Towards Human-Level Adaptation, Reasoning, and Efficient Spatial Intelligence

Shibo Zhao, Ph.D. Candidate in Computer Science, Robotics Institute, CMU

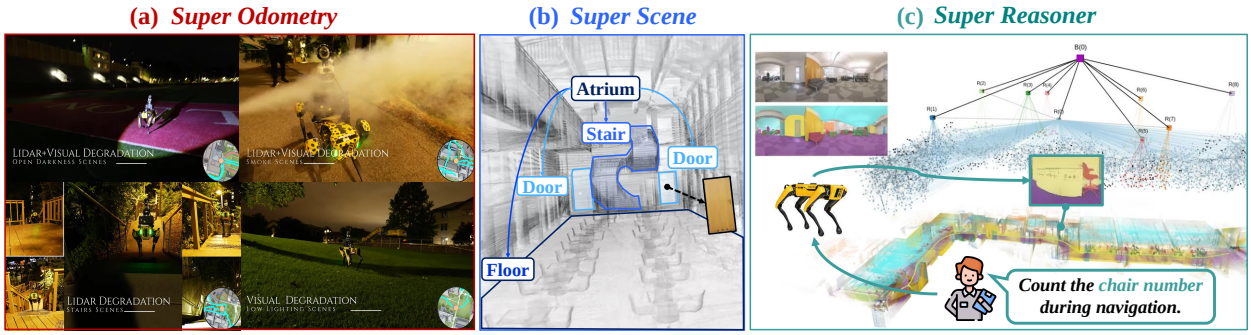


Figure 1: ***Towards Real-time, Onboard Spatial Intelligence.*** *To enable scalable spatial intelligence, we propose: (**a**) **Super Odometry**: Real-time ego-motion estimation via adaptive fusion of visual, LiDAR, and inertial sensors; (**b**) **Super Scene**: Spatial-temporal 4D scene reconstruction for spatial reasoning; (**c**) **Super Reasoner**: Physical world understanding for efficient and intelligent autonomy.*

Introduction

The development of autonomous systems capable of human-level reasoning, comprehension, and interaction within complex environments remains a significant challenge in robotics and artificial intelligence. While advancements in Large Language Models (LLMs), such as ChatGPT, and large-scale world modeling frameworks like Genie have demonstrated promising progress toward achieving Artificial General Intelligence (AGI), these approaches remain insufficient for robotics applications. Their limitations stem from an inability to address critical real-world tasks like reliable state estimation, rich environmental representation, and spatial reasoning, essential for robots to understand and interact effectively with the physical world.

To address these challenges, we believe robots need a comprehensive framework capable of reliably adapting to any environment, flexibly representing surroundings with varying levels of detail, and reasoning about physics efficiently. These capabilities are essential for achieving long-term autonomy and enabling robots to exhibit behavior approaching human-level performance. To this end, we introduce the concept of *Spatial Intelligence* (see Fig. 1), which incorporates a *human-level adaptive sensor fusion* framework for robust state estimation, a *4D scene representation* with flexible levels of detail, and *scene reasoning* to support autonomy.

Research Problems

1 **Human-Level Sensing Adaptation:** The primary challenge of current robotic systems is their inability to adapt effectively to various environments, threatening the safety of robots. Our team addresses this limitation by introducing a hierarchical adaptation framework (***Super Odometry***) to seamlessly adjust different sensing modalities, which strikes a balance between redundancy and operational efficiency.

2 **Spatio-Temporal 4D Reconstruction:** A major limitation of current scene reconstruction pipelines is that they generate a world model based on a single snapshot in time. However, robots need the ability to reason about the state of scenes at previous moments and predict how these scenes may evolve in the future. To address this challenge, our team proposes the development of spatial-temporal 4D reconstruction (***Super Scene***), which incrementally builds a world model over time. This approach not only enables detailed, photorealistic reconstructions but also facilitates abstract object-level mapping.

3 **Human-Level Scene Reasoning:** The primary objective is to discover, learn, and transfer spatial common sense by analyzing the above 4D reconstruction with symbolic (language, physical, logical, and geometrical) reasoning abilities. Our approach (***Super Reasoner***) allows robots to have an integrated perception-plan-act pipeline by combining the powerful expressiveness of 4D model with explainable and rule-based reasoning. Such an integration will help robots comprehend physical relationships in their surroundings and make informed decisions based on these learned physical and logical rules.

Proposed Spatial Intelligence Framework

Fig. 2 illustrates the *Proposed Spatial Intelligence* framework, which comprises three key modules: *Super Odometry*, *Super Scene*, and *Super Reasoner*. The *Super Odometry* module ensures robust localization across diverse environments. The robot's state is then transmitted to the *Super Scene* module, which reconstructs the surrounding environment in a temporal-spatial context. Meanwhile, the *Super Reasoner* module stores the robot's state and the 4D scene representation to facilitate understanding. Together, these modules deliver expressiveness, efficiency, and reliable representations for the perception-planning-action pipeline, enabling a wide range of robotic tasks.

