The Application of Semantic-enhanced 3D Laser SLAM in Mobile Robots

Ruizhi Feng

College of Automation Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, 210016, China

ruizhi_feng@nuaa.edu.cn

Abstract. For indoor environments with low Global Navigation Satellite System (GNSS) and dynamic changes, such as hospitals and shopping malls, traditional 3D LiDAR Simultaneous Localization and Mapping (SLAM) is prone to mismatch, false loops, and map drift under repeated topology, occlusion, and dynamic interference. Introducing semantic priors into the front-end, back-end, and loop closure detection of SLAM can significantly improve the robustness of localization and the interpretability of the task. This article outlines the workflow of semantic augmented SLAM, starting with ordinary 3D laser SLAM. It summarizes the architecture of front-end odometry, back-end optimization, loop closure, and map, and concludes the front-end dynamic point culling and class weighting, back-end semantic consistency, loop closure, and instance anchoring. It also introduces commonly used semantic augmentation methods, such as SuMa++ and dynamic-static object recognition, as well as their roles in motion segmentation and mapping. Based on service robot scenarios such as hospital corridors, supermarket restocking, and warehouse handling, this article summarizes common evaluation and practical experience related to localization and semantics, providing a reference for the application of semantic information in practical SLAM systems.

Keywords: Semantic augmented SLAM, 3D LiDAR; mobile service robot; dynamic scene.

1. Introduction

Mobile robots play a key role in areas such as hospital drug delivery, warehouse cargo handling, and public facility inspection. Sales of mobile service robots for professional purposes reached nearly 200,000 units, an increase of nearly 9%, with the key driver of sales growth being the shortage of employees [1]. However, mobile robots mostly operate in indoor environments with low GNSS, repetitive environmental topology, and dense pedestrian traffic. These factors limit the robustness and global consistency of 3D laser SLAM, which is highly dependent on geometric and illumination intensity information. By introducing semantic priors, the system can distinguish between movable and immovable objects at the front end and suppress motion interference. At the back end, semantic consistency is used to improve the quality of loop closure checks and relocation. Labeled elements are built on the map side to support path planning and target retrieval, thereby improving the accuracy of positioning in complex indoor environments and shortening the mapping time. Semantic augmentation further provides interpretable information to the task layer, enabling robots to provide more reliable navigation and obstacle avoidance in complex indoor spaces and improving long-term maintainability.

With the rapid advancements in scene understanding using machine learning and convolutional neural networks (CNNs), much research on semantic SLAM technology tends to favor methods such as camera + IMU and RGB-D for collecting semantic information. 3D lasers have advantages such as high geometric recognition intensity and reliable loop closure, and have also been the subject of extensive research. SuMa++ introduces semantic segmentation on the basis of SuMa surface meta-mapping, performs spherical projection on LiDar point cloud, uses this information to filter dynamic objects, and adds semantic constraints in ICP to improve the robustness and accuracy of pose estimation [2]. Moving Object Segmentation (MOS) calculates cross-frame residual maps by obtaining distance maps from continuous scanning, and uses the temporal features of CNNs to directly distinguish dynamic objects from static backgrounds, thus obtaining clean maps in dynamic environments. This article outlines the workflow of semantically enhanced 3D laser SLAM,

introduces commonly used semantic enhancement methods such as SuMa++, MOS, and neural implicit semantic mapping, and discusses the application prospects of semantically enhanced 3D laser SLAM in specific real-world scenarios.

2. 3D Laser SLAM

Both laser SLAM and visual SLAM largely follow a hierarchical architecture of front-end odometry and back-end optimization. Fig. 1 shows a commonly used working framework for SLAM. For 3D laser SLAM, the front end uses LiDAR as the sensor to continuously collect cloud point maps of the environment. In the front-end stage, IMU is used to assist in data preprocessing such as noise removal, distortion removal, de-skew, and time synchronization. Loop closure detection is used to determine whether the robot has reached a previous position, eliminate accumulated error offsets, and improve global consistency. The backend receives information from the frontend regarding loop closure detection and performs optimization using methods such as Kalman filtering and histogram filtering to determine its own position. After completing these steps, a corresponding map is built based on the estimated trajectory.

