Semantic Visual-inertial SLAM for Automated Valet Parking

Seungwon Oh[1][0009−0008−0952−9740], Junghoon Seo[1], Jungho Park[1], Viswanath Veera[2], Jersha Felix[2], Midhun Menon[2], and Chinmay Shinde[2]

[1] Samsung Electronics, Suwon, Republic of Korea
{victor.oh, richy.seo, j-alan.park}@samsung.com
[2] Samsung Research and Development Institute Bangalore, Bangalore, India
{viswanath.v, jersha.felix, midhun.menon, chinmay.ms}@samsung.com

Abstract. Enhancing localization and mapping accuracy in constrained environments like parking lots is critical for autonomous driving. This paper introduces a novel visual-inertial Simultaneous Localization and Mapping (SLAM) approach tailored for automated valet parking (AVP). By incorporating semantic information such as various objects and markings found in parking lots, our method significantly enhances the robustness and precision of the localization process. These semantic features provide essential information for the automated parking system to understand the structure and rules of the parking environment, enabling more accurate navigation and decision-making. We developed a hybrid algorithm that integrates traditional key-point feature-based localization with semantic feature-based localization. The evaluations conducted in the CARLA simulator demonstrated a 54% reduction in position error compared to state-of-the-art methods, achieving an average trajectory error of 0.19 meters. These advancements are vital for improving AVP system and facilitating the broader adoption of autonomous parking solutions. Future research will focus on scaling the approach to various urban environments and addressing challenges presented by dynamic conditions.

Keywords: Visual-inertial SLAM · Semantic feature · Parking · AVP.

1 Introduction

Visual SLAM is the one of the fundamental components of autonomous driving in unknown environments. Particularly in indoor scenarios like parking lots, visual SLAM plays a critical role in ensuring safe automated parking. In addition to automated parking, extensive research is currently being conducted on Automated Valet Parking (AVP) systems in order to fulfill the requirements of drivers. Accurate localization and detailed mapping are essential for the successful implementation of AVP, ensuring safe autonomous driving.

However, it is not easy to recognize an exact location and create an useful map in a dynamic environment without prior map information during autonomous

driving. Although HD maps can provide centimeter-level accuracy with multiple layers of detailed road environment information [12, 2], visual SLAM approaches independent of infrastructure like HD maps should be considered to ensure generalizability across various parking environments without relying on HD maps. In addition, vehicle dynamics or sensor systems are non-deterministic because various errors or noises exist. Moreover, moving objects such as other vehicles or pedestrians should also be considered [9]. Therefore, to overcome these constraints, it is possible to leverage robust information tailored specifically to parking lot environments.

Therefore, we propose an improved visual-inertial SLAM methodology that leverages semantic information that is helpful in understanding parking lot environments as shown in Fig. 1. The parking lots contain various objects and markings for parking, such as parking lines, parking space availability, parking guide signs, speed limit markings, parking directions, parking zone markers, and various indicators. We aim to utilize this information for accurate localization and mapping to provide useful data for path planning in AVP.

Firstly, we identify the various objects and markings as beneficial semantic information within parking lots. By utilizing the features extracted from the semantic information, we solve the localization problem in unknown environments. Furthermore, we combine traditional key-point feature-based localization results with semantic feature-based localization results to develop a more reliable localization algorithm. This enables us to achieve more accurate localization through the utilization of semantic guided landmarks with fixed locations. Moreover, we generate a multi-layer occupancy grid map using the semantic features in real-time during vehicle movement. The resulting map data can play a critical role in path planning for AVP, providing essential input for safe and efficient navigation.

To demonstrate the superiority of the proposed approach compared to several state-of-the-art (SOTA) SLAM algorithms, we created a parking lot environment using the CARLA simulator and obtained evaluation data from this simulated environment. Our algorithm reduced position errors by approximately 54% compared to other algorithms, achieving a localization accuracy of less than 20 centimeters.

Our key contributions are summarized as follows:

- We propose enhancing visual-inertial SLAM by incorporating semantic features.
- We construct a multi-layer occupancy grid map for path planning.
- We create a customized simulated parking lot environment using the CARLA simulator for comprehensive evaluation.
- We achieve SOTA localization performance for AVP, outperforming previous approaches

We believe that our solution will ultimately contribute to the improvement of AVP performance and expedite the commercialization of AVP.

