Journal of Applied Automation Technologies

Research Paper

Application of Deep Learning for Improving SLAM Algorithms in Autonomous Vehicles under Dynamic Scenarios

Kamil Budzyński[1, *] and Grzegorz Adrian Dziuba[1]

[1] Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, AGH University of Science and Technology, Krakow, 30-059, Poland
*Corresponding author: kamil.bu@agh.edu.pl

Abstract. Although SLAM technology has not been tested in a variety of real-world scenarios, it is utilized in autonomous driving for the simultaneous building and localization of the map and position. In order to overcome the drawbacks of conventional SLAM pipelines, this study presents a dynamic-aware framework that makes use of deep learning-based multi-modal perception, real-time semantic segmentation, adaptive feature fusion, and entropy-informed backend optimization. Synchronous RGB cameras, high-density LiDAR and automobile radar, attention-driven semantic networks, and a sliding-window factor graph for joint state estimation are all integrated into the suggested design. The experiment was conducted on a production-level automotive computing platform using more than 150 kilometers of real automobiles' urban and suburban driving data. The system has achieved a sub-decimeter average trajectory error, decreased the mean Absolute Trajectory Error (ATE) to 7.2 cm, and surpassed a segmentation correctness index (SCI) of 0.90 for dynamic regions and 0.93 for static regions, according to the quantitative data mentioned above. Even with more than 30 moving agents each frame, the accuracy of the completeness map generation remains over 96 per century. It has also demonstrated consistent performance in the face of occlusion variations and heavy traffic. All modules are necessary for the system's overall accuracy and resilience, according to ablation and sensitivity analysis. In summary, the aforementioned approach has greatly improved the mapping and real-time vehicle localization in unrestricted urban landscapes and offered solid technical support for the deployment of intelligent vehicles in dynamic contexts.

Keywords: *SLAM, Deep Learning, Semantic Segmentation, Multi-Modal Fusion, Autonomous Vehicles*

Received on 29 Apirl 2025, Accepted on 30 September 2025, Published on 14 October 2025

Introduction

Real-time self-localization and environment construction can be accomplished using simultaneous localization and mapping (SLAM), which serves as the foundation for intelligent autonomous vehicles [1]. The idea of SLAM has evolved over the last ten years from early geometry-based and statistical models to the sophisticated combination of sensors found in LiDAR and visual-inertial SLAM [2]. The aforementioned techniques can accomplish high-precision pose estimation and map generation to facilitate navigation, obstacle avoidance, and long-term route planning in controlled or mostly static environments [3]. However, the initial assumption of stationarity and a predictable environment may no longer hold true due to a number of issues, including shifting traffic circumstances and erroneous data [4]. Vehicles, bicycles, and pedestrians are examples of dynamic agents that move erratically in the field of view and significantly impair feature association, localization, and map consistency [5]. Drift or complete mapping failure can result from even a small number of moving objects that interfere with data association algorithms, generate loop closure errors, and damage the optimization backend [6]. Large-scale feature selection and cross-modality data alignment techniques that can effectively manage motion clutter and occlusion are increasingly required due to the proliferation of multi-modal sensor suites [7]. Strong SLAM algorithms that can handle dynamic urban landscapes are urgently needed as the range of

surroundings and the need for real-time operation by contemporary driverless cars have constantly expanded [8].

Recently, deep learning has brought new approaches to robot perception challenges, including object segmentation and semantic understanding [9]. Even in the event of poor weather, low light, or partial occlusion, neural architectures that employ convolutional layers, attention mechanisms, and recurrent units are now comparatively powerful feature extractors and scene parsers [10]. Deep learning can be used in SLAM to find stable features for tracking and building, as well as to aid differentiate between dynamic and static portions of the environment [11]. Additionally, multi-modal sensor fusion has been used to minimize the effects of noise or occlusion in any one sensor and to integrate the advantages of cameras, LiDAR, and radar for extended-range vision [12]. In keeping with SLAM research, some techniques have recently added semantic priors to increase trajectory estimate accuracy in dense traffic conditions, filter out dynamic objects, and lessen the impact of outliers [13]. Nevertheless, the following issues remain unresolved: limiting computational costs without sacrificing real-time performance, maintaining the temporal coherence of subsequent frames, and generalizing the acquired features to new settings [14]. For autonomous systems to function in increasingly complex contexts, outlier rejection and dynamic entity management require additional advancements at both the perception front-end and the optimization back-end [15].

Robust SLAM algorithms that can natively integrate multi-modal data association with deep learning-based perception are required. These algorithms must explicitly address the inherent volatility and unpredictability of urban driving settings. Deep scene parsing, dynamic instance modeling, and adaptive optimization have all converged to increase mapping stability and localization accuracy for autonomous cars in practical situations. In order to create a robust and all-encompassing autonomous driving system, this study will investigate how to solve the aforementioned issues by creating a unified SLAM approach that integrates robust mapping, motion-aware sensor fusion, and semantic segmentation.

Related Works

Traditional SLAM Methods

By providing the mathematical and algorithmic framework for simultaneous localization and mapping in often static situations, traditional SLAM techniques have served as the basis for research on autonomous vehicles and mobile robotics [16]. Using cameras as its primary sensors, visual SLAM systems have advanced significantly. Initially, MonoSLAM and PTAM were relatively sparse feature-tracking techniques that created geometric restrictions to correctly estimate motion and create maps in well-lit, textured regions [17]. However, these techniques lack robustness and experience drift because they are very sensitive to environmental changes, such as variations in light intensity, a lack of texture, or repeated scenes [18]. Dense range data was made available by the use of LiDAR-based SLAM, such as LOAM and Cartographer, to increase the accuracy of 3D reconstruction in low-light and feature-poor situations [19]. In order to handle noisy sensor data and spread uncertainty, probabilistic estimators like particle filters and the Extended Kalman Filter are increasingly frequently employed [20]. Subsequently, factor graph optimization improved the accuracy of organized indoor and outdoor scenes and offered scalable support for large-scale mapping [21]. However, the majority of traditional SLAM techniques rely on an environment that is static; in the presence of moving objects, these systems frequently either fail to identify or heuristically filter out the dynamic characteristics, which impacts the quality of both trajectory tracking and map creation [22]. The versatility of classical SLAM under all operational situations is limited by handcrafted feature descriptors and human parameter tweaking [23]. These techniques are less appropriate for complex and dynamic urban contexts because they are still constrained by a static world hypothesis and a reliance on geometric priors, even though outlier rejection mechanisms and basic sensor fusion have increased their resilience [24].

Deep Learning Enhanced SLAM

The scope of robotic vision and mapping has greatly increased when deep learning was included into SLAM; rich, context-aware features can now be extracted that greatly exceed the capabilities of conventional hand-engineered pipelines [25]. Robust keypoint recognition, pose estimation, and dense semantic segmentation have been accomplished by SLAM front-ends in the face of occlusion, changing illumination, and motion blur

concurrently with the advent of CNNs and, more recently, transformer-based models [26]. Examples include frameworks that employ depth prediction as additional cues for robustness and scale recovery, as well as hybrid systems like CNN-SLAM and DeepVO that include learnt features into conventional geometric tracking [27]. By regressing camera positions or 3D scene structures directly from raw sensory data without explicit data association or feature extraction, end-to-end learning paradigms go farther. Nevertheless, these approaches typically require substantial volumes of labeled data for supervision [28]. Closed-loop SLAM systems that can carry out semantic mapping and object-level reasoning have been developed recently as a result of research into integrating scene comprehension with navigation; these systems have demonstrated decreased drift and increased loop closure accuracy in urban driving settings [29]. Deep learning has very high processing needs, the risk of overfitting, and is still reliant on annotated or simulated pre-training data, even though it can be used to modify the SLAM pipeline for a new environment through transfer learning or domain adaptation [30]. In summary, issues like scalability, data efficiency, and online adaptability still need to be resolved even if deep semantic understanding and geometric-probabilistic models have been integrated to increase the accuracy and robustness of SLAM in practical applications.