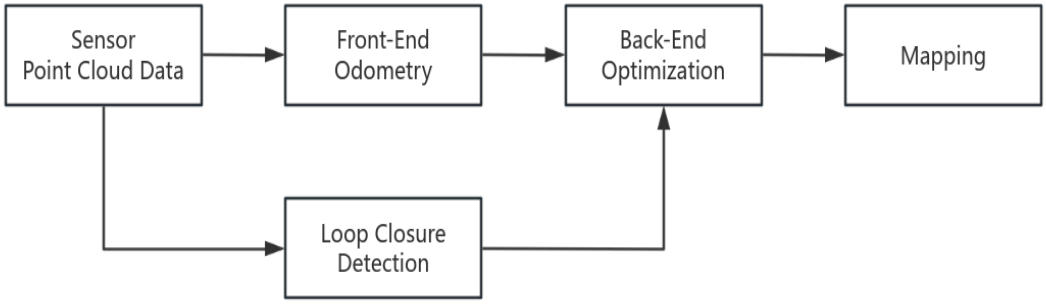

Fig. 1 SLAM working architecture (Original)

2.1. Motivation and Working Principle of Semantic Enhancement

Semantic enhancement refers to introducing category priors and instance-level information in the front-end registration, back-end graph optimization, and map representation stages. Through selective filtering and weighting, as well as instance anchors and global semantic consistency, it reduces erroneous loops in repetitive structures. The specific approach involves the front end identifying and removing or downweighting movable point clouds and reducing the weight of cross-class matching. The back end utilizes semantic consistency and object anchors to improve the discriminative power of loop closure verification and strengthen graph optimization constraints. Labeled surfel is used for incremental maintenance to provide interpretable high-level semantics for navigation and retrieval, serving downstream tasks such as navigation and retrieval. 3D laser SLAM can demonstrate high performance in static environments and can handle large-scale scenarios using consumer-grade CPUs [3].

2.2. Dynamic-Static Object Recognition

2.2.1 Dynamic-Static Object Recognition (LiDAR-MOS)

LiDARMOS aims to distinguish moving points from stationary points at frame rates, reducing outlier matching at the source and avoiding short-term dynamic writes to the map. Based on the SemanticKITTI dataset, Chen et al. projected continuous point clouds into range images and transformed the point clouds of the first N frames to the current coordinate system to generate residual images. With existing mature segmentation network structures, only the input and data labels need to be modified. Experimental results show that the SalsaNext model combined with residual plots achieves an IoU of 62.5% on SemanticKITTI, which is significantly better than existing methods [4].

2.2.2 SD-SLAM

SD-SLAM uses DBSCAN to cluster the results of point-by-point semantic segmentation into instances, and only performs state tracking on semantically consistent instances in adjacent frames [5][6]. It uses a constant rotation rate motion model in conjunction with Kalman filtering to estimate the center position and velocity, thereby classifying objects into three categories: dynamic, semi-static, and purely static. This achieves suppression of moving external points and preservation of available constraints. Only semi-static and purely static instances are used in the pose estimation and registration stages; dynamic instances are discarded. During the loop-through phase, matching is performed only on purely static instances, and semantic consistency is checked. In the map maintenance phase, only pure static instances are fixed to obtain a stable static semantic map. These strategies together reduce false loops under repetitive structures and improve global consistency [5]. Li reported in KITTI Odometry that the method achieved lower ATE and RPE in dynamic scenes compared to various comparison methods, while keeping the computational cost within an acceptable range, verifying the effectiveness of identifying who is moving, using layered methods, and only writing reliable parts into the map [5].

3. Applications and Prospects of Semantic Enhancement 3D Laser SLAM in Service Robots

3.1. Evaluation Metrics and Benchmarks

This section focuses on the evaluation and reproduction of 3D laser SLAM based on publicly available benchmarks. In the vehicle scenario, KITTI proposes to statistically analyze the relative pose error according to the subsequence length and vehicle speed to decompose the sources of translation, rotation and drift and form comparable odometer evaluations. At the same time, KITTI provides annotations for three-dimensional target detection and orientation estimation, and supports the removal of dynamic targets in semantic layer analysis [7]. The Hilti dataset is designed for engineering applications and construction scenarios, covering indoor (offices, laboratories) and outdoor spaces (construction sites, parking lots). It features weak texture/few features, variable lighting, and rapid/sudden motion. Each sequence in the dataset is provided with a motion capture system or total station, resulting in millimeter-scale sparse ground truth [8]. To supplement the diverse scenarios, this section also provides a reference for the Newer College Dataset. This dataset features more intense motion and stronger lighting contrast. It focuses on handheld platforms, uses cameras with built-in IMUs and LiDAR for data acquisition, and obtains the time offset between each measurement through spatiotemporal joint calibration. It is suitable for examining challenges such as low texture and sudden motion [9].

To facilitate horizontal comparison, the differences between the three publicly available benchmarks in terms of sensor configuration, truth source, main indicators, and typical challenges are summarized on one page, as shown in Table1.