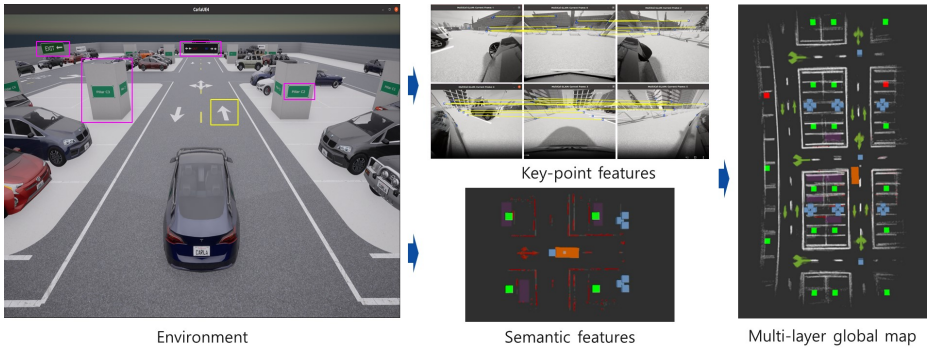


Fig. 1: Semantic visual-inertial SLAM for AVP

2 Related Work

2.1 Key-point Feature-based Visual SLAM

Traditional key-point feature-based visual SLAM technologies primarily rely on recognizing specific patterns or shapes within images to estimate the robot's position. This approach is fast, efficient, and works well under most conditions. Representative algorithms are ORB-SLAM2 [8], DSO [3], Stereo DSO [18], and LSD-SLAM [4]. ORB-SLAM2 [8] uses a monocular camera and extracts ORB features to estimate the vehicle's location, orientation, and speed. ORB features are detected quickly within the image and have the advantage of being robust to rotation and scale changes. However, key-point feature-based visual SLAM has some limitations. First, since features are based on specific shapes or patterns within images, if the image is too simplistic or lacks patterns, it becomes challenging to find sufficient features for estimation. In parking lots, even if features are detected, they may be repeating or visually similar features. This leads to ambiguity in feature descriptors, which in turn causes imprecise tracking and incorrect estimates. This leads to inaccurate or less precise location or pose estimation. Moreover, precisely identifying the positions of features is challenging, leading to cumulative errors that decrease positional estimation accuracy over time. To address these issues, VI-SLAM, which combines a camera sensor and an Inertial Measurement Unit (IMU), has been proposed. The integration of visual and IMU inputs provides robustness to low-quality texture, motion blur and occlusions, while enabling scale observability in monocular systems. Representative algorithms are OKVIS [5], ORB-3 SLAM [1], VINS-Mono [11], VI-DSO [17], and Usenko et al. [16]. In particular, the SOTA algorithm, ORB-3 SLAM [1] utilizes IMU data to maintain temporal consistency and recover from sudden movements. By improving the loop closing module to detect overlapping areas and reflect them in map updates, it showed better performance and stability than the previous ORB-SLAM [8]. Usenko et al. [16] introduced sliding window optimization technology for real-time processing and stable operation in diverse lighting

environments. Moreover, the system also includes functions like scale recovery, re-localization, and loop closure, ensuring high reliability. However, it is difficult to accurately estimate the actual trajectory. Therefore, semantic features should be considered, especially in environments such as parking lots.

2.2 Semantic Feature-based Visual SLAM

In parking lots environments with low lighting or challenging conditions for feature extraction, traditional key-point feature-based SLAM approaches have difficulty extracting meaningful features. Therefore, Shao et al. [13], Zhao et al. [19], MOFISSLAM [14], AVM-SLAM [7], and AVP-SLAM [10] proposed semantic SLAM approaches using panoramic surround-view images. With these images, semantic objects such as parking lines and slots on the ground can be reliably and consistently detected regardless of varying perspectives and low lighting conditions. However, they only consider semantic features on the ground using four surround-view cameras. Moreover, there is limitation in providing high-level information needed for interaction with surrounding environments. Therefore, various semantic features specific to parking lots should be considered.