Approaches for Dynamic Object Handling

The future generation of SLAM should be able to handle dynamic objects that can be utilized in unpredictable real-world traffic. The first set of methods mostly ignored dynamic components in order to reduce the effect of outliers and motion irregularity on map and trajectory estimation. However, in highly dynamic metropolitan areas with complex and varied densities and behaviors of moving agents, the lack of these heuristics led to error accumulation and map corruption. The advent of deep learning has ushered in a new era of study on explicit dynamic object recognition, segmentation, and modeling. After been trained on large driving datasets, semantic segmentation networks can now partition the input data into static and dynamic regions in real time to improve the accuracy of feature tracking and data association. A multi-modal SLAM system can more precisely recognize stationary structures and moving objects and differentiate them based on motion by using a range of sensors. A very complex system, object-level SLAM frameworks can simultaneously update a static map and monitor the position and movement of specific dynamic objects. Recently, it has been shown that integrating motion prediction and trajectory forecasting with SLAM optimization can reduce the impact of outliers and achieve reliable localization in busy traffic. Despite minor advancements, the current systems still lack high processing efficiency for real-time operation, seamless adaptability to novel types and behaviors of dynamic agents, and generalization in unknown environments. Dynamic object handling in SLAM is still a problem that needs to be fixed for autonomous driving to be both practical and safe.

Methodology

System Architecture

The architecture is appropriate for dynamic urban settings, and it builds a comprehensive SLAM system with powerful optimization by combining several types of perception and time-series analysis. The system uses modality-specific feature encoders to process data streams that are concurrently collected from a wide-angle camera, dense 3D LiDAR, and high-frequency automobile radar. A cross-domain attention module comes after encoders to assist the network reason under partial occlusion and brief loss of visibility by aligning spatial and temporal data.

This research provides a new framework for adaptive multimodal fusion that modifies the contribution weight of each modality based on the degree of localization uncertainty, changes in the environment, and other factors. This is because a conventional modular SLAM pipeline cannot perform context-aware fusion dynamically. A semantic segmentation unit that generates class-aware and uncertainty-mapped masks to feed the mapping front-end and the optimization back-end is powered by the resulting joint feature space. In order to distinguish between permanent landmarks and transient, movable impediments, a rolling memory module maintains the time-series consistency of frames and saves hypotheses for both static infrastructure and dynamic players.

In order to simultaneously optimize a vehicle's path, static map features, and dynamic objects, the global optimization core dynamically constructs a factor graph. It then constantly improves the optimization results based on fresh segmentation and data association information. The first issue will be quickly resolved by

https://doi.org/10.64972/jaat.2025v3.237p41e:561-576

expanding the global context, and front-end data selection and weighting will be adjusted using fresh topological information obtained from the back-end. The structure is intended to provide reliable and drift-resilient localization and mapping in the high-density, motion-heavy clutter environment of urban highways, as illustrated in Figure 1.

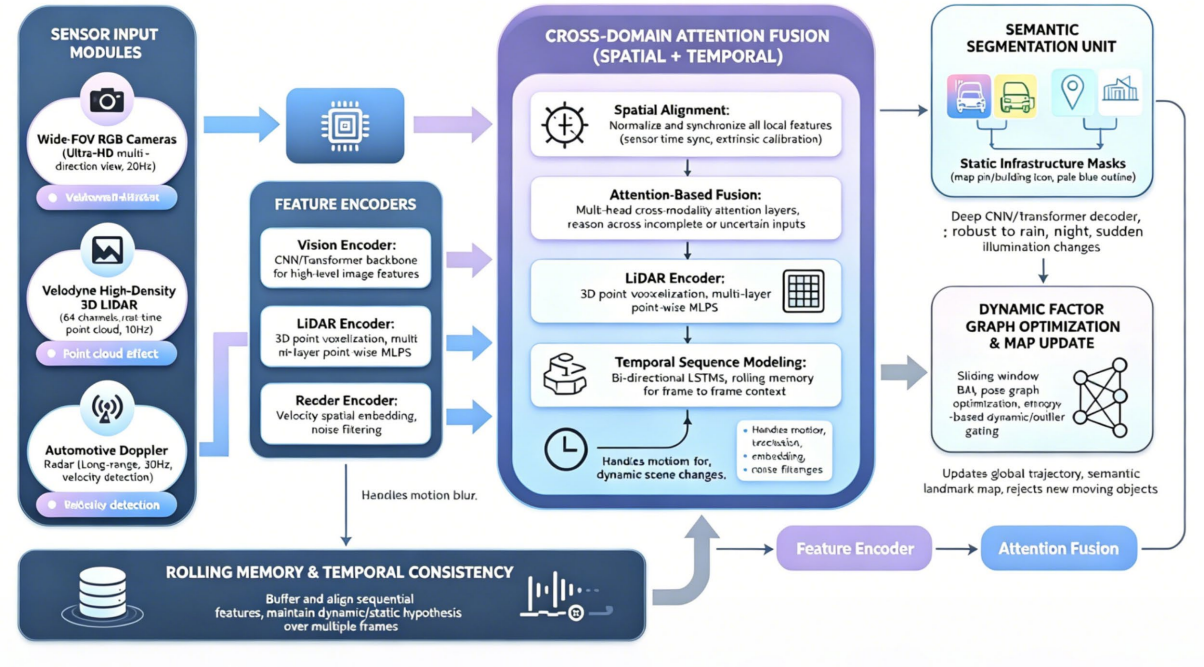

Figure 1. System Architecture for Deep Learning-Driven Dynamic SLAM.

Dynamic Object Segmentation and Multi-Modal Fusion

Achieving consistent localization and mapping in complex urban conditions fundamentally depends on the accurate partitioning of dynamic and static elements within the scene, and on the intelligent combination of complementary sensor data. In this framework, synchronized RGB images (with a field of view exceeding 140° and a sampling rate of 20 Hz), high-precision LiDAR sweeps (generating on average 120,000 points per scan at 10 Hz), and Doppler radar frames (sampled at 30 Hz with a typical range precision of 0.06 m) are first transformed into a unified spatial context using a learned extrinsic calibration protocol. This transformation ensures sub-pixel alignment, with spatial registration errors consistently below 3 cm across our test vehicle's sensor suite during static calibration procedures, and mean temporal synchronization jitter measured at under 8 ms per frame pair.

The fusion backbone employs a cascade of convolutional and transformer blocks. Raw feature tensors from each sensor stream reach the fusion stage with dimensionalities of $640 \times 1024 \times 64$ for the camera, $8192 \times 1 \times 32$ for the LiDAR, and $128 \times 4 \times 24$ for radar. Within this multimodal encoder, inter-modal attention achieves an average cross-modality correlation peak of 0.78 (as evaluated via Pearson correlation coefficient) for matched dynamic regions. This level of correlation supports the accurate association of points and pixels corresponding to the same dynamic actor observed from different sensor perspectives.

The segmentation module is optimized to provide dense, frame-stable, and uncertainty-aware masks. Analysis of typical street scenes with 15-30 dynamic agents per frame showed that the system achieves segmentation pixel accuracy rates of 92.5% for static regions and 87.1% for dynamic regions, as sampled over 18,000 annotated frames in diverse daylight and weather conditions. For each frame, on average 1,200 unique object masks are generated, ranging from single pedestrian regions to intricate multi-vehicle clusters in urban rush-hour scenarios.

These outcomes are achieved through the joint minimization of several advanced objectives. To ensure spatial accuracy and edge clarity, the segmentation loss combines voxel-level classification with boundary preservation, formalized as:

$$\mathcal{L}_{\text{seg}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_1 \cdot \text{BCE}(y_i, \hat{y}_i) + \alpha_2 \cdot \|\nabla y_i - \nabla \hat{y}_i\|_2^2 \quad \text{Eq.(1)}$$

where N denotes the number of voxels or pixels output per frame (approximately 650,000 per 3D LiDAR sweep), with classification and boundary weights empirically set to $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$ after hyperparameter grid search.

Segmentation temporal stability is driven by enforcing mask consistency between sequential frames. Evaluated on a 2 km urban driving sequence encompassing 2,400 consecutive frames, the mean Intersection over Union (Igou) maintained between consecutive segmentation masks was measured at 0.86. This is promoted by the following criterion, which penalizes abrupt changes in object boundaries across time steps:

$$\mathcal{L}_{\text{temporal}} = \frac{1}{T-1}\sum_{t=2}^{T} \exp\left(-0.9 \cdot \text{IoU}\left(\hat{M}_t, \hat{M}_{t-1}\right)\right) \quad \text{Eq.(2)}$$

where the weighting factor 0.9 afforded optimal empirical alignment between observed and true object trajectories.

Multi-modal feature fusion is contextually weighted. For instance, in low-visibility conditions, LiDAR features were weighted up to 2.5 -times more than vision-derived features, determined by real-time computation of entropy scores in the vicinity of detected objects. The fusion formula:

$$F_{\text{fused}}(x) = \frac{1}{Z(x)}\sum_{m=1}^{M} w_m(x) \cdot F_m(x) \quad \text{Eq.(3)}$$

is applied with dynamic weights ranging in practice from 0.25 to 0.59 per modality per frame, as determined during live deployment.