Table1. Common SLAM datasets and their characteristics

Benchmark	Sensors	Truth value	Evaluation Suggestions	Applicable scenarios	challenge
KITTI	High-resolution video camera, LiDAR, positioning system	Stereometry and optical flow, visual odometry/SLAM, 3D target ground truth	ATE RPE loop closure assessment, long-distance consistency, high-speed motion	Vehicle-mounted platforms for cars, vans, trucks, etc.	Non-Lambertian surfaces, high-speed motion, diverse textures, and changes in lighting.

Hilti	Vision, LiDAR, and inertial sensors	Total Station/MoCap Millimeter-Level Sparse True Values	ATE/RPE equivalence error, segmented scoring, map geometric consistency	Indoor (office, laboratory) Outdoor (construction site, parking lot)	Weak feature regions and changes in illumination
Newer College Dataset	Binocular vision module; multi-line 3D LiDAR (both with built-in IMU)	High-precision 6DoF truth value	ATE/APE RPE Loopback Retrieval Robustness	Handheld high-intensity exercise equipment	Low texture, strong lighting variations, fast and irregular motion

3.2. Typical Application Cases

Service robots mostly operate in low GNSS and geometrically repetitive environments, or in indoor spaces with high personnel flow. The value of semantically enhanced 3D laser SLAM is mainly reflected in three aspects. First, semantic and motion segmentation are used at the front end to suppress dynamic interference. The backend improves loop closure detection capability through semantic consistency to reduce false loops under similar structures. The map side provides searchable representations of passable and impassable objects and functional objects and links them with the cost map. Such representations are often implemented using probability occupancy and hierarchical cost maps [10][11].

3.2.1 Hospital delivery robots

Hospital environments often consist of long corridors, repetitive topologies of intersections and ward doors, and the high mobility of pedestrians and hospital beds poses a significant challenge to traditional SLAM. Semantic enhancement is introduced in 3D laser SLAM to perform point-level or instance-level motion segmentation on the mapping side and solidify only the pure static layer. The loop closure stage checks semantic consistency in addition to geometric similarity to reduce mismatches in similar corridors [12][13]. PuduBot2 is a representative example of a multi-sensor SLAM and semantic recognition system used to complete the delivery of medicines and supplies in densely populated wards. Official disclosures show that PUDU VSLAM+ can reduce deployment time by about 75% without marking, and can operate stably in spaces with a maximum ceiling height of 30 meters. The actual test claimed that it could complete mapping of more than 1,000 m^2 in 15 minutes, complete full map configuration in one hour, and support mapping of large venues up to 40,000 m^2. These capabilities are in line with the typical low GNSS, high ceilings and repetitive corridors of hospitals, and are also suitable for indoor environments with high traffic.

3.2.2 Warehouse handling robots

The arrangement of shelves in warehouses is very complex, and the reflective properties of materials and the high-frequency dynamic changes of goods also pose challenges to traditional laser SLAM. During mapping, the shelves are treated as semi-static bodies for positioning rather than static bodies, and online motion segmentation is performed on the moving pallets. The output side maps reachable channels and temporary obstacle buffer layers to local paths and speed limit constraints, and actively adjusts the avoidance strategy using object-aware semantic cost graphs [11][13].

Simbe Tally spends a long time inspecting shelves in the store aisles and generating semantic layer inventory and display status. Simbe Toll has an RFID read/write rate of over 700 tags/second and an accuracy of over 99%, while also providing information about the scanned target. This type of high-frequency, low-latency semantic data collection allows replenishment paths and out-of-stock

warnings to be updated in near real-time, making it suitable for supermarkets and shopping malls where there are frequent changes in customer traffic and goods during peak periods.

4. Introducing the Limitations and Improvements of Semantic Enhancement

Limitation

Semantic segmentation requires supervised learning, which means that a large amount of data needs to be manually labeled. The data processing cost is very high and it is difficult to cover diverse environments and different targets. Domain shifts can easily occur across sensors and scenes, leading to a decrease in the model's generalization ability in new cities, new seasons, or new installation poses. Furthermore, semantic categories are not equivalent to motion states [14]. Semantic segmentation can explicitly indicate potential dynamic targets, thereby filtering moving objects in the feature tracking and mapping module and obtaining more accurate pose estimation. However, existing methods still need further optimization [15]. Potential dynamic classes may be static at present. Simply removing them by category will destroy the integrity of the map and lead to over-filtering and under-filtering. It needs to be combined with motion consistency or visibility methods [14].

4.1. Improvement Measures

4.2.1 Cross-modal UDA (Xmuda)

For robots that use SLAM, the collection of 3D data is often multimodal. Cross-modal UDA (xMUDA) employs a decoupled dual-branch architecture and a learning scheme that mimics each other. The modal complementarity of 3D points and 2D images enables knowledge transfer between modalities. By learning the 2D and 3D branches separately, and setting a main head and a mimicry emulator head for each. The main head is used to estimate the best prediction for this modality, while the mimicry head is used to estimate the other modality. This decouples the cross-modal imitation target from the main segmentation target, avoiding the discarding of private information to reduce cross-modal loss [16].