2.3 Multi-layer Occupancy Grid Map

In addition to key-point feature-based approach and semantic feature-based approach, Li et al. [6] and Zhao et al. [19] propose a method for constructing a semantic map based on visually recognizable landmarks. These approaches identify and extract visually recognizable symbols that humans can interpret and utilize the symbols for creating a semantic map. The proposed technology significantly improves the accuracy of positioning and efficiency of path planning compared to the conventional feature-based SLAM approaches. However, relying solely on information marked on pillars is limited in providing sufficient information for path planning. Therefore, we propose a method to divide various landmark information into multiple layers based on their attributes including parking slots, parking directions, parking zone markers, and obstacles to enhance localization accuracy and path planning for AVP.

3 Method

Our approach addresses the problem of localization in parking lot environments. The sensor configuration for our target system includes six cameras and one IMU for surround sensing, which is a common setup for modern autonomous vehicles. As shown in Fig. 2, the overall architecture can be divided into two main parts. The first part is the localization module, where we implement traditional key-point feature-based visual-inertial SLAM and integrate it with visual SLAM that utilizes semantic features. The second part involves map generation, where we construct a multi-layer occupancy grid map based on the information used for localization and path planning. Specifically, we propose methods for utilizing

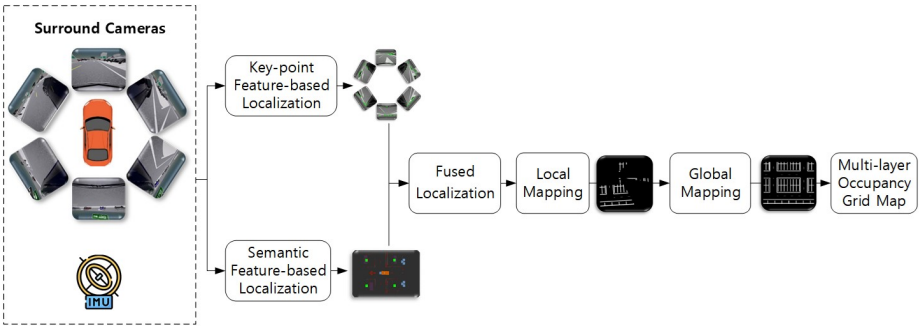


Fig. 2: Overall architecture

semantic information tailored to parking lot environments and integrating it to achieve more accurate localization performance. To verify the performance of the proposed algorithm, we built a simulator environment and collected data for algorithm development.

3.1 Key-point Feature-based Localization

First, we implement a basic vehicle odometry using cameras and IMU. IMU can provide essential information for calculating the vehicle's odometry regarding the vehicle's current velocity and angular position, crucial for the initial stages of localization. However, this not only includes noise errors but also leads to cumulative errors that occur during driving. Therefore, sensor fusion techniques are necessary to minimize these errors. A general visual odometry is required to estimate the motion of the camera using images from the camera as follows,

$$E = R[t]_x, \tag{1}$$

where E is the essential matrix. [R, t] is the pose from visual odometry. With the visual odometry, it is necessary to utilize the estimated pose changes provided by the IMU as follows,

$$\Delta V_{x,y,z} = \iint a_{x,y,z}\, dt\, dt. \tag{2}$$

As a traditional visual SLAM, key-point features are extracted from the camera images as shown in Fig. 3. This process involves identifying distinct and recognizable points within the surround camera's visual field, which are used as reference points in subsequent matching and localization processes. Because the robust detection of these features is critical for the accurate tracking of movement and orientation changes in unknown environments, we implement key-point feature-based localization by applying OKVIS [5] algorithm in a multi-camera system. Based on the implemented visual inertial odometry, the vehicle coordinate of the features are transferred into the world coordinate as follows,