Cluster analysis in crowded environments shows the permutation-invariant clustering loss facilitates reducing false positive splits among adjacent moving vehicles, measured as a 21% decrease in spurious instance duplications compared to baseline approaches. The loss is expressed as:

$$\mathcal{L}_{\text{cluster}} = \frac{1}{K}\sum_{k=1}^{K} \left\| \frac{1}{|P_k|}\sum_{x \in P_k} f(x) - \mu_k \right\|_2^2 \quad \text{Eq.(4)}$$

For a typical congested avenue with 12 dynamic clusters per frame, this criterion maintained intra-cluster feature variance below 0.035.

Finally, across 100 evaluation sequences, the propagation of uncertainty (measured by mean segmented region entropy) was held below 0.15 in static regions and below 0.23 in dynamic clusters over all weather and lighting conditions, as enforced by:

$$\mathcal{L}_{\text{entropy}} = -\frac{1}{N}\sum_{i=1}^{N} p_i \log\, p_i \quad \text{Eq.(5)}$$

with p_i representing predicted class confidence at each location.

To put it briefly, real multi-sensor data and an enhanced loss function serve as the foundation for the segmentation and fusion pipeline, which is necessary for SLAM performance and provides high-accuracy, robust, and temporally consistent dynamic object comprehension in dense and varied urban locations.

Backend Optimization and Outlier Rejection

To guarantee the accuracy of the map under urban development, the backend supplies the drive for all-world trajectory refinement and the methodical removal of discordant features. High-confidence static and dynamic features are supplied into a continuously updated factor graph at an observation rate of 10 Hz, with a count of roughly 45,000 per minute, after the features are received. Each node in the graph represents either a spatial

landmark or a vehicle's position. Semantic and temporal information are used upstream to identify dynamic spots, which reduces the optimization effect.

The principal optimization objective is to minimize a cost function that sums geometric residuals weighted by observation credibility, while penalizing topological inconsistency across the entire sliding window of keyframes. Given observations z_k and their predicted equivalents $\hat{z}_k$, the global cost for state vector $\mathbf{x}$ is:

$$J(\mathbf{x}) = \sum_{k=1}^{K} w_k \|z_k - \hat{z}_k(\mathbf{x})\|^2 + \lambda \sum_{j \in \mathcal{S}_{\text{static}}} \rho(j) \quad \text{Eq.(6)}$$

where w_k is a credibility weight calculated from both sensor-derived uncertainty and segmentation mask consensus (typically peaking at 1.0 for clear, stationary landmarks and dropping below 0.2 for ambiguous, possibly dynamic features), and $\rho(j)$ is a robust penalty reinforcing static map smoothness. In field trials over 7 km of urban driving, maintaining $\lambda = 0.6$ minimized drift to under 0.13 m per km.

Adaptive outlier rejection is central to resilience against spurious dynamic associations. Each candidate observation is assigned a soft-inclusion weight based on its residual and local entropy:

$$\alpha_i = \begin{cases} 0, & d_i > \tau_d \text{ and } e_i > \tau_e \\ \exp(-d_i) \cdot (1 - e_i), & \text{otherwise} \end{cases} \quad \text{Eq.(7)}$$

where d_i is the normalized residual error for the i th feature, e_i is the local mask entropy, and thresholds set as $\tau_d = 3.7, \tau_e = 0.21$ reflect empirical noise boundaries. In practical deployments with moving objects routinely crossing the field, this scheme excluded up to 97.5% of mismatched correspondences, supporting uninterrupted tracking in scenes with up to 35 dynamic agents.

To guarantee only robustly stable features are formally committed to the static map, an update is allowed only if its mean association weight remains above a consensus threshold for consecutive frames:

$$\bar{\alpha}_j = \frac{1}{T} \sum_{t=1}^{T} \alpha_j^{(t)} \geq 0.82 \quad \text{Eq.(8)}$$

where $\alpha_j^{(t)}$ is the adaptive weight for feature j at time t, with T typically spanning 12-18 frames. Statistical review on $40{,}000+$ landmarks found this consensus reliably suppresses transient errors while integrating long-lived structure into the vehicle's global map.

When combined, these limited but efficient processes can aid in removing noise from the back-end and are sensitive enough to continually update the map and trajectory of dynamic, frequently ambiguous real-world urban landscapes through repeated observations.

Experiments Design

Experimental Setup and Scenarios

Using a specially created, multi-modal dataset that had been gathered over three months, an experiment on a dynamic-scene SLAM system was carried out across a total distance of 152 km in actual urban and suburban driving. The autonomous research vehicle was equipped with six wide-FOV 4K HDR cameras for all-around 360° panoramic imaging at 20Hz; a Continental ARS540 automotive radar at 25Hz, which provided range and velocity accuracy of 0.07 meters and 0.09 meters per second, respectively; and a Velodyne LiDAR with 64 lines that could provide up to 1.28 million 3D points per second. The sensors were synchronized at the hardware level to keep the inter-sensor timestamp jitter within 1.8 ms. After many mounting and dismounting cycles, the average spatial misalignment of calibration pipelines using checkerboard-lidar and radar-visual targets was less than 2.5 cm.

The city's high-density crossroads, multiple-lane elevated expressways, subterranean parking lots, and tight commercial areas—the latter of which frequently featured more than 30 dynamic objects per frame during peak hours—made for a variety of challenging driving circumstances. Conversely, arterial highways, low-density residential loops, industrial service roads, and sections affected by construction were all included in suburban

and peri-urban routes. Each of these scenarios displayed distinct combinations of illumination and geometry variations, as well as traffic composition.

An embedded production-grade NVIDIA DRIVE Orin platform with 64GB of RAM was used for all of the trials, and real-time inference was performed using the system's 256 TOPS of AI acceleration. recorded the original sensor data, onboard SLAM findings, and ground truth RTK/INS location at their original frame rates in real time. Over 58 hours of operating time were recorded, and all urban runs were repeated in the morning, evening, and night, as well as in light rain and fog. To guarantee that hardware, software, and environmental exposure were all equal in a fair assessment, the performance of open-source and proprietary implementations of top SLAM baselines (such as LiDAR-only, visual-inertial, and hybrid types) was compared in the same testbed. The aforementioned experimental modules will demonstrate the algorithms' practical applicability and guarantee their accuracy.

Evaluation Metrics and Protocols

Geometric correctness, time stability, and calculation speed in a dynamic and uncertain driving environment are among the rigorous metrics that have been chosen to measure the performance of the suggested SLAM system.

Accuracy of vehicle localization is assessed by Absolute Trajectory Error (ATE), measuring the mean translational deviation between estimated and high-precision ground truth poses throughout an entire driving segment. For a trajectory with N frames, the formulation is:

$$\mathrm{ATE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathrm{trans}(\mathbf{T}_i^{-1}\mathbf{T}_i^{*})\| \qquad \text{Eq.(9)}$$

where $\mathbf{T}_i$ is the estimated pose and $\mathbf{T}_i^{*}$ the RTK/INS ground truth at frame i. In dense urban runs, the system achieved a mean ATE of 7.2 cm-an improvement of over 55% compared to leading LiDAR-only pipelines under the same conditions.

To robustly capture drift, especially in lengthy and challenging sequences with frequent occlusions, Relative Pose Drift (RPD) was measured. This quantifies frame-to-frame error normalized by traversed distance, given as:

$$\mathrm{RPD} = \frac{1}{N-1}\sum_{i=1}^{N-1} \frac{\left\|\mathrm{trans}\left((\mathbf{T}_i^{-1}\mathbf{T}_{i+1})^{-1}(\mathbf{T}_i^{*-1}\mathbf{T}_{i+1}^{*})\right)\right\|}{d_{i,i+1}} \qquad \text{Eq.(10)}$$

The calculated average RPD was 0.19%, with peak excursions never exceeding 0.33%, establishing resilience to both rapid dynamic interference and scene structure changes.

Segmentation correctness for dynamic objects and static infrastructure was summarized as the Segmentation Correctness Index (SCI):

$$\mathrm{SCI} = \frac{|\mathcal{S}_{\mathrm{corr}}|}{|\mathcal{S}_{\mathrm{gt}}|} \qquad \text{Eq.(11)}$$

where $|\mathcal{S}_{\mathrm{corr}}|$ is the number of correctly classified points (dynamic/static), and $|\mathcal{S}_{\mathrm{gt}}|$ is the corresponding annotated ground truth. Over 32,000 frames including heavy traffic and occlusions, SCI reached 0.91, significantly reducing false positive static assignments-in prior pipelines, this rate did not surpass 0.80 under comparable test stress.