4.2.2 Graffiti-style labeling

Without reducing annotation costs, 3D semantic segmentation will be difficult to apply on a large scale. Graffiti-style annotation, as a form of weakly supervised training, first introduces consistent supervision to the majority of unlabeled points using an average teacher, then generates pseudo-labels through sub-training, and uses a pyramid-style local semantic context to improve the boundaries and small targets, and finally removes additional feature dependencies through a round of distillation [15]. Experiments show that near-fully supervised accuracy (relative performance of about 95.7%) is achieved with about 8% annotation, significantly reducing annotation costs and improving cross-scene robustness [15].

5. Conclusion

This paper summarizes the methods and applications of semantically enhanced 3D laser SLAM in service robots, outlining the roles of front-end suppression and weighting, back-end semantic factors and loop closure detection, and map-side semantic representation. It also summarizes benchmarks such as SemanticKITTI, Hilti, and Newer College, and illustrates these applications using scenarios like hospital delivery, supermarket replenishment, and warehouse handling. Introducing semantically enhanced SLAM requires extensive manual annotation of data and exhibits poor generalization ability. Methods such as cross-membrane UDA and doodle-style annotation can reduce the required annotation quantity, facilitating the large-scale application of semantically enhanced SLAM. The conclusion is that incorporating semantic priors into the front-end, back-end, and loop closure detection stages of SLAM can improve robustness and global consistency in low GNSS, highly dynamic, and geometrically repetitive scenarios.

References

[1] International Federation of Robotics. Service robots see global growth boom. Press Release, 2025, Oct 7. Available: https://ifr.org/ifr-press-releases/news/service-robots-see-global-growth-boom (accessed: Oct 24, 2025).

[2] Chen X, Milioto A, Palazzolo E, Giguère P, Behley J, Stachniss C. SuMa++: Efficient LiDAR-based semantic SLAM. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 4530–4537.

[3] Chen X, Li S, Mersch B, Wiesmann L, Gall J, Behley J, Stachniss C. Moving object segmentation in 3D LiDAR data: A learning-based approach exploiting sequential data. IEEE Robotics and Automation Letters, 2021, 6(4): 6529–6536.

[4] Li F, Fu C, Sun D, Li J, Wang J. SD-SLAM: A semantic SLAM approach for dynamic scenes based on LiDAR point clouds. Unpublished Manuscript, College of Mechanical and Vehicle Engineering, Chongqing University; State Key Laboratory of Intelligent Vehicle Safety Technology, Chongqing Changan Automobile Co., Ltd.

[5] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of KDD, 1996: 226–231.

[6] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of CVPR, 2012.

[7] Helmberger M, Morin K, Berner B, Kumar N, Cioffi G, Scaramuzza D. The Hilti SLAM challenge dataset. IEEE Robotics and Automation Letters, 2022.

[8] Ramezani M, Wang Y, Camurri M, Wisth D, Mattamala M, Fallon M. The Newer College dataset: Handheld LiDAR inertial and vision with ground truth. Proceedings of IROS, 2020.

[9] Hornung A, Wurm KM, Bennewitz M, Stachniss C, Burgard W. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. Autonomous Robots, 2013, 34(3): 189–206.

[10] Nav2 Team. Costmap 2D. Nav2 Documentation, accessed 2025. [Online]. Available: https://navigation.ros.org/configuration/packages/costmap2d

[11] Rosinol A, Abate M, Chang Y, Carlone L. Kimera: An open-source library for real-time metric-semantic localization and mapping. Proceedings of ICRA, 2020: 5791–5798.

[12] Peng H, Zhao Z, Wang L. A review of dynamic object filtering in SLAM based on 3D LiDAR. Sensors, 2024, 24(2): 645.

[13] Wang W, You X, Zhang X, Chen L, Zhang L, Liu X. LiDAR-based SLAM under semantic constraints in dynamic environments. Remote Sensing, 2021, 13(18): 3651.

[14] Jaritz M, Vu TH, de Charette R, Wirbel É, Pérez P. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 12605–12614.

[15] Chen X, Milioto A, Palazzolo E, Giguère P, Behley J, Stachniss C. SuMa++: Efficient LiDAR-based semantic SLAM. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019: 4530–4537.

[16] Li F, Fu C, Sun D, Li J, Wang J. SD-SLAM: A semantic SLAM approach for dynamic scenes based on LiDAR point clouds. State Key Laboratory of Intelligent Vehicle Safety Technology, 2025.