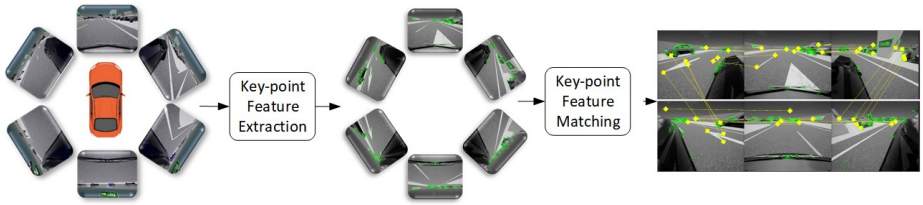


Fig. 3: Key-point feature-based localization

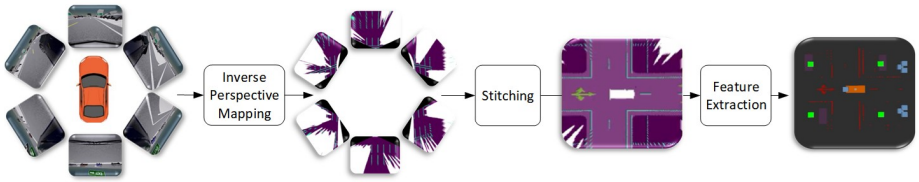


Fig. 4: Semantic feature-based localization

$$\begin{bmatrix} x^w \\ y^w \\ z^w \end{bmatrix} = R \begin{bmatrix} x^v \\ y^v \\ 0 \end{bmatrix} + t, \tag{3}$$

where [R, t] is the pose from visual odometry. Key-point feature matching is performed to compare unique key-point features extracted from adjacent cameras to identify the same features. These matches are crucial for assessing the consistency and accuracy of the feature points detected across different camera views. Through feature matching, we can obtain robust features from multiple cameras, which enhances the localization performance .

3.2 Semantic Feature-based Localization

Semantic feature-based localization consists of three steps as shown in Fig. 4. Six surround-view cameras are used in this system. Inverse perspective mapping transforms the view from each camera into a common frame of reference, typically a bird's-eye-view using cameras' extrinsic parameters. This standardized perspective is essential for aligning images from different cameras, facilitating more effective comparison and integration of visual data. This Inverse Perspective Mapping (IPM) projection process is fundamental in preparing the data for precise stitching and subsequent semantic analysis.

Following inverse perspective mapping, the stitching process integrates the transformed views from multiple cameras into a single, unified image as shown in Fig. 5. By combining images, stitching helps overcome the limitations of individual camera perspectives, creating a continuous, seamless image using smoothing filters and providing a complete visual representation of the parking area. As

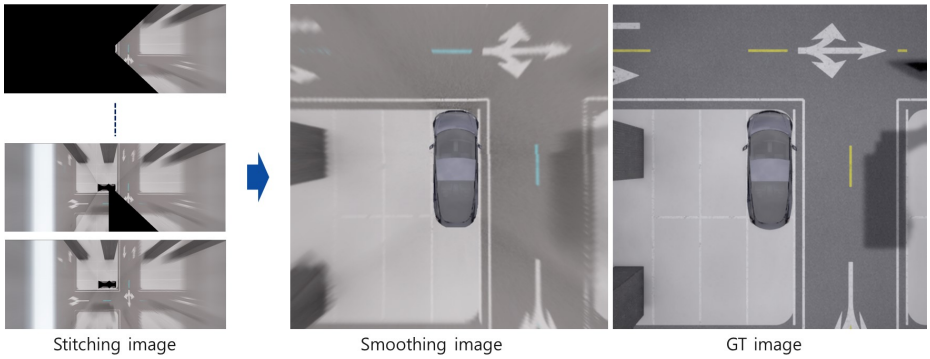


Fig. 5: Image stitching

shown in Fig. 5, it was possible to generate images similar to the ground truth image. This consolidated view is useful for extracting semantic features from the observed space.

3.3 Fused Localization

We integrate the results of different localization procedures. The position information of each feature complements each other. In other words, the position information calculated from the same camera pose information can complement each other to estimate the optimal position. By utilizing an Extended Kalman Filter (EKF) suitable for such circumstances, we realize fused localization, allowing for more accurate localization in complex and dynamic environments such as parking lots. We extract the same features from two different localization results and fuse landmarks' position information using an EKF as follows:

$$\mu_t = \mu_t + K_t(Z_t - h(\mu_t)), \tag{4}$$

where μ_t is a state vector, K_t is Kalman Gain, Z_t is position information of key-point feature-based landmarks, and $h(\mu_t)$ represents the position information of the semantic features matched with the key-point features. This fusion allows for an accurate and robust localization system essential for autonomous driving with a localization accuracy of less than 20 centimeters.