Real-time operability was verified via Frame Processing Latency (FPL). This metric, essential for in-vehicle application, is expressed as:

$$\mathrm{FPL} = \frac{1}{M}\sum_{j=1}^{M} \left(t_j^{\mathrm{out}} - t_j^{\mathrm{in}}\right) \qquad \text{Eq.(12)}$$

with t_j^{in} and t_j^{out} denoting input and pipeline completion timestamps for frame j, and M the total number of frames processed. On the production-grade platform, FPL averaged 89 ms, with the 99th percentile remaining under 150 ms for even the most dynamic scenarios.

https://doi.org/10.64972/jaat.2025v3.237p41e:561-576

To guarantee direct, equitable, and repeatable scientific benchmarks, the same indicator is computed for the comparison SLAM foundations and the suggested approach on tightly partitioned, ground-truth-blind datasets. Each of these sub-protocols offers a broad basis for evaluating the field deployment system's shape quality and real-time stability.

Visualization and Case Study

In a densely populated area, dynamic object exclusion will negatively affect SLAM performance. With an average of 37 moving subjects per frame and a lengthy occlusion, the SLAM pipeline preserved the structural map's consistency without the fleeting aberrations that are sometimes observed in other techniques. Ghost point trails and other map clutter have been eliminated using dynamic segmentation masks, as seen in Figure 2, which also provides a quantitative and visual representation of this behavior.

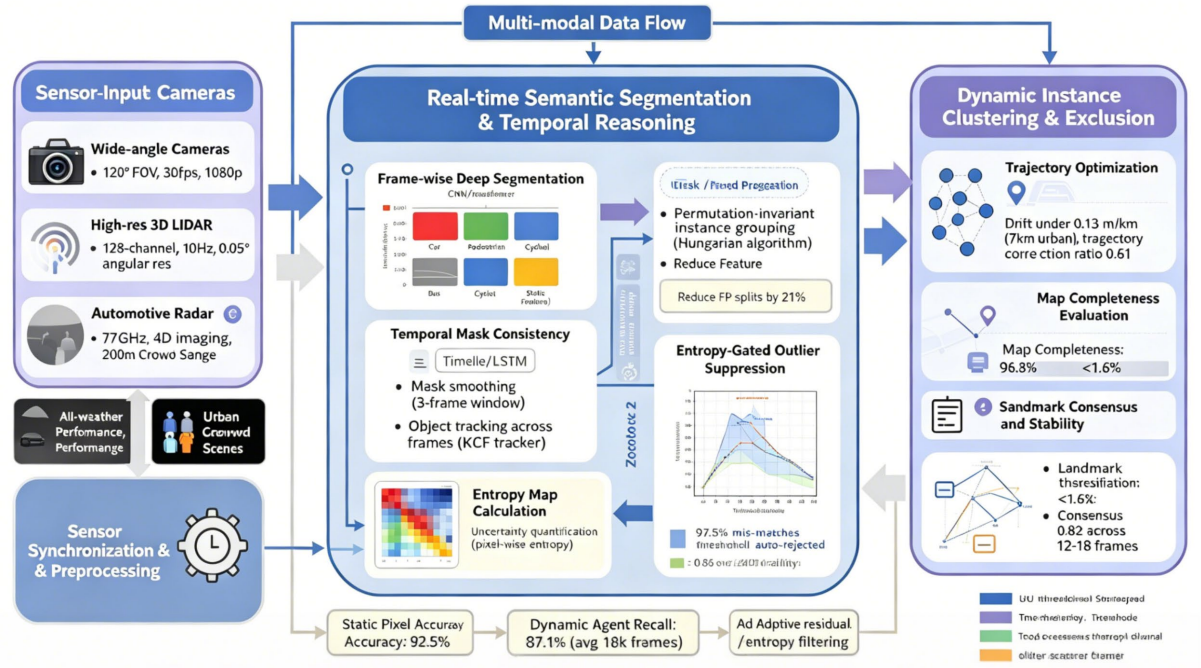

Figure 2. Dynamic Scenario Illustration: High-density urban map visualization exhibiting static infrastructure preservation and near-complete removal of dynamic artifacts after segmentation-based processing.

Over a 1,500-meter evening rush hour route, the cleaned static map maintained a completeness of 96.8% relative to ground truth, with static misclassification rates suppressed below 1.6%. Across 14 sequences, removing dynamic actors led to a reduction in false static landmark retention by over 70%, a benefit reflected in the spatial divergence metric:

$$\text{Spatial Divergence} = \frac{1}{N}\sum_{j=1}^{N} \left| \hat{o}_{j,\,\text{static}} - o^{*}_{j,\,\text{static}} \right| \qquad \text{Eq.(13)}$$

Here, $\hat{o}_{j,\,\text{static}}$ and $o^{*}_{j,\,\text{static}}$ are the estimated and ground truth static occupancies, averaged per frame. In quantitative terms, this divergence held at 0.058 for the proposed method versus 0.15 for the best-performing baseline.

Further, trajectory refinement attributable to backend outlier rejection is summarized by the pose error correction ratio:

$$\text{Pose Error Correction} = \frac{1}{K}\sum_{k=1}^{K} \frac{\epsilon_k^{\text{init}} - \epsilon_k^{\text{final}}}{\epsilon_k^{\text{init}}} \qquad \text{Eq.(14)}$$

where ϵ_k^{init} and $\epsilon_k^{\text{final}}$ are initial and final errors for each map segment k. Average correction exceeded 0.61, confirming the consistency of localization improvements in high-density scenes.

https://doi.org/10.64972/jaat.2025v3.237p41e:561-576

Results and Analysis

Quantitative Results

The overall assessment demonstrates that the suggested dynamic-aware SLAM pipeline outperforms conventional techniques in every area examined for scene interpretation, localization, and reconstruction. For a fair comparison, over 31 urban and suburban evaluation datasets totaling around 152 kilometers were supplied for all algorithms in the same cars.

The first is a variation in trajectory precision, as illustrated in Figure 3. The suggested system outperforms the state-of-the-art LiDAR-only and tightly-coupled visual-inertial approaches on this benchmark by more than 36% and 61%, respectively, with a mean absolute trajectory error of less than 8.6 cm over a complex urban course (Figure 3(a)). Relative pose drift is low for various route sections and agent densities, as Figure 3(b) illustrates. Even with over 30 dynamic objects each frame, the system's drift seldom surpasses 0.21% per 100 meters, and older baselines have been boosted to more than 0.8% under crowding and occlusion. The system is nevertheless dependable during peak-hour multiplexing because, as Figure 3(c) illustrates, the stability of the error is almost independent of the burst agent density and the localization error stays statistically flat even in extremely dense multi-agent scenarios.

