Research Statement and Plan

Giseop Kim

July 22, 2024

1 Vision: Self-evolving Robot-Web Navigation Intelligence

Robots are now undeniably leaving the lab and exploding in number. Google's Waymo expanded its robotaxi service across San Francisco in 2024. According to a 2024 report from Morgan Stanley, the global number of humanoid robots is expected to increase from 40,000 in 2030 to 8 million in 2040, and to 63 million by 2050.

The rapid advancement of GPU hardware and deep learning technology over the 10 years following AlexNet in 2012 has played a significant role to excel the robot explosion. Before 2018, the invention of Transformer, the deep learning-based perception was difficult to generalize and thus challenging to apply in real-world robotics. However, recent Foundation Model (FM) research, such as Segment Anything [1] and DINOv2 [2], is alleviating these concerns. These foundation deep models provide strong prior knowledge about intelligence for robot autonomy.

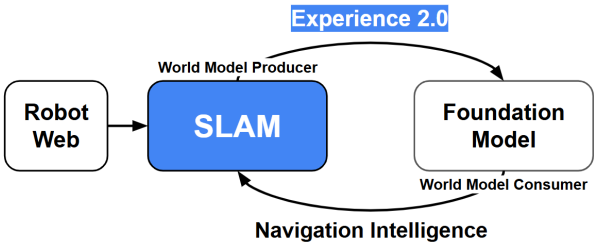


Figure 1: The mission of my research group, **Autonomy and Perceptual Robotics Lab (APRL)**. We achieve a positive feedback loop by producing Experience 2.0 (Def. 3) through advanced SLAM from the robot web, contributing to the improvement of foundation models for perception, and consequently enhancing the navigation intelligence of robots (RS. 3).

The recent advancements in **Simultaneous Localization and Mapping (SLAM)** [3] are another key factor enabling robot autonomy in the real world. Because, SLAM serves as a producer of a world model data while a deep FM is a data consumer that learns that world model. SLAM is a reconstruction machine [4] that transforms raw and noisy heterogeneous sensor data into a consistent world model as in Fig. 2. SLAM is often technically defined as either the co-optimization of robot pose and map state [3], or as a combination of odometry [5] and place recognition [6]. But here, I propose to define the term SLAM at a higher level, as

> **Definition 1.** *SLAM* := An Automated Experience Reconstruction Machine

Therefore, it is natural that we should also provide a definition of the term experience. Traditionally, until around 2020, SLAM research has primarily been limited to the following definition:

> **Definition 2.** (As is) *Experience 1.0* := A globally-aligned spatial model of the world, from a single robot's single journey.

As a result, in the SLAM academic community, research has primarily focused on improving efficiency and accuracy from the perspective of 'mapping', rather than on management and expansion of the notion of the experience.

My research group, **Autonomy and Perceptual Robotics Lab (APRL)**, aims to extend the current definition of Def. 2 to Def. 3. This is defined as a series of challenging subproblems of the larger question of how data, produced by

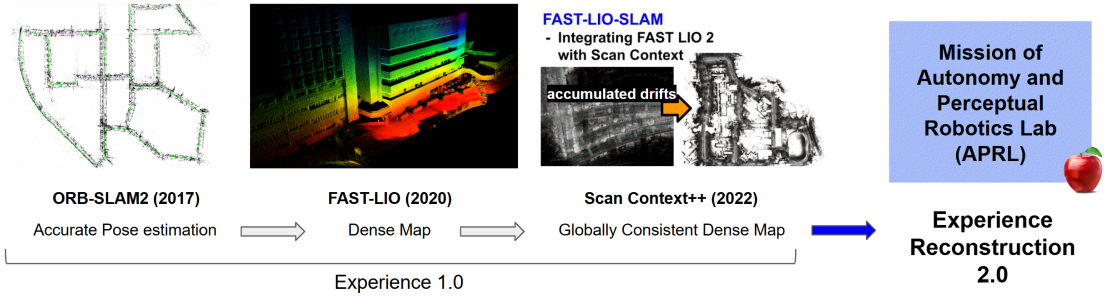


Figure 2: Traditional SLAM has evolved towards improving the accuracy of dense maps (we call it Experience 1.0 era). My research group aims to develop the next generation of SLAM to produce and reconstruct the Experience 2.0 mentioned in Def. 3.

various forms of robots from different manufacturers, operating in diverse environments, and generated at different times asynchronously (and even their initials are unknown), can be fused together **to enhance the knowledge required for robots to understand the world better**. We call this **Experience 2.0** to distinguish it from the existing limited single agent's spatial-only map reconstruction. This acts as a mission that unifies the research conducted in my lab. As shown in Fig. 2, my lab aims to carry out research on robot perception and navigation that contributes to Experience 2.0 of Def. 3, of course including the traditional Experience 1.0.