3.4 Multi-layer Occupancy Grid Map

A multi-layer occupancy grid map is created, as shown in Fig. 6. With a stitched image in place, semantic feature extraction proceeds by identifying higher-level contextual elements such as parking lines, parking space availability, parking guide signs, speed limit markings, parking directions, parking zone markers, and various indicators. These features were categorized into five layers according to their purpose. First, Layer 1 and Layer 2 represent parking lines and parking

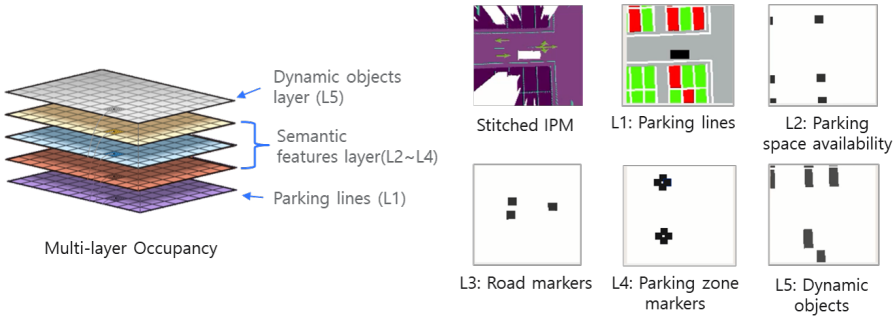


Fig. 6: Multi-layer occupancy grid map produced by our SLAM algorithm, providing a rich representation of the environment for subsequent prediction and planning stages.

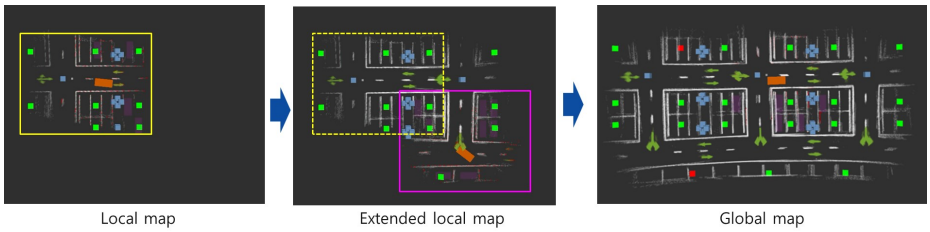


Fig. 7: Local map and global map

availability, respectively, to identify available parking spaces. Next, Layer 3 and Layer 4 depict road markers and parking zone markers, which are used to interpret various information displayed in the parking lot for parking. Finally, there is a dynamic object layer that defines pedestrians and vehicles to prevent collisions within the parking lot. Fig. 7 illustrates the local maps and the global map. The vehicle continuously captures and tracks ground plane features within the parking area, using these tracked features to estimate a ground plane map. As the vehicle moves, it expands the map in real-time by matching and stitching the semantic features of the road surface.

The proposed multi-layer occupancy grid map provides critical contextual information that complements the position data obtained from extracted features, thereby enhancing the system's ability to effectively interpret complex environments. By providing additional cues that are particularly useful in structured environments such as parking lots, semantic feature extraction plays a pivotal role in improving the system's interpretative capabilities and path planning.

4 Experiments

In order to evaluate the performance of the proposed approach, we utilized the CARLA simulator to create a custom parking lot environment, as illustrated

Fig. 8: Parking map for CARLA Simulator

in Fig. 8. We designed the parking lot to demonstrate the proposed concept and measured the localization accuracy of the algorithm while the vehicle was traveling along planned routes. Due to the lack of publicly available benchmark datasets for parking environments, we developed our own custom environment to obtain the necessary benchmark data. This environment is also expected to be valuable for future algorithm training. We measured the localization accuracy and compared the results against two SOTA visual SLAM algorithms, ORB-SLAM3 and AVP SLAM. The proposed approach achieved an average localization accuracy of 0.19 meters, outperforming the reference algorithms. In addition, we performed ablation studies to assess the impact of each individual component.