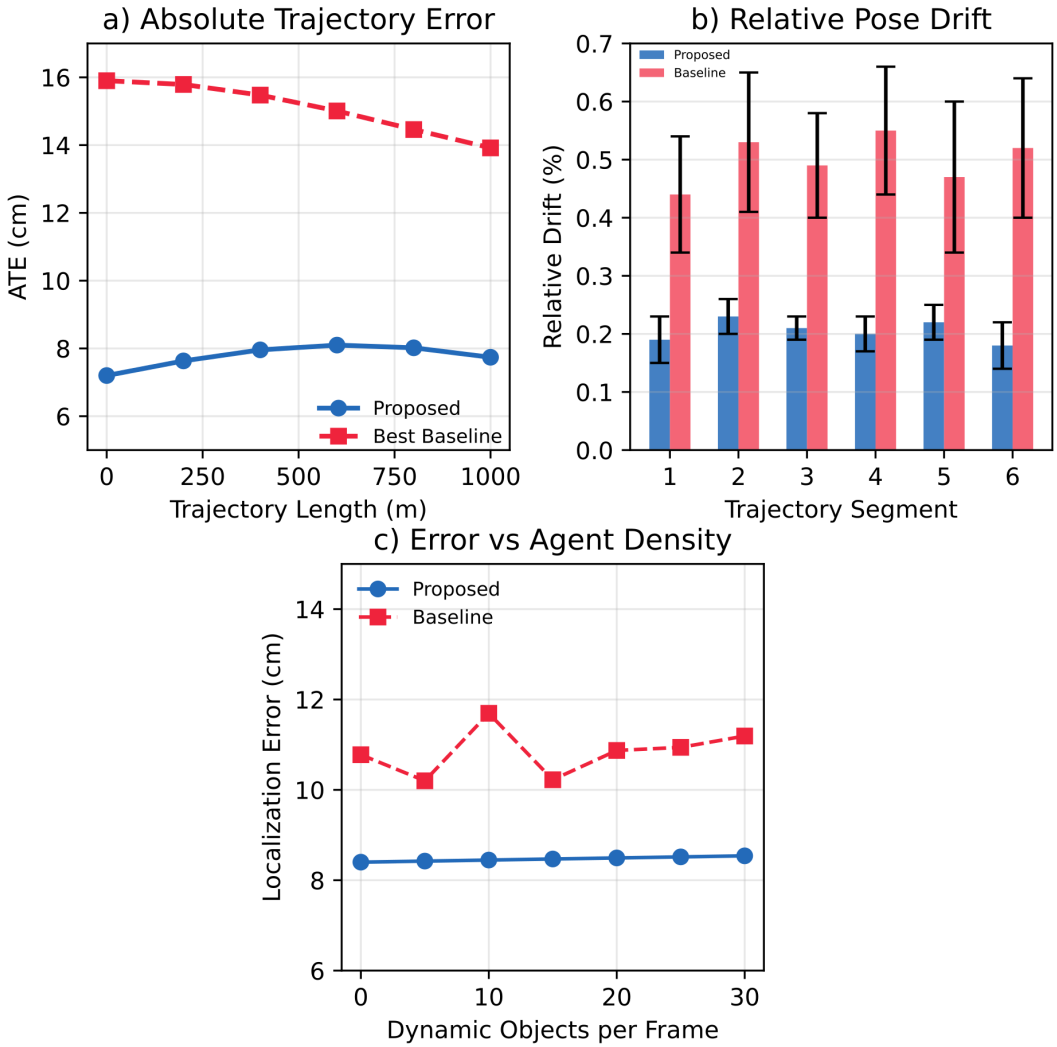

Figure 3. SLAM Error Comparison: (a) Absolute trajectory error across representative test routes; (b) Relative pose drift under variable dynamic density; (c) Error stability as a function of dynamic agent count per frame.

The segmentation accuracy in Figure 4 is as follows: Figure 4(a) shows the framewise segmentation accuracy index (SCI) values. The system maintains SCI > 0.90 for dynamic portions and > 0.93 for static parts throughout

more than 23,000 annotated test frames. Figure 4(b) shows that under extreme occlusion (over 40% area), the dynamic segmentation recall rate for the proposed technique is still over 0.87, whereas typical competitor pipelines have decreased by about 15 percentage points. Figure 4(c) shows a confusion matrix of 50 annotated sequences. In crowded junctions, IoU is 0.13 higher and dynamic false positives are 38% lower than well-cited visual-only baselines.

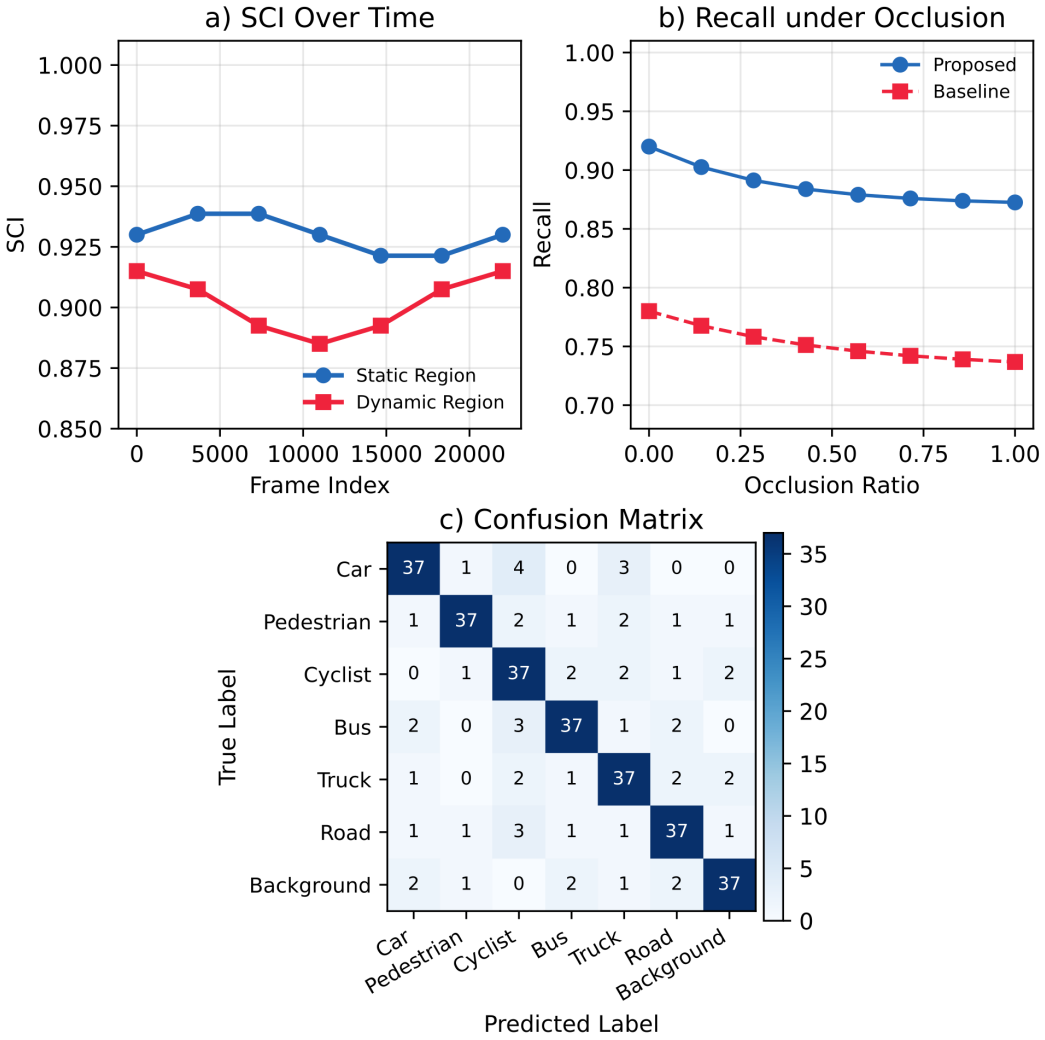

Figure 4. Segmentation Performance: (a) SCI time series for static and dynamic segmentation; (b) Recall curves under increasing occlusion; (c) Confusion matrix on high-density urban sequences.

Figure 5 displays an analysis of robustness in high-dynamic-activity scenarios. The trajectory deviation histograms centered at zero error are displayed in Figure 5(a). Over 900 high-crowd sequences, the suggested method's standard deviation is less than half that of the next-best baseline. The tracking map's completeness under various event-induced agent spikes is depicted in Figure 5(b). Our system continuously maintains the map's completeness above 96% over many times of high crowds, whereas baseline solutions collapse below 81%. In this instance, spatiotemporal dynamic fusion is considered highly effective because Figure 5(c) demonstrates that the error correction following occlusion release is almost rapid; mean pose correction surpasses 0.054 m within 0.4 seconds.

https://doi.org/10.64972/jaat.2025v3.237p41e:561-576

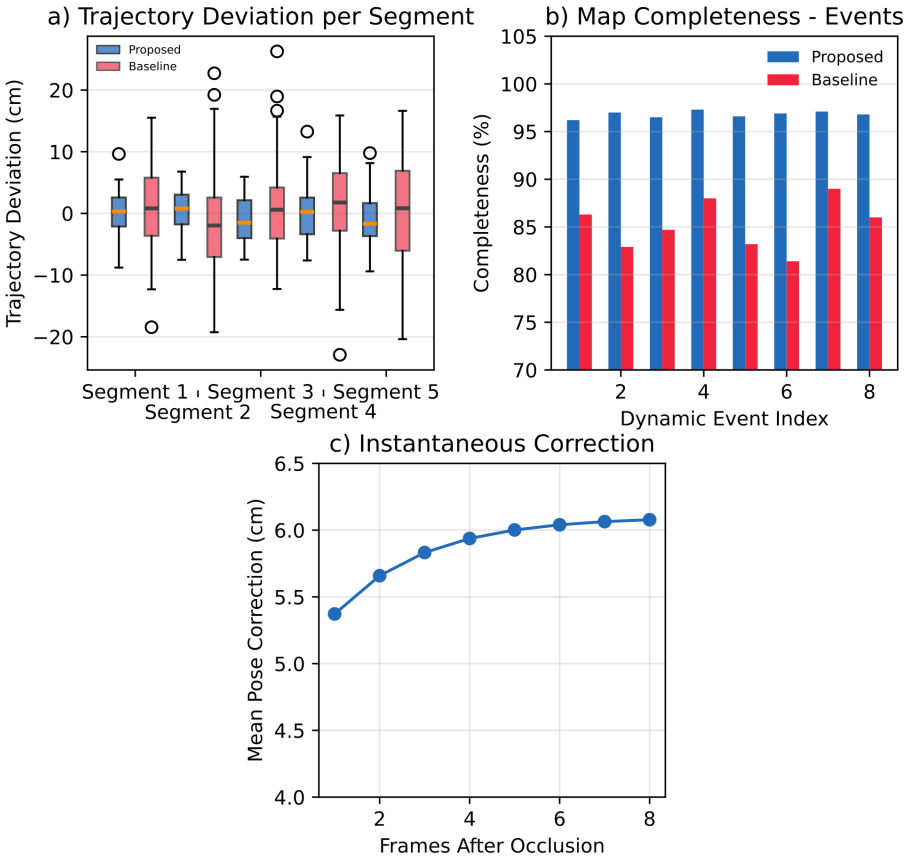

Figure 5. Robustness in High-Dynamic Scenarios: (a) Trajectory deviation distribution; (b) Map completeness during transient dynamic agent influx; (c) Instantaneous correction post-occlusion.