Definition 3.

(To be) *Experience 2.0* := A globally-aligned spatial
+ and feature-semantic model of the world
+ reconstructed from heterogeneous sensor measurements
+ collected by numerous heterogeneous robots'
+ during multiple asynchronous journeys over time,
+ despite unknown initial conditions and inherent noise characteristics,
+ leveraging foundation priors about the world and sensors.

In summary, my labs's first research statement is as follows:

Research Statement 1. Expanding the definition of SLAM, **the robot web collaborates to build and manage a comprehensive world model, named Experience 2.0** ,

where details about the key subproblems to achieve this will be provided in the Future Research (Sec. 2). Recently, the term 'Robot Web' has been used to refer to distributed optimization for multiple robots [7]. We aim to expand this definition to encompass a variety of heterogeneous robots performing their respective roles in diverse environments (e.g., indoors, city outdoors, unstructured forests, construction sites, etc.). These robots can take various forms, such as mobile phones, VR/AR equipment, wheeled robots, drones, or humanoids.

My research group's second research objective is to consider the purpose behind the generation of Experience 2.0. I believe that the foundation of modern numerical computational engineering follows the Bayesian equation below.

Definition 4. Posterior = Prior × Likelihood

This is not merely a definition but a philosophy [8, 9]. Supporting information for current decision-making, or posterior, is derived through a weighted sum of the prior knowledge up to the previous moment and current sensor measurements, or likelihood. Here, the prior can be exemplified by the physical laws of a known environment (e.g., gravity constant is 9.806, state-space models of Inertial Measurement Unit (IMU) kinematics) or the recent advancements in foundation models [10, 11]. This can be rephrased as follows:

Definition 5. A robot agent's navigation intelligence

= Compile-time Prior knowledge about the human, other robots, and the world up to time $t-1$
× Run-time sensor measurements at time t

In short, a robot uses prior knowledge generated by other robots (the Robot Web) and fuses it with its own runtime on-board measurements to reconstruct the Experience 2.0 or assist in decision-making. I can refer to this as a Bayesian update for the robot's navigation intelligence, as illustrated in Fig. 3.
As shown in this example, both prior and likelihood have their respective strengths and weaknesses. Def. 5 can also be seen as bridging the gap between the modern computer vision (-only) community and the conventional robotics (including navigation) community. By fusing these two approaches, we aim to continuously enhance a method for reconstructing Experience 2.0. Thus, my lab's first objective, RS. 1, is expanded to this second statement as follows:

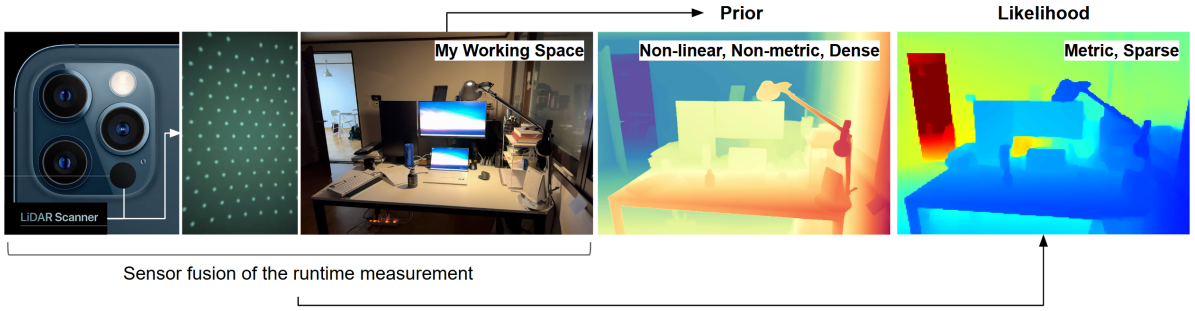


Figure 3: A single RGB frame and corersponding sparse depth image are captured by iPhone 15 Pro. The prior is generated from the latest foundation model, DepthAnything [11]. This model provides high-resolution but nonlinear data with an unknown metric scale. The likelihood, on the other hand, is a low-resolution metric depth image obtained by fusing RGB images with sparse LiDAR depth measurements. For a robot to understand the scene accurately as well as dealing with uncertainty, it is crucial to effectively combine these two sources of knowledge.

