Robust Localization based on Multi-Modal Implicit Maps

Proposed by Junqiao Zhao

Tongji University
2014.12

Abstract

Autonomous localization serves as the foundational cornerstone for intelligent decision-making in unmanned systems. Addressing the limitations of existing methods—where SLAM-based explicit maps suffer from high multi-modal feature storage costs, difficulties in dynamic updates, and poor robustness, while place recognition approaches struggle with long-term environmental evolution in urban-scale scenarios due to unstructured and highly redundant sample storage—this proposal introduces a novel paradigm of "Robust Localization Based on Multi-Modal Implicit Maps." The research encompasses: 1) Constructing an implicit localization map model integrating long-term memory and working memory mechanisms, proposing hierarchical geospatial grid-based memory storage and freshness-driven update/forgetting strategies to achieve structured management of location memories; 2) Developing a local-global decoupled cross-modal data alignment framework coupled with anti-forgetting continuous learning mechanisms to enable synergistic updates between map encoders and implicit maps; 3) Establishing an uncertainty-driven fast-slow system inference framework that dynamically switches between rapid place retrieval and refined localization through local feature similarity and spatiotemporal consistency, balancing precision and efficiency. This research aims to build learnable, updatable, and inferable implicit localization maps with corresponding methodologies, providing theoretical and technical foundations for national strategic initiatives in autonomous unmanned systems.

Research Statement

Localization is the foundation for autonomous decision-making in intelligent systems such as autonomous driving, unmanned aerial vehicles (UAVs), and robotics (Liu & Luo, 2022). Its reliability directly determines the task execution capability—and even the survival ability—of unmanned systems operating in complex environments. With the rapid development of smart cities, the low-altitude economy, and strategies for autonomous unmanned systems in China, achieving long-term robust localization in large-scale urban environments has become a critical challenge for the wide deployment of such systems (Li, 2024).

Localization based on Global Navigation Satellite Systems (GNSS) has been extensively applied and is relatively mature. However, its performance is constrained by the availability and stability of satellite and differential signals, which makes it unreliable in dense urban areas or indoor environments. Although GNSS can be complemented with inertial measurement units and odometry, these approaches still struggle to provide long-term reliability in the presence of signal loss or interference. Consequently, most unmanned systems today rely on Simultaneous Localization and Mapping (SLAM) to autonomously build maps from sensory data and perform localization. While SLAM-based explicit maps represent the mainstream approach to autonomous localization, they inherently suffer from high cross-modal storage costs, difficulty in dynamic updates, and limited robustness, making them unsuitable for kilometer-scale urban scenarios and long-term deployment across seasonal cycles (Figure 1, left). Addressing the contradiction between "storage efficiency in large-scale environments" and "adaptability to long-term environmental changes," and establishing a new paradigm of intelligent localization that is robust, accurate, and evolvable (Liu & Luo, 2022), remains an urgent scientific problem across multiple fields including unmanned systems, navigation, surveying and mapping, and artificial intelligence.

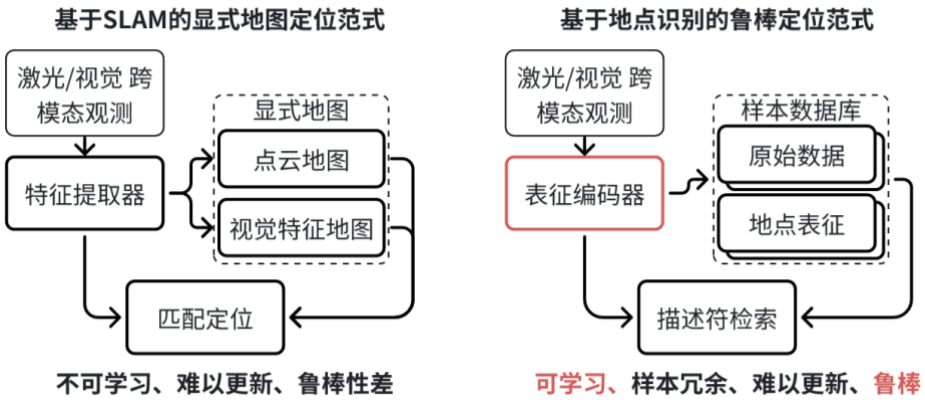

Figure 1 SLAM-based mapping and localization (left)； Robust localization based on place recognition. (right)