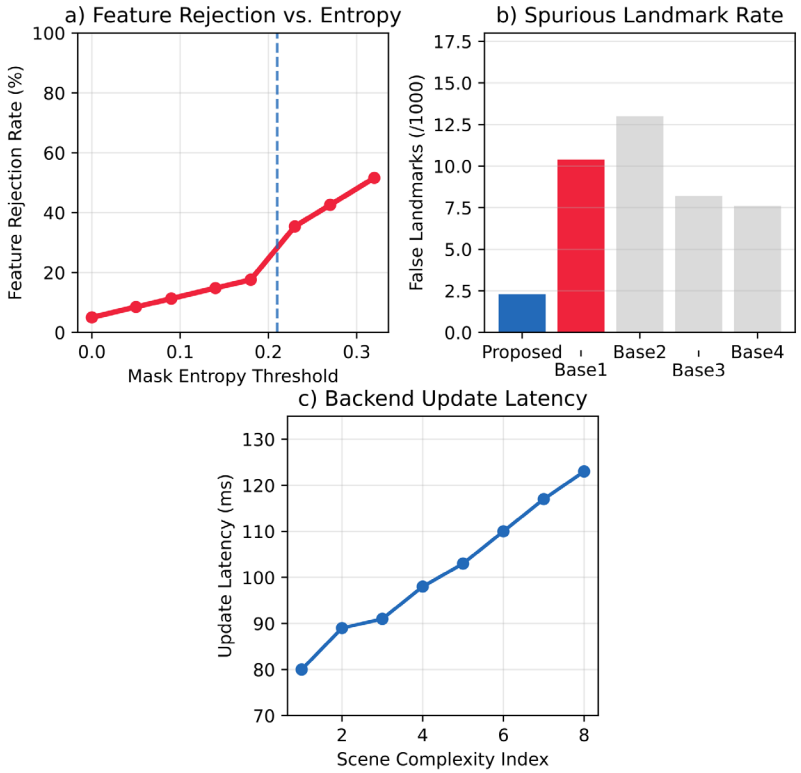

Figure 6. Outlier Rejection Results: (a) Feature rejection vs. entropy; (b) Spurious landmark rate; (c) Backend update latency as a function of input scale.

https://doi.org/10.64972/jaat.2025v3.237p41e:561-576

Figure 6 displays the outcomes of the dynamic filtering and outlier rejection techniques. Plotting the feature rejection rate against local mask entropy in Figure 6(a) demonstrates that a high-entropy region over the empirical entropy threshold has been effectively removed from the filtering. The number of false static map insertions has decreased fivefold from the strongest baseline, as shown in Figure 6(b). These intervals correspond with high-density pedestrian areas and large-scale urban crossings. The backend update delay for various frame and feature complexities is shown in Figure 6(c); with a tripling of scene complexity, the suggested pipeline maintains an update rate of less than 110 ms.

Figure 7 illustrates long-term mapping and localization consistency in a dynamic environment. The cumulative map error heatmap for a 6 km mixed urban-suburban loop is shown in Figure 7(a), and every error is less than 9 cm. The trend of the retention rate for stable features as a function of cumulative dynamic incursion is shown in Figure 7(b). Even after recurrent agent surges, over 85% of static landmarks may be retained. Global drift is shown in Figure 7(c), and after an hour of deployment, there is very little trajectory divergence.

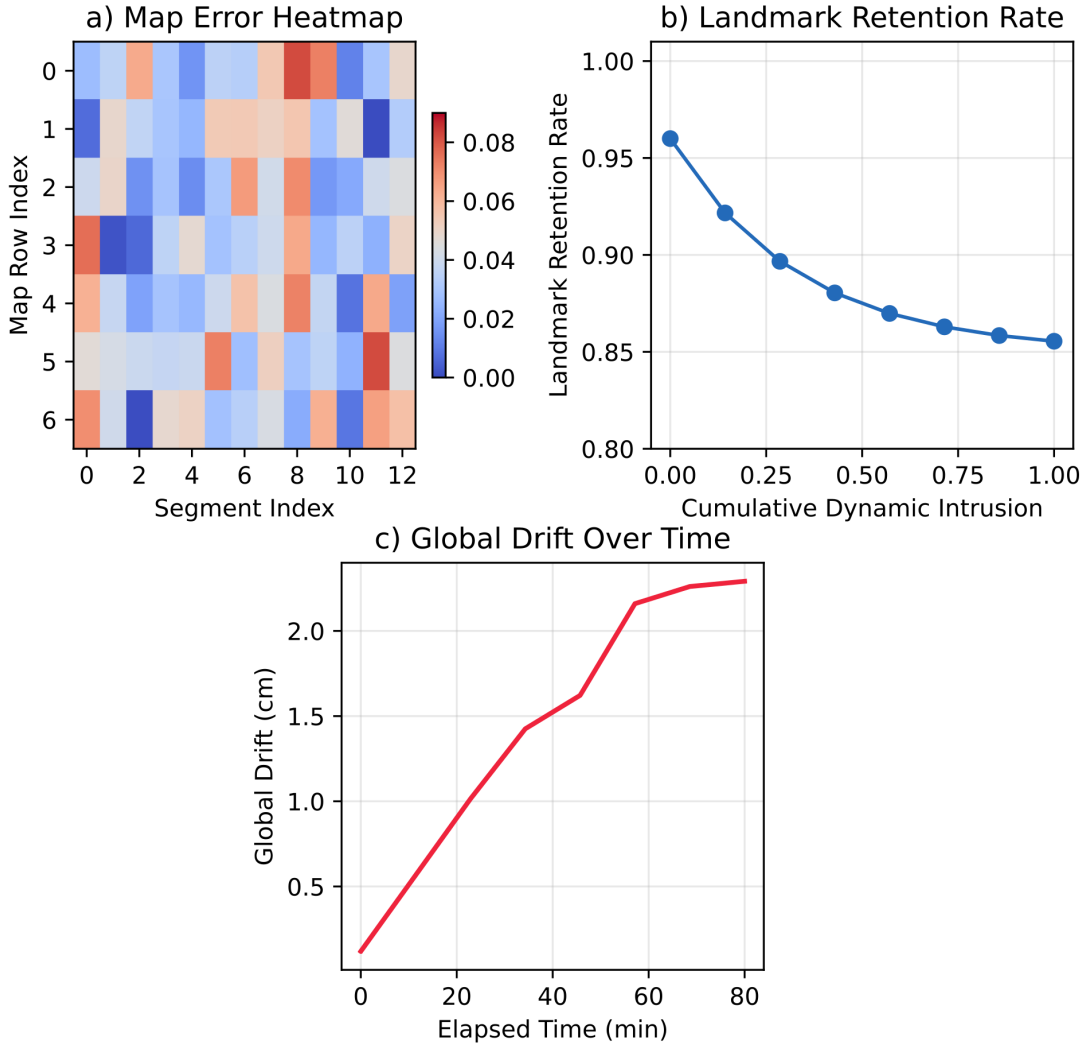

Figure 7. Mapping Accuracy over Sequences: (a) Map error heatmap over extended routes; (b) Landmark retention after dynamic interference; (c) Global drift over long-duration trials.

Ablation and Sensitivity Study

To ascertain the individual and combined contributions of the suggested SLAM architecture, conduct a sensitivity analysis and an all-out ablation investigation. The three primary components of the system under study were backend entropy-gated outlier rejection, temporal dynamic segmentation, and multi-modal feature fusion.