Research Statement 2. Expanding the definition of SLAM, the robot web collaborates to build and manage a comprehensive world model, **enabling continual Bayesian updates of robot intelligence**.

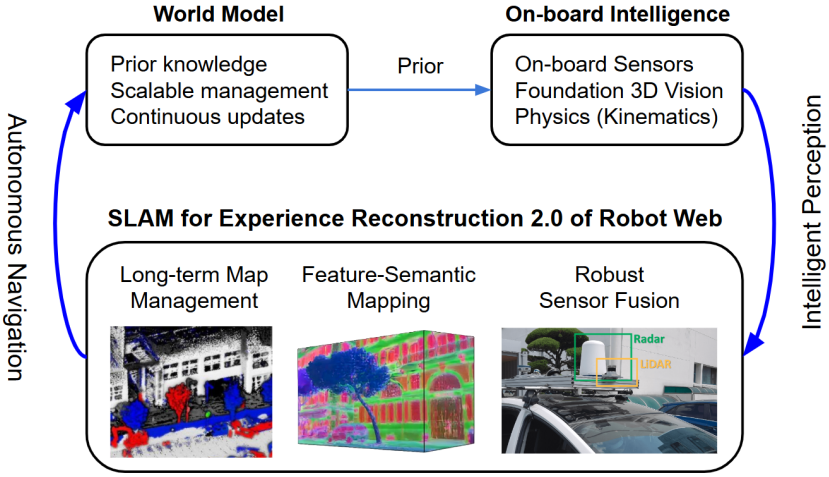


Figure 4: The iterative process of self-evolving robot-web navigation intelligence and the pipeline for its construction. The visualizations of the robust sensor fusion, feature-semantic mapping, and long-term map management are from [12], [13], [14], respectively.

The beauty of the Bayesian philosophy lies in the fact that the posterior at the current time t then serves as the prior for the next timepoint $t+1$. Therefore, I believe that achieving RS. 2 represents leads to the acceleration of creating a reconstruction machine of Experience 2.0. I refer to this as self-evolving robot-web navigation intelligence:

Definition 6. *Self-evolving Robot-Web Navigation Intelligence*

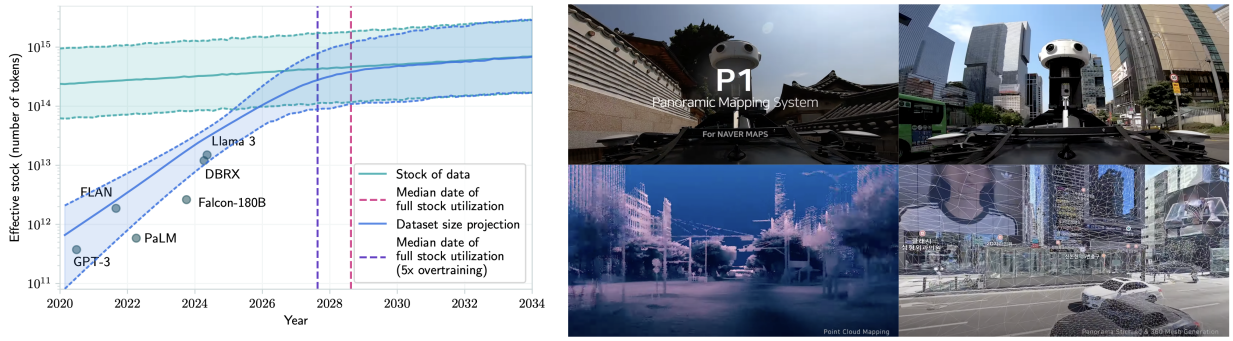
:= A distributed and resilient SLAM pipeline to generate the Experience 2.0 aforementioned
+ ensuring the continual updating of the robots' prior knowledge of the world,
+ which iteratively enhances next time's robot navigation and its experience reconstructions.

Therefore, ultimately, building this is the final goal of my research group.

Research Statement 3. In turn, we contributes to the development of a **self-evolving robot-web navigation intelligence pipeline** by enabling the reconstruction of the Experience 2.0, aimed at creating better cities and improving human lives.

