, Victor Reijgwart, Renaud Dubé, Roland Siegwart, and Andrei Cramariuc. Dynamic object aware lidar slam based on automatic generation of training data. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11641–11647. IEEE, 2021.

[19] Jingxing Qian, Veronica Chatrath, Jun Yang, James Servos, Angela P. Schoellig, and Steven L. Waslander. POCD: Probabilistic object-level change detection and volumetric mapping in semi-static scenes.

[20] Jingxing Qian, Veronica Chatrath, James Servos, Aaron Mavrinac, Wolfram Burgard, Steven L. Waslander, and Angela P. Schoellig. POV-SLAM: Probabilistic object-aware variational SLAM in semi-static environments. In *Robotics: Science and Systems*, 2023.

[21] Russell Reinhart, Tung Dang, Emily Hand, Christos Papachristos, and Kostas Alexis. Learning-based path planning for autonomous exploration of subterranean environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1215–1221. IEEE, 2020.

[22] Friedrich M Rockenbauer, Jaeyoung Lim, Marcus G Müller, Roland Siegwart, and **L. Schmid**. Traversing mars: Cooperative informative path planning to efficiently navigate unknown scenes. *IEEE Robotics and Automation Letters*, 10(2):1776 – 1783, December 2025.

[23] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *Intl. J. Robotics Research*, 39 (8):895–935, 2020.

[24] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conf. on Computer Vision (ECCV)*, pages 683–700, 2020.

[25] Tim Salzmann, Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley. Robots that can see: Leveraging human pose for trajectory prediction. *IEEE Robotics and Automation Letters*, 2023. doi: 10.1109/LRA.2023.3312035.

[26] Magnus Selin, Mattias Tiger, Daniel Duberg, Fredrik Heintz, and Patric Jensfelt. Efficient autonomous exploration planning of large-scale 3-d environments. 4(2):1699–1706. ISSN 2377-3766.

[27] Jiadai Sun, Yuchao Dai, Xianjing Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. Efficient spatial-temporal information fusion for lidar-based 3d moving object segmentation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11456–11463. IEEE, 2022.

[28] **L. Schmid**, Michael Pantic, Raghav Khanna, Lionel Ott, Roland Siegwart, and Juan Nieto. An Efficient Sampling-Based Method for Online Informative Path Planning in Unknown Environments. *IEEE Robotics and Automation Letters*, 5(2):1500–1507, April 2020. doi: 10.1109/LRA.2020.2969191.

[29] **L. Schmid**, Victor Reijgwart, Lionel Ott, Juan Nieto, Roland Siegwart, and Cesar Cadena. A Unified Approach for Au-

tonomous Volumetric Exploration of Large Scale Environments Under Severe Odometry Drift. *IEEE Robotics and Automation Letters*, 6(3):4504–4511, July 2021. ISSN 2377-3766. doi: 10.1109/LRA.2021.3068954.

[30] **L. Schmid**, Mansoor Nasir Cheema, Victor Reijgwart, Roland Siegwart, Federico Tombari, and Cesar Cadena. SC-Explorer: Incremental 3D scene completion for safe and efficient exploration mapping and planning. 2022.

[31] **L. Schmid**, Jeffrey Delmerico, Johannes L. Schönberger, Juan Nieto, Marc Pollefeys, Roland Siegwart, and Cesar Cadena. Panoptic Multi-TSDFs: a Flexible Representation for Online Multi-resolution Volumetric Mapping and Long-term Dynamic Scene Consistency. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 8018–8024, May 2022. doi: 10.1109/ICRA46639.2022.9811877.

[32] **L. Schmid**, Chao Ni, Yuliang Zhong, Roland Siegwart, and Olov Andersson. Fast and Compute-Efficient Sampling-Based Local Exploration Planning via Distribution Learning. *IEEE Robotics and Automation Letters*, 7(3):7810–7817, July 2022. ISSN 2377-3766. doi: 10.1109/LRA.2022.3186511.

[33] **L. Schmid**, Olov Andersson, Aurelio Sulser, Patrick Pfreundschuh, and Roland Siegwart. Dynablox: Real-time detection of diverse dynamic objects in complex environments. *IEEE Robotics and Automation Letters*, 8(10):6259–6266, October 2023.

[34] **L. Schmid**, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments. In *Robotics: Science and Systems (RSS)*, 2024. (**Outstanding Systems Paper Award**).

[35] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019.

[36] Delong Zhu, Tingguang Li, Danny Ho, Chaoqun Wang, and Max Q-H Meng. Deep reinforcement learning supervised autonomous exploration in office environments. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7548–7555. IEEE, 2018.

[37] Liyuan Zhu, Shengyu Huang, Konrad Schindler, and Iro Armeni. Living scenes: Multi-object relocalization and reconstruction in changing 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28014–28024, 2024.

[38] Yufei Zhu, Andrey Rudenko, Tomasz P Kucner, Luigi Palmieri, Kai O Arras, Achim J Lilienthal, and Martin Magnusson. Cliff-lhmp: Using spatial dynamics patterns for long-term human motion prediction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3795–3802. IEEE, 2023.

[39] René Zurbrügg, Hermann Blum, Cesar Cadena, Roland Siegwart, and **L. Schmid**. Embodied Active Domain Adaptation for Semantic Segmentation via Informative Path Planning. *IEEE Robotics and Automation Letters*, 7(4):8691–8698, October 2022. ISSN 2377-3766. doi: 10.1109/LRA.2022.3188901.