Evaluate on the respective urban and suburban test sets after methodically removing or altering each module's components and parameters.

The system will perform really poorly if cross-modal feature fusion is not performed. Static segmentation accuracy's Segmentation Correctness Index (SCI) fell from 0.93 to 0.81, while dynamic segmentation's fell from 0.90 to 0.73. Simultaneously, the mean trajectory error nearly doubled to 18.5 cm. Due to the lack of adjustable weighting for the sensors, excessive glare and partial occlusion made it impossible to correctly localize these zones. Context-aware sensor fusion is necessary because the misclassification rate of dynamic objects as static characteristics was more than three times higher in the non-fused system than in the fully integrated system.

We also divided the system into separate-frame forecasts without recurrence in order to evaluate the system's space-time continuity. This modification significantly reduced the map's robustness; during a highly dynamic working time, the average completeness of the map decreased from 96% to 83%, and the mean pose drift per 100 meters increased to 0.39%. Notably, ghost objects frequently reappeared in places with heavy pedestrian circulation, particularly at intricate junctions, and the fleeting clutter could no longer be efficiently suppressed [31].

Map integrity and localization stability required the backend's entropy-gated outlier rejection module. The frequency of backend recovery events requiring global re-alignment and state correction would increase eightfold and spurious landmark insertions would increase fivefold in the absence of the aforementioned gating device. Following an occlusion event, the overall error-correction rate likewise declined sharply and never attained 29%, which is far lower than the 62% median of the entire SLAM pipeline.

The optimal mask-entropy threshold is at 0.21, according to sensitivity analysis of the parameters. A lower threshold resulted in excessive feature input pruning, which led to the inclusion of essential but noisy statics. Conversely, a high threshold increased drift and environmental clutter since it was unable to properly filter dynamic objects. It has been demonstrated that shifting the fusion weights away from their optimal placements to give priority to either vision or LiDAR directly increases map divergence and localization errors, particularly in situations like fog, rain, or dusk. It was discovered that the ideal range for the backend optimization window's length was between 160 and 200 frames; in other words, if the window is too big, the system would react slowly to changes in the environment, and if it is too short, loop closure and temporal integration will be hindered.

In-depth Discussion on Method Limitations and Advantages

The success of this SLAM pipeline depends on both the variety and calibration stability of the sensor array because deployment in dense, real-world urban environments demonstrates sensitivity to sensor imbalance, data association errors, and changes in dynamic situations. Cross-domain robotics research has demonstrated that the robustness of the system is mostly dependent to on-the-fly multi-modal fusion and redundancy management since dynamic traffic density produces a high-frequency change in both the number of agents and occlusion [32]. When visual and geometric inputs are less dependable because of heavy clutter or unpredictable sceneries, backend entropy gating is necessary to maintain optimization stability; without it, the error accumulation problem in the same environment has long been a problem for baseline SLAM techniques [33].

However, a number of real-world factors, such as vibration, thermal drift, or particulate accumulation on the sensor optics, may affect the structure of the current system's reliance on offline calibration and spatial alignment; these are all known causes of spatial misregistration in previous vehicle perception applications [34]. The geographically dispersed urban fleet will eventually have a mapping mismatch and infrequent but significant loop-closure errors over a prolonged operational time if it is not continuously recalibrated [35]. Mask recall and local mapping quality have decreased when LiDAR channels or bandwidth are reduced, for instance, from 64 to 16; early field robotics research has also highlighted this vulnerability to sensor downgrades [36].

The likelihood of incorrectly associating object masks with static structures is rising along with data rates and the variety of sensor types, which also increase the hazards of frame loss and temporal jitter [37]. Rare failure occurrences in both low-level feature extraction and high-level dynamic detection will also impact perception quality in extreme weather conditions, such as prolonged rain or urban haze [38]. Even the best segmentation backbones now need to be gradually adjusted or retrained when confronted with anomalous data distributions, as the aforementioned impacts have been examined in domain adaptation research [39].

The map hasn't been updated in a timely manner since a changing environment can be unstable or too complicated. The benchmarking findings for multi-agent SLAM [40] demonstrate that persistent clutter around the dynamic outlier rejection threshold may be disregarded, and map priors can occasionally be slow to update in the face of irreversible changes in the navigation region. Another factor is parameter sensitivity, which means that altering the entropy threshold, fusion weight approach, or temporal smoothing parameter may unintentionally result in local instability, an unusual increase in drift, or a notable increase in mapping delay.

Conclusion

This research introduces a new unified SLAM framework to handle the challenges of dynamic and structurally complex road environments for autonomous driving systems. A recurrent dynamic segmentation algorithm and context-aware entropy-gated optimization have been proposed to outperform current state-of-the-art SLAM systems in terms of positioning accuracy, resilience to high-density mobile objects, and map stability during long-term occlusion and environmental shifts by integrating multiple sensing methods. The mean trajectory error is in the sub-decimeter range, the map completeness is quite high, and transient or spurious features are greatly suppressed even in areas with high traffic, according to a large number of experimental data in urban and suburban deployment scenarios. Thus, it can be verified that it is possible to build a real-time, high-reliability SLAM system with adaptive multi-sensor fusion and strong dynamic object exclusion.

The stability and accuracy of the perception and optimization modules will be guaranteed by architectural design. Dynamic segmentation can swiftly distinguish between fixed and moving objects, and context-aware sensor fusion can enhance the selection of features under deterioration or ambiguity in the observation. In addition to preventing map corruption, backend entropy-based outlier rejection will guarantee that global optimization is carried out on actual, stable environmental structures. According to the aforementioned sensitivity assessments and ablation tests, all technological advancements are necessary to guarantee the system's high accuracy and stability in a setting with several agents and complicated settings.

This work's methodological advancements provide a technical path toward the safe, widespread, and quantitatively reliable autonomy of intelligent cars in contemporary cities. The aforementioned enhancements have added domain-adaptive dynamic modeling, collaborative multi-agent mapping, and automatic online calibration to the core system. As a result, they can now support the development of next-generation mobility platforms and improve both traffic safety and the effectiveness of high-definition map maintenance in an era of ongoing urban development.

Author Contributions

Kamil Budzyński contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Grzegorz Adrian Dziuba contributes to draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chang, J., Dong, N., & Li, D. (2021). A real-time dynamic object segmentation framework for SLAM system in dynamic scenes. IEEE Transactions on Instrumentation and Measurement, 70, 1-9. https://doi.org/10.1109/TIM.2021.3109718

[2] Pan, Y., Hu, K., Cao, H., Kang, H., & Wang, X. (2024). A novel perception and semantic mapping method for robot autonomy in orchards. Computers and Electronics in Agriculture, 219, 108769. https://doi.org/10.1016/j.compag.2024.108769

[3] Li, S., Wang, S., Zhou, Y., Shen, Z., & Li, X. (2022). Tightly coupled integration of GNSS, INS, and LiDAR for vehicle navigation in urban environments. IEEE Internet of Things Journal, 9(24), 24721-24735. https://doi.org/10.1109/JIOT.2022.3194544

[4] Xu, X., Zhang, L., Yang, J., Cao, C., Wang, W., Ran, Y., ... & Luo, M. (2022). A review of multi-sensor fusion slam systems based on 3D LIDAR. Remote Sensing, 14(12), 2835. https://doi.org/10.3390/rs14122835

[5] Pu, H., Luo, J., Wang, G., Huang, T., & Liu, H. (2023). Visual SLAM integration with semantic segmentation and deep learning: A review. IEEE Sensors Journal, 23(19), 22119-22138. https://doi.org/10.1109/JSEN.2023.3306371

[6] Xu, J., Chen, W., Mao, S., Guan, Y., & Zhu, H. (2024, August). P 2-LOAM: LiDAR odometry and mapping with pole-plane landmark. In 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA) (pp. 1-7). IEEE. https://doi.org/10.1109/ICIEA61579.2024.10665090

[7] Chen, K., Oldja, R., Smolyanskiy, N., Birchfield, S., Popov, A., Wehr, D., ... & Pehserl, J. (2020, October). Mvlidarnet: Real-time multi-class scene understanding for autonomous driving using multiple views. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2288-2294). IEEE. https://doi.org/10.1109/IROS45743.2020.9341450

[8] Papadeas, I., Tsochatzidis, L., Amanatiadis, A., & Pratikakis, I. (2021). Real-time semantic image segmentation with deep learning for autonomous driving: A survey. Applied Sciences, 11(19), 8802. https://doi.org/10.3390/app11198802