As robots gain a better understanding of cities, they can contribute to creating better cities. In turn, the process of building better cities enhances the robots' understanding of them, creating a positive feedback loop that continues to repeat. This process is visualized in Fig. 4. The reason this is possible is that, unlike the concerns about human-generated text data

depletion in Fig. 5a, the diversity and sustainability of sensor data acquisition through which a robot perceives the real-world have no limitations. For example, Fig. 5b shows a project I participated in during my time in the industry. The new generation of street-view vehicle project (P1, Panoramic Mapping System) at NAVER LABS continuously records various spatio-temporal events nationwide. SLAM for building Experience 2.0 will function as such an unlimited diverse world model producer and will also reinforce itself thanks to the future works of our lab.



(a) Concerns about text data depletion [15].

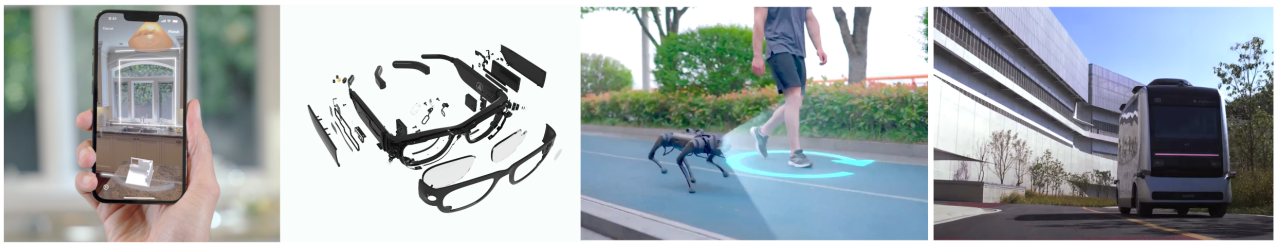
(b) The P1 project at NAVER LABS.

Figure 5: (a) Concerns about the potential depletion of human-generated text data for learning models (source: image courtesy of [15]). (b) The P1 project, in which I am involved, is a new panoramic robotics sensor system for NAVER MAPS developed by NAVER LABS. Using data from multiple sensors, it successfully reconstructs 3D world models (e.g., point clouds, panoramic meshes). Robots gathering sensor data in the real world are unlikely to encounter concerns about data depletion.

2 Future Research: Towards Self-evolving Navigation Intelligence

To summarize the future works of our lab in a single mission statement again: We aim to innovate SLAM technology and combine it with foundation models to iteratively build *self-evolving navigation intelligence* by reconstructing Experience 2.0. To achieve this, we need to innovate in three key areas.

1. Input: 3D Vision Data Explosion and Opportunities. As mentioned in Fig. 5, there is no limit to the amount of 3D data obtained by robots interacting with the real world using real sensors. The variety and number of robots are rapidly increasing, and sensors are becoming very affordable. For example, 10 years ago, the Velodyne 64 ray LiDAR cost over 100 million KRW, but now the Livox Mid-360 LiDAR, which can obtain a denser point cloud, costs less than 1 million KRW. In other words, it has become 100 times cheaper. 3D data will also explode in non-autonomous vehicle areas. For example, as shown in Fig. 6, a lot of data is obtained from everyday devices such as iPhones and Smart eyeglasses (e.g., Project Aria of Meta), which will be used to reconstruct our surrounding environments and serve as ingredients for robots to perceive their environments. Therefore, the quantity and diversity of 3D data will explode in all domains. I call it the **Robot Web era** inspired by [7].



Apple iPhone LiDAR RoomPlan | Project Aria from Meta | Unitree Go1 Legged Robot | NAVER LABS Unmanned Shuttle (ALT-B)

Figure 6: In the near future, there will be an explosive generation of robot sensor data from various platforms and diverse situations.

Our research group aims to predict the potential problems that will arise from this fact and seek research topics that can enhance robot navigation intelligence using this data to achieve the feedback loop in Fig. 4 effectively and efficiently.

