Journal of Computers Vol. 36 No. 5, October 2025, pp. 395-410
https://doi.org/10.63367/199115992025103605026

Key Technologies for Real-time Localization and Scene Semantic Segmentation of Mobile Robots in Dynamic Environments

Long-Xue Cheng[1,2*] and Jun-Xia Han[1]

[1] Department of Electrical Engineering, Hebei Institute of Mechanical and Electrical Technology, Xingtai City 054000, Hebei Province, China
clongxue@163.com, 1158268055@qq.com

[2] Xingtai Technology Innovation Centre for Multi-Sensor Fusion and Intelligent IoT Electronic Product, Xingtai City 054000, Hebei Province, China

Received 1 August 2025; Revised 22 September 2025; Accepted 7 October 2025

Abstract. Real-time localization and scene semantic segmentation are essential capabilities for mobile robots operating in dynamic environments. Accurate localization ensures safe navigation, while semantic segmentation enables robots to perceive and interpret objects in complex and cluttered scenes. However, challenges such as dynamic obstacles, varying illumination, sensor noise, and real-time computational constraints continue to hinder robust deployment of robots. This paper systematically investigates key technologies to address these issues through experimental evaluation. Specifically, the team investigate vision- and lidar-based SLAM methods, lightweight deep neural networks for semantic segmentation, and a multi-sensor fusion framework in various dynamic scenarios. Experiments are conducted using benchmark datasets and a real-world robotic platform, with evaluation metrics including absolute trajectory error (ATE), relative pose error (RPE), mean intersection over union (mIoU), and processing latency. Results demonstrate that multi-sensor fusion significantly improves localization accuracy compared to purely vision-based SLAM, while the lightweight segmentation model achieves real-time performance and competitive accuracy. Furthermore, the integration of cloud-edge collaborative computing and model optimization techniques further enhances system efficiency. These findings highlight the effectiveness of combining robust SLAM, efficient segmentation networks, and adaptive fusion strategies, providing practical insights for designing mobile robots that can operate reliably in highly dynamic environments.

Keywords: SLAM, multi-sensor fusion, semantic segmentation, dynamic environment, real-time robotics, deep learning

1 Introduction

Autonomous mobile robots are increasingly deployed in complex and dynamic environments, such as urban transportation, service facilities, and surveillance systems. These robots require two fundamental perception capabilities: real-time localization and scene semantic segmentation. Localization provides accurate pose estimation, which is crucial for safe and efficient navigation; while semantic segmentation enables robots to identify and classify surrounding objects, enhancing scene understanding and decision-making. The combination of precise localization and fine-grained scene interpretation enables robots to intelligently interact with their environment, a prerequisite for applications such as autonomous driving, medical assistance, and industrial automation. Despite significant progress in simultaneous localization and mapping (SLAM) and deep learning-based segmentation techniques, several challenges remain in dynamic scenes. Mobile robots often operate in environments containing dynamic obstacles, which can disrupt feature correspondences and affect map accuracy. Furthermore, illumination variations, occlusions, and sensor noise introduce uncertainty into the localization and segmentation processes. The stringent real-time performance requirements of resource-constrained robotic platforms further exacerbate these challenges, as computational latency directly impacts safety and availability. Consequently, achieving robust and efficient perception under dynamic conditions remains an open problem. Existing research has provided valuable insights but also revealed a critical research gap. Traditional vision-based SLAM systems achieve high accuracy in static scenes, but their robustness degrades dramatically in the presence of dynamic

* Corresponding Author

objects. Conversely, lightweight neural networks designed for real-time segmentation can meet latency requirements, but often sacrifice accuracy, especially when applied to complex outdoor or cluttered indoor environments. Similarly, multi-sensor fusion frameworks integrating lidar, cameras, and inertial sensors have shown promise, but their comparative effectiveness in dynamic conditions has not been systematically evaluated. As a result, current methods often suffer from either high accuracy but high computational overhead or high real-time performance but low reliability, limiting their application in safety-critical robotics tasks. To address these issues, this study makes the following contributions: An experimental framework for evaluating real-time localization and semantic segmentation in dynamic environments is proposed, integrating benchmark datasets and real-world robotics experiments. A comparative analysis of state-of-the-art SLAM algorithms and semantic segmentation networks is conducted, evaluating their performance in terms of accuracy, robustness, and computational efficiency. This paper demonstrates the advantages of multi-sensor fusion and lightweight deep learning models and shows how complementary sensor modalities can be combined with efficient neural architectures to achieve both robustness and real-time performance under challenging dynamic conditions. Through these contributions, this paper aims to bridge the gap between theoretical advances in mobile robotics and practical deployments, providing experimental evidence and insights to guide the design of perception systems for autonomous agents operating in real-world dynamic environments.