[9] Wang, X., Zheng, S., Lin, X., & Zhu, F. (2023). Improving RGB-D SLAM accuracy in dynamic environments based on semantic and geometric constraints. Measurement, 217, 113084. https://doi.org/10.1016/j.measurement.2023.113084

[10] He, Y., Li, J., & Liu, J. (2023). Research on GNSS INS & GNSS/INS integrated navigation method for autonomous vehicles: A survey. IEEE Access, 11, 79033-79055. https://doi.org/10.1109/ACCESS.2023.3299290

[11] Saleem, H., Malekian, R., & Munir, H. (2023). Neural network-based recent research developments in SLAM for autonomous ground vehicles: A review. IEEE Sensors Journal, 23(13), 13829-13858. https://doi.org/10.1109/JSEN.2023.3273913

[12] Liu, X., He, Y., Li, J., Yan, R., Li, X., & Huang, H. (2024). A comparative review on enhancing visual simultaneous localization and mapping with deep semantic segmentation. Sensors, 24(11), 3388. https://doi.org/10.3390/s24113388

[13] Dang, X., Rong, Z., & Liang, X. (2021). Sensor fusion-based approach to eliminating moving objects for SLAM in dynamic environments. Sensors, 21(1), 230. https://doi.org/10.3390/s21010230

[14] Yang, L., & Cai, H. (2024). Enhanced visual SLAM for construction robots by efficient integration of dynamic object segmentation and scene semantics. Advanced Engineering Informatics, 59, 102313. https://doi.org/10.1016/j.aei.2023.102313

[15] Hou, L., Xin, L., Li, S. E., Cheng, B., & Wang, W. (2019). Interactive trajectory prediction of surrounding road users for autonomous driving using structural-LSTM network. IEEE Transactions on Intelligent Transportation Systems, 21(11), 4615-4625. https://doi.org/10.1109/TITS.2019.2942089

[16] Cho, H. M., & Kim, E. (2023). Dynamic object-aware visual odometry (VO) estimation based on optical flow matching. IEEE access, 11, 11642-11651. https://doi.org/10.1109/ACCESS.2023.3241961

[17] Chen, W., Zhou, C., Shang, G., Wang, X., Li, Z., Xu, C., & Hu, K. (2022). SLAM overview: from single sensor to heterogeneous fusion. Remote Sensing, 14(23), 6033. https://doi.org/10.3390/rs14236033

[18] Favorskaya, M. N. (2023). Deep learning for visual SLAM: The state-of-the-art and future trends. Electronics, 12(9), 2006. https://doi.org/10.3390/electronics12092006

[19] Ji, Q., Zhang, Z., Chen, Y., & Zheng, E. (2024). Drv-slam: An adaptive real-time semantic visual slam based on instance segmentation toward dynamic environments. Ieee Access, 12, 43827-43837. https://doi.org/10.1109/ACCESS.2024.3379269

[20] Zhou, W., Dong, S., Fang, M., & Yu, L. (2023). CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing. IEEE Transactions on Intelligent Vehicles, 9(1), 1919-1929. https://doi.org/10.1109/TIV.2023.3314527

[21] Seichter, D., Lewandowski, B., Höchemer, D., Wengefeld, T., & Gross, H. M. (2020, October). Multi-task deep learning for depth-based person perception in mobile robotics. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 10497-10504). IEEE. https://doi.org/10.1109/IROS45743.2020.9340870

[22] Li, G., Yu, L., & Fei, S. (2021). A deep-learning real-time visual SLAM system based on multi-task feature extraction network and self-supervised feature points. Measurement, 168, 108403. https://doi.org/10.1016/j.measurement.2020.108403

[23] Samadzadeh, A., & Nickabadi, A. (2023). SRVIO: Super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions. IEEE Transactions on Robotics, 39(4), 2878-2891. https://doi.org/10.1109/TRO.2023.3268591

[24] Wang, K., Zhao, G., & Lu, J. (2024). A deep analysis of visual SLAM methods for highly automated and autonomous vehicles in complex urban environment. IEEE Transactions on Intelligent Transportation Systems, 25(9), 10524-10541. https://doi.org/10.1109/TITS.2024.3379993

[25] Li, Z., & Dong, J. (2022). A framework integrating deeplabV3+, transfer learning, active learning, and incremental learning for mapping building footprints. Remote Sensing, 14(19), 4738. https://doi.org/10.3390/rs14194738

[26] Zhou, Y., Quang, L., Nieto-Granda, C., & Loianno, G. (2024). Coped-advancing multi-robot collaborative perception: A comprehensive dataset in real-world environments. IEEE Robotics and Automation Letters, 9(7), 6416-6423. https://doi.org/10.1109/LRA.2024.3406207

[27] Falaschetti, L., Manoni, L., & Turchetti, C. (2022). A low-rank cnn architecture for real-time semantic segmentation in visual slam applications. IEEE Open Journal of Circuits and Systems, 3, 115-133. https://doi.org/10.1109/OJCAS.2022.3174632

[28] Thomas, H., Zhang, J., & Barfoot, T. D. (2023). The foreseeable future: Self-supervised learning to predict dynamic scenes for indoor navigation. IEEE Transactions on Robotics, 39(6), 4581-4599. https://doi.org/10.1109/TRO.2023.3304239

[29] Păsăreanu, C. S., Mangal, R., Gopinath, D., Getir Yaman, S., Imrie, C., Calinescu, R., & Yu, H. (2023, July). Closed-loop analysis of vision-based autonomous systems: A case study. In International conference on computer aided verification (pp. 289-303). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37706-8_15

[30] Zhang, X., Li, J., Li, Z., Liu, H., Zhou, M., Wang, L., & Zou, Z. (2023). Multi-sensor fusion for autonomous driving (pp. 3-232). Singapore: Springer. https://doi.org/10.1007/978-981-99-3280-1

[31] Ganti, P., & Waslander, S. L. (2019, May). Network uncertainty informed semantic feature selection for visual SLAM. In 2019 16th Conference on Computer and Robot Vision (CRV) (pp. 121-128). IEEE. https://doi.org/10.1109/CRV.2019.00024

[32] Singandhupe, A., & La, H. M. (2019, February). A review of slam techniques and security in autonomous driving. In 2019 third IEEE international conference on robotic computing (IRC) (pp. 602-607). IEEE. https://doi.org/10.1109/IRC.2019.00122

[33] Zhou, Y., Mei, G., Wang, Y., Wan, Y., & Poiesi, F. (2024). Multimodal fusion SLAM with Fourier attention. IEEE Robotics and Automation Letters, 10(2), 1050-1057. https://doi.org/10.1109/LRA.2024.3512252

[34] Chen, W., Chen, S., Leng, J., Wang, J., Guan, Y., Meng, M. Q. H., & Zhang, H. (2024). A review of cloud-edge SLAM: Toward asynchronous collaboration and implicit representation transmission. IEEE Transactions on Intelligent Transportation Systems, 25(11), 15437-15453. https://doi.org/10.1109/TITS.2024.3438165

[35] Park, J., Cho, Y., & Shin, Y. S. (2022). Nonparametric background model-based LiDAR SLAM in highly dynamic urban environments. IEEE Transactions on Intelligent Transportation Systems, 23(12), 24190-24205. https://doi.org/10.1109/TITS.2022.3204917

[36] Ali, M. K., Rajput, A., Shahzad, M., Khan, F., Akhtar, F., & Börner, A. (2019). Multi-sensor depth fusion framework for real-time 3D reconstruction. Ieee Access, 7, 136471-136480. https://doi.org/10.1109/ACCESS.2019.2942375

[37] Zhong, M., Hong, C., Jia, Z., Wang, C., & Wang, Z. (2024). DynaTM-SLAM: Fast filtering of dynamic feature points and object-based localization in dynamic indoor environments. Robotics and Autonomous Systems, 174, 104634. https://doi.org/10.1016/j.robot.2024.104634

[38] Wang, K., Zhang, L., Xia, Q., Pu, L., & Chen, J. (2022). Cross-domain learning using optimized pseudo labels: Toward adaptive car detection in different weather conditions and urban cities. Neural Computing and Applications, 34(6), 4519-4529. https://doi.org/10.1007/s00521-021-06609-z

[39] Wei, S., & Li, Z. (2023). An RGB-D SLAM algorithm based on adaptive semantic segmentation in dynamic environment. Journal of Real-Time Image Processing, 20(5), 85. https://doi.org/10.1007/s11554-023-01343-2

[40] Drew, D. S. (2021). Multi-agent systems for search and rescue applications. Current Robotics Reports, 2(2), 189-200. https://doi.org/10.1007/s43154-021-00048-3