4.1 Implementation Details

The CARLA simulator is an open-source simulation platform that supports the development, training, and validation of autonomous driving systems. It provides freely usable digital assets, including urban layouts, buildings, and vehicles. The CARLA simulator offers flexibility and full control over all static and dynamic actors. The proposed solution was evaluated using synthetic data generated from the CARLA simulator. For evaluation purposes, agents were spawned within a custom parking lot map in the CARLA simulator, and required sensors were mounted on the ego vehicle as shown in Fig. 8. Table 1 provides details such as dimensions, the number of parking spaces, and other relevant metrics for our test environment.

4.2 Results

To evaluate the performance of the algorithm, we used two visual SLAM algorithms as references: ORB-SLAM3 and AVP SLAM. ORB-SLAM3 is a SOTA SLAM algorithm based on a monocular camera and an IMU. And AVP SLAM is a SOTA SLAM algorithm for automated parking. By comparing the proposed

Table 1: Test Environment. We simulated various environmental conditions to test system robustness.

Element	Specification
Total area	$9,000m^2$
Parking space	157
Lighting	$10 \sim 1,000lux$
Occupancy	$20 \sim 95\%$
Dynamic obstacle	vehicles and pedestrians

approach with two SOTA algorithms, we successfully demonstrated its superior performance.

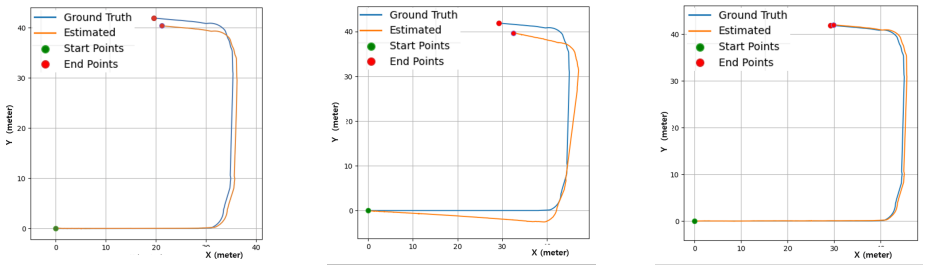
To compare performance differences across various driving paths, we collected data from the simulator on five different routes and performed a comparative evaluation. The localization accuracy was calculated as the mean position error over the entire trajectory, as described by the following equation [15].

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

As shown in Table 2, our proposed approach achieved an average localization accuracy of 0.19 meters. An accuracy level below 0.2 meters is deemed acceptable for executing autonomous parking maneuvers in parking environments.

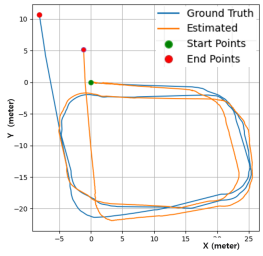
Fig. 9 illustrates the actual driving trajectories and the estimated positions for each algorithm on five different test routes. It can be observed that the proposed method consistently demonstrates stable performance with minimal errors across the entire trajectory.

Additionally, to clearly identify the performance contributions of semantic features and the multi-camera setup in the proposed approach, we defined and compared the following three different configurations, as shown in Table 3. First, we evaluated the performance of Ours(VIS) using traditional key-point features based on a single camera and an IMU. Second, we evaluated the performance