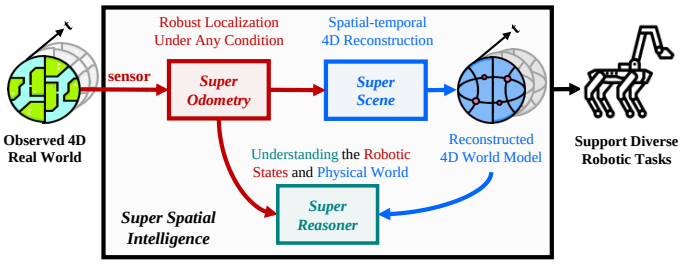


Figure 2: Pipeline overview of Spatial Intelligence.

Thrust #1: Human-level Sensing Adaptation Towards All-weather Environments. The core challenge addressed by ***Super Odometry*** is maintaining high precision in ego-motion estimation under varying environmental conditions. We propose a hierarchical adaptation framework that dynamically reconfigures sensor fusion processes to respond to challenges such as occlusions, sensor degradation, or environmental factors like low light, confined spaces, and dust. The system is dynamically reconfigurable, functioning as a multi-level scheme and each level focuses on different types of degradation. Lower levels handle basic, fast, and resource-efficient adaptation and higher levels handle complex and computationally-demanding processes. If low-level adaptation mechanisms fail, high-level mechanisms can intervene to support recovery. This hierarchical method enhances the resilience of state estimation, allowing the robot to effectively navigate through all forms of challenges. This approach ensures that ***Super Odometry*** can provide continuous, high-precision ego-motion estimation, even in the most challenging and resource-constrained environments. The summarized work is submitted to Science Robotics.

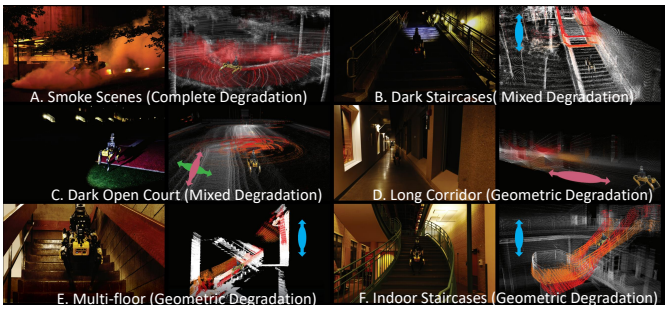
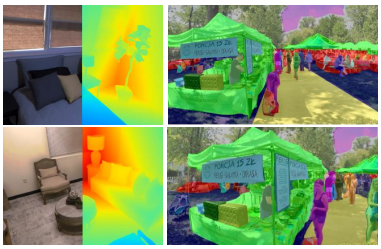


Figure 3: Our past work of robust odometry [1, 2].

Thrust #2: Online Spatial-Temporal Scene Reconstruction for 4D Environments. To achieve more adaptive and intelligent autonomy, a dynamic, spatial-temporal 4D representation is critical for autonomous robots to reason about the evolving interactions and temporal movements of objects in 3D space. Despite pioneering efforts [5] in modeling spatial dynamics, current methods fall short of meeting real-time processing requirements. To address this limitation, we focus on online 4D scene reconstruction. Building upon our advancements in real-time 3D reconstruction and object-centric semantics (see Fig. 4), we present the ***Super Scene*** module, designed to achieve metric-semantic 4D reconstruction from sensor observations and poses estimated by ***Super Odometry***. ***Super Scene*** provides an open-vocabulary, graph-structured 3D representation adaptable to varying detail levels across time. This representation supports diverse applications, from dense geometric data essential for mobility and manipulation to abstract semantic and object-level affordances critical for task planning. The system leverages 2D foundation models, integrating their outputs into 3D representations while tracking object-centric changes over time.

Future Direction #3: Large Language Model for Spatial Understanding. We propose ***Super Reasoner***, a 4D large language model specifically designed to process 3D point cloud data over time. We propose to generates structured 3D scene understanding, which includes identifying architectural elements such as walls, doors, and windows, as well as producing oriented object bounding boxes along with their corresponding semantic categories. By advancing spatial reasoning, *Super Reasoner* will significantly enhance applications in embodied robotics, autonomous navigation, and a variety of 3D scene analysis tasks.



(a) High-Quality Spatial 3D Reconstruction (b) Long-term Temporal Semantics Understanding

Figure 4: Our past work of spatial 3D reconstruction [3] and semantics understanding [4].

References

[1] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, "Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8729–8736.

[2] S. Zhao, P. Wang, H. Zhang, Z. Fang, and S. Scherer, "Tp-tio: A robust thermal-inertial odometry with deep thermalpoint," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4505–4512.

[3] X. Xu, F. Xue, S. Zhao, Y. Pan, S. Scherer, and X. Huang, "Mac-ego3d: Multi-agent gaussian consensus for real-time collaborative ego-motion and photorealistic 3d reconstruction," *arXiv preprint arXiv:2412.09723*, 2024.

[4] X. Xu, J. Wang, X. Ming, and Y. Lu, "Towards robust video object segmentation with adaptive object calibration," in *ACM International Conference on Multimedia (ACM MM)*, 2022.

[5] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 310–20 320.