Recent advances in deep learning–based place recognition have opened promising avenues to address these challenges. By leveraging large-scale data to train neural networks that extract discriminative representations of places, such methods enable robust localization across different times of day and weather conditions (Barros et al., 2022; Shi et al., 2023; Xie et al., 2024; Yin et al., 2024). However, current methods generally adopt a "mapless" architecture, relying solely on similarity between query and historical data representations for retrieval (Figure 1, right). This leads to several limitations: (1) the unstructured storage of historical data results in redundant place descriptors for repeatedly captured locations, leading to complex management and inefficient retrieval; (2) existing methods lack adaptive update mechanisms in dynamic environments. Seasonal changes or roadworks cause distributional drift in place representations, requiring full retraining of encoders, which prevents incremental

adaptation; (3) there is a trade-off between localization efficiency and accuracy. At city scale, place descriptors can reach billions, and the difficulty of recognition varies across scenes, making it hard for existing methods to balance computational efficiency with recognition accuracy. These limitations hinder the ability of current place recognition approaches to meet the requirements of robust, continuous, and efficient localization at urban scale.

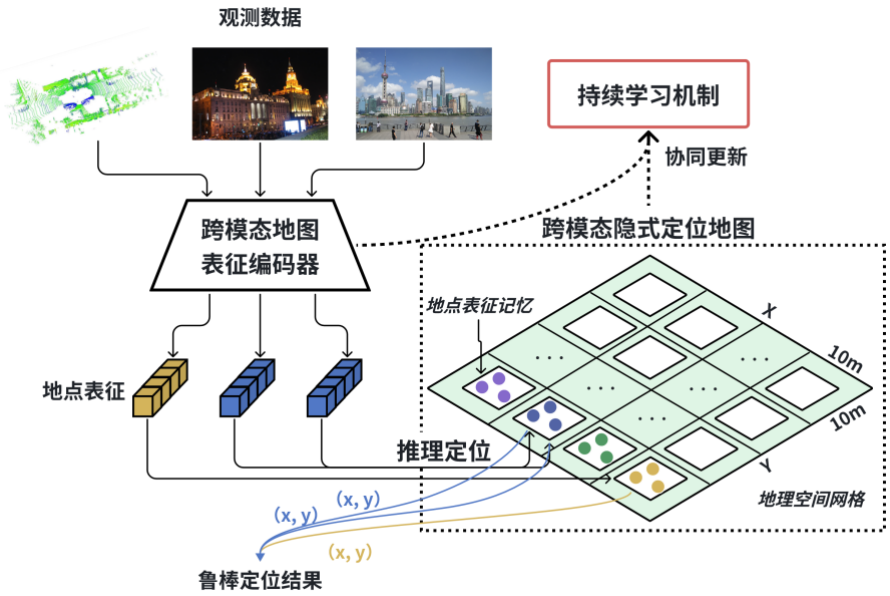

Figure 2 The proposed robust localization based on cross-modal implicit map

To overcome these challenges, the project proposes a new paradigm of robust localization based on cross-modal implicit maps (Figure 2). The central innovation lies in building a learnable, updatable, and inferable implicit map model, and developing robust localization methods upon it.

To address the issue of structured place representation management, we propose a neural implicit map model that embeds geospatial units with implicit representations. The urban space is hierarchically partitioned into grid cells, each embedding a unified and compressed cross-modal representation of LiDAR and visual data, forming retrievable implicit memory nodes. A streaming working memory pool is further employed to dynamically cache real-time observations, which, combined with freshness-based evaluation of representations, enables incremental map updates and forgetting of redundant memories—effectively breaking the "storage-update" bottleneck of explicit feature maps.

To tackle the challenges of cross-modal representation construction and continual updates, we design a decoupled local-global representation alignment mechanism and a continual learning strategy that integrates memory replay and knowledge distillation. Local and global feature alignment enforces consistency between visual and LiDAR modalities across multiple scales. In parallel, memory replay of implicit map

representations and knowledge distillation suppress catastrophic forgetting, enabling both the encoder and the map to adapt to evolving environments through collaborative updates.

For robust localization in city-scale scenarios, we propose a fast-slow system for uncertainty-driven cooperative inference. Based on the uncertainty of place representations, the system adaptively switches between fast and slow modes: the fast system retrieves candidate locations rapidly from implicit memory for easily recognizable scenes, while the slow system performs fine-grained ranking and verification for ambiguous scenes using local feature similarity and spatiotemporal consistency. Furthermore, lightweight strategies for both the encoder and map representations are introduced to optimize computational cost and inference efficiency while preserving localization accuracy, achieving a balanced trade-off among precision, cost, and efficiency.