2. Algorithm: Continual World Model Optimization with Unknown Initials. The second future work is to technically build Experience 2.0 itself using the aforementioned data. This may require overcoming practical issues related to computation, such as observability, solution convergence, high-performance computing, etc. These challenges are primarily due to unknown correspondences and unknown initial solutions arising from various robots collecting data at different times in diverse physical environments. We also cannot know in advance where and to what degree there are imperfections (e.g., timestamp inversion, missing data, unsynced sensors) in all or part of the sensor data. In LT-mapper [14] (ICRA 2022), as a starting point for addressing these issues, we spatially aligned multi-session trajectories with unknown prior global alignment into a common coordinate system, as shown on the left side of Fig. 7. Furthermore, in [16] (IJRR 2024), we also released a dataset that promotes place recognition among heterogeneous LiDAR sensors as well as multi-session scenarios.

Even after overcoming the challenges of unknown data distribution and imperfections, it is still necessary to define what is the thing the Robot Web cooperatively contributes to (e.g., world model). The right image in Fig. 7 illustrates an example

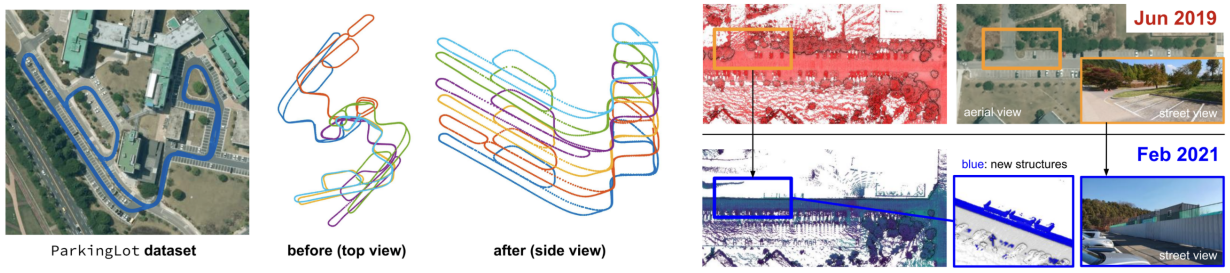


Figure 7: Left: Unknown Initials in real-world multi-robot applications, Right: An example of change detection and update when the world model is defined as 3D point cloud.

of updating the world model using the most primitive form, 3D points. Recent trends emphasizes new directions in research topics through neural implicit representation, shedding light on emerging forms of world representation. However, there is still limited number of discussion on their long-term management and continual updates. Therefore, our research group aims to explore neural representation when it comes to Experience 2.0.

3. Application: Robot As Infrastructure with Self-evolving Navigation Intelligence. Finally, I propose a new concept called Robot As Infrastructure (RAI). The self-evolving navigation intelligence mentioned in the first section (Sec. 1) is a kind of tool. What is the purpose of that tool? What does our research group aim to contribute towards based on that navigation intelligence made by Experience 2.0 reconstruction machine? RAI is the answer. I believe that robots would not just be limited to personal properties or household appliances. The true value of robots lies in their ability to perform tasks that are difficult for humans (e.g., delivering heavy objects, night patrols across large areas in a city) or impossible for humans to do (e.g., exploring areas with radiation leaks, space exploration). To achieve this, the investment and operation of robots need to occur at the city or national level rather than by individuals. We call such things infrastructure.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[3] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016.

[4] Cyrill Stachniss. *Robotic mapping and exploration*, volume 55. Springer, 2009.

[5] Wei Xu and Fu Zhang. Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter. *IEEE Robotics and Automation Letters*, 6(2):3317–3324, 2021.

[6] Giseop Kim, Sunwook Choi, and Ayoung Kim. Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments. *IEEE Transactions on Robotics*, 38(3):1856–1874, 2022.

[7] Riku Murai, Joseph Ortiz, Sajad Saeedi, Paul HJ Kelly, and Andrew J Davison. A robot web for distributed many-device localisation. *IEEE Transactions on Robotics*, 2023.

[8] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[9] Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*. Cambridge university press, 2013.

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[12] Giseop Kim, Yeong Sang Park, Younghun Cho, Jinyong Jeong, and Ayoung Kim. Mulran: Multimodal range dataset for urban place recognition. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 6246–6253. IEEE, 2020.

[13] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024.

[14] Giseop Kim and Ayoung Kim. Lt-mapper: A modular framework for lidar-based lifelong mapping. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7995–8002. IEEE, 2022.

[15] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? Limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

[16] Minwoo Jung, Wooseong Yang, Dongjae Lee, Hyeonjae Gil, Giseop Kim, and Ayoung Kim. Helipr: Heterogeneous lidar dataset for inter-lidar place recognition under spatiotemporal variations. *The International Journal of Robotics Research*, 2023.