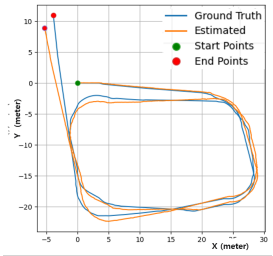


(a) ORB-SLAM3:Test1 (b) AVP SLAM:Test1 (c) Ours:Test1

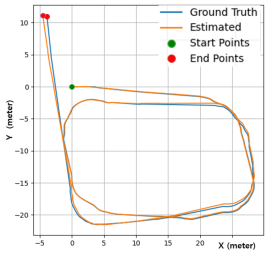
Fig. 9: Estimated trajectory results



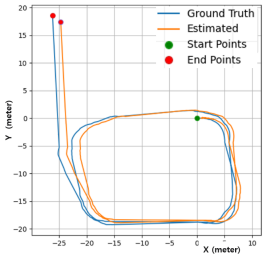
(d) ORB-SLAM3:Test2



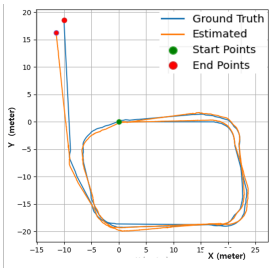
(e) AVP SLAM:Test2



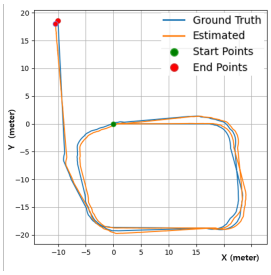
(f) Ours:Test2



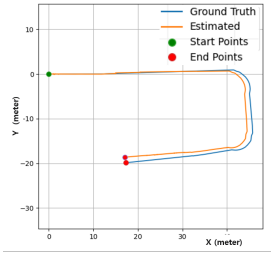
(g) ORB-SLAM3:Test3



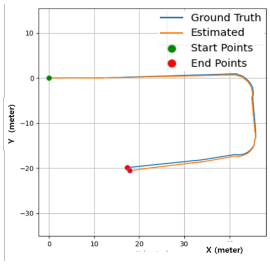
(h) AVP SLAM:Test3



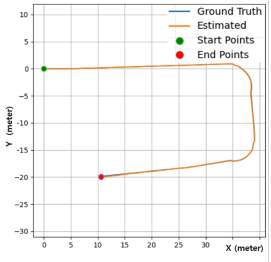
(i) Ours:Test3



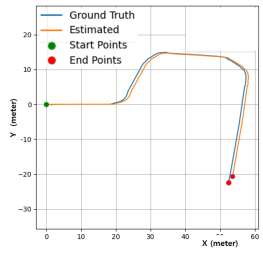
(j) ORB-SLAM3:Test4



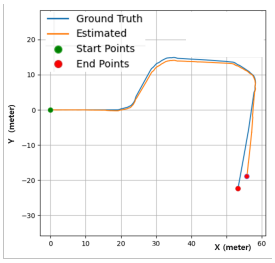
(k) AVP SLAM:Test4



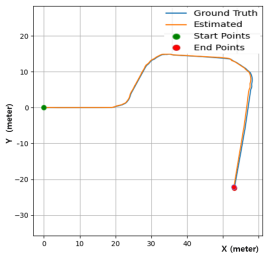
(l) Ours:Test4



(m) ORB-SLAM3:Test5



(n) AVP SLAM:Test5



(o) Ours:Test5

Fig. 9: Estimated trajectory results

Table 2: Evaluation Result. We measured the root mean square error (RMSE) for algorithm comparison evaluation. ORB-SLAM3 is a key-point feature-based SLAM using a single camera and an IMU. AVP-SLAM utilizes semantic features. Test cases can be referenced in Fig. 9. For each test case, the ground truth and estimated results by the algorithm are provided.

Method	Test1	Test2	Test3	Test4	Test5	Avg.
ORB-SLAM3	0.66	2.78	0.46	0.58	0.44	0.98
AVP-SLAM	0.35	0.39	0.24	0.40	0.68	0.41
Ours	0.14	0.19	0.21	0.28	0.15	0.19

Table 3: Model Ablation. Ours(VIS) is a traditional visual inertial SLAM using a single camera and an IMU. Ours(Semantic) is a semantic feature-based SLAM using a single camera and an IMU. Ours(Fused) is the proposed fused SLAM using multi cameras and an IMU.

Method	Test1	Test2	Test3	Test4	Test5	Avg.
Ours(VIS)	0.93	0.74	0.78	2.21	1.36	1.20
Ours(Semantic)	0.37	0.40	0.25	0.44	0.69	0.43
Ours(Fused)	0.14	0.19	0.21	0.28	0.15	0.19

of Ours(Semantic) using only semantic features from a single camera. While the performance of the mono camera configurations was not as good as the multi-camera setup, the results highlighted the significant contribution of semantic features to overall performance.

5 Conclusion

In this study, we presented an innovative visual-inertial SLAM approach specifically designed for AVP scenarios. Our method leverages semantic features to improve localization accuracy within parking environments, integrating traditional key-point feature-based localization with advanced semantic feature-based techniques. The proposed algorithm, evaluated in the CARLA simulator, achieved a significant 54% reduction in position error over current SOTA visual SLAM algorithms, with an average error of 0.19 meters. In addition, We constructed the multi-layer occupancy grid map by categorizing semantic information based on its attributes. This map can serves as valuable information for AVP in parking environments without prior map information. These results demonstrate the effectiveness of our approach in enhancing AVP performance, paving the way for more reliable and commercially viable autonomous parking solutions. Future work will aim to extend the scalability of our method to diverse urban settings and address the challenges in complex environments including dynamic objects such as moving vehicles and pedestrians.

References

1. Campos, C., Elvira, R., Rodriguez, J.J.G., M. Montiel, J.M., D. Tardos, J.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics **37**(6), 1874–1890 (Dec 2021)
2. Elhousni, M., Huang, X.: A survey on 3d lidar localization for autonomous vehicles. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1879–1884 (2020)
3. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3), 611–625 (2018)
4. Engel, J.J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: European Conference on Computer Vision (2014)
5. Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual–inertial odometry using nonlinear optimization. The International Journal of Robotics Research **34**(3), 314–334 (2015)
6. Li, S., Xing, H., Zhao, J., Huang, T., Xiong, L., Yu, Z.: Semantic mapping and localization for autonomous indoor parking based on human-readable visual landmark. In: 2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI). pp. 1–6 (2021)
7. Li, Y., Yang, W., Tao, J., Wang, Q., Cui, Z., Qin, X.: Avm-slam: Semantic visual slam with multi-sensor fusion in a bird's eye view for automated valet parking (2023)
8. Mur-Artal, R., Tardos, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics **33**(5), 1255–1262 (Oct 2017)
9. Oh, S., Hahn, M., Kim, J.: Dynamic ekf-based slam for autonomous mobile convergence platforms. Multimedia Tools and Applications **74**(16), 6413–6430 (Aug 2015)
10. Qin, T., Chen, T., Chen, Y., Su, Q.: Avp-slam: Semantic visual mapping and localization for autonomous vehicles in the parking lot (2020)
11. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics **34**(4), 1004–1020 (2018)
12. Seif, H.G., Hu, X.: Autonomous driving in the icity—hd maps as a key challenge of the automotive industry. Engineering **2**(2), 159–162 (2016)
13. Shao, X., Zhang, L., Zhang, T., Shen, Y., Li, H., Zhou, Y.: A tightly-coupled semantic slam system with visual, inertial and surround-view sensors for autonomous indoor parking. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 2691–2699. MM '20, Association for Computing Machinery, New York, NY, USA (2020)
14. Shao, X., Zhang, L., Zhang, T., Shen, Y., Zhou, Y.: Mofisslam: A multi-object semantic slam system with front-view, inertial, and surround-view sensors for indoor parking. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4788–4803 (2022)
15. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 573–580 (2012)
16. Usenko, V., Demmel, N., Schubert, D., Stuckler, J., Cremers, D.: Visual-inertial mapping with non-linear factor recovery. IEEE Robotics and Automation Letters **5**(2), 422–429 (Apr 2020)
17. Von Stumberg, L., Usenko, V., Cremers, D.: Direct sparse visual-inertial odometry using dynamic marginalization. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE (May 2018)

18. Wang, R., Schwörer, M., Cremers, D.: Stereo dso: Large-scale direct sparse visual odometry with stereo cameras (2017)
19. Zhao, J., Huang, Y., He, X., Zhang, S., Ye, C., Feng, T., Xiong, L.: Visual semantic landmark-based robust mapping and localization for autonomous indoor parking. Sensors **19**(1) (2019)