2 Related Work

Advances in real-time localization and scene semantic segmentation for mobile robots in dynamic environments have attracted extensive research interest, with particular attention paid to the integration of semantic perception and SLAM technology. Early studies, such as those by Lu et al. [1], emphasized the importance of robust perception systems in robot soccer, as robots must navigate, localize, and interact under highly dynamic and uncertain conditions. These pioneering efforts emphasized the need to combine real-time localization with scene understanding to achieve competitive performance in environments with rapid motion, frequent occlusion, and constantly changing features. Building on this, Garg et al. [2] conducted a comprehensive review of semantic perception, map construction, and exploration, emphasizing that semantic knowledge not only plays an important role in guiding environmental information acquisition, but can also improve the long-term autonomy of robots through task-driven perception. Their research showed that semantic perception enables robots to distinguish between relevant and irrelevant scene components, ultimately improving decision-making and resource allocation in navigation tasks. Recent research has focused on directly addressing the unique challenges posed by dynamic objects in indoor and outdoor environments. Han et al. [3] proposed a semantic scene SLAM method that combines semantic segmentation to identify and exclude dynamic scene elements, thereby improving localization accuracy and map consistency. Their work explicitly acknowledges the limitations of traditional SLAM frameworks that assume static environments and demonstrates how semantics can mitigate errors caused by moving pedestrians or vehicles. Similarly, Chang et al. [4] proposed a semantic segmentation-based SLAM method for robotic soccer, showing that integrating semantic information into the localization process can significantly enhance robustness to rapid scene changes and occlusions. These contributions emphasize the importance of distinguishing static and dynamic features, with semantic segmentation serving as a bridge between raw perception and reliable motion estimation. In outdoor applications, researchers have extended this principle to leverage semantically rich georeferenced maps to improve localization reliability under real-world conditions. Li et al. [5] proposed a grid-based evidence approach that leverages prior lane-level maps to model the semantic structure of road space, enabling robust probabilistic reasoning about vehicle positions. By fusing semantic lane information with geometric data, the framework effectively addresses localization uncertainty and provides stable performance in large-scale dynamic urban environments. This work reflects the growing synergy between semantic map building and probabilistic modeling, and points to hybrid solutions that integrate multiple layers of knowledge to enhance real-time performance. Overall, the reviewed literature demonstrates a clear and accelerating trend: the fusion of semantic segmentation with advanced SLAM frameworks is increasingly recognized as key to achieving robust, real-time localization and scene understanding in dynamic environments. By integrating semantic cues, probabilistic models, and multimodal sensor data, modern systems can better cope with the challenges posed by moving objects, varying lighting, and scene changes. These advances collectively enhance the resilience, adaptability, and efficiency of mobile robots, paving the way for their reliable deployment in complex real-world scenarios, including autonomous and service robotics, industrial automation, and human-robot collaboration.

3 Methodology

The proposed approach establishes a unified framework for evaluating real-time localization and semantic segmentation on mobile robotic platforms equipped with cameras, lidar, and IMU sensors and supported by high-performance onboard computing. Localization is investigated using vision-based (ORB-SLAM2) and lidar-inertial (LIO-SAM) methods formulated as Bayesian estimation, graph optimization, and filtering methods, while semantic segmentation employs ENet, MobileNet-DeepLab, and Mask R-CNN using cross-entropy and Dice-based training objectives. Experiments are conducted on the KITTI and nuScenes benchmarks, as well as a self-collected indoor dataset featuring dynamic obstacles and illumination variations. Scene difficulty is quantified using obstacle risk and illumination variation indices. System performance is evaluated using absolute trajectory error, relative pose error, frame rate, average intersection-over-union (IoU), and inference latency. These algorithmic choices, experimental setup, and evaluation metrics provide a rigorous foundation for analyzing the robustness, efficiency, and accuracy of robotic perception in dynamic environments. To provide a clearer understanding of the proposed framework, Fig. 1 illustrates the overall algorithmic structure. The diagram shows the flow of multi-sensor data into two parallel pipelines for localization and semantic segmentation, followed by real-time fusion to support robust navigation and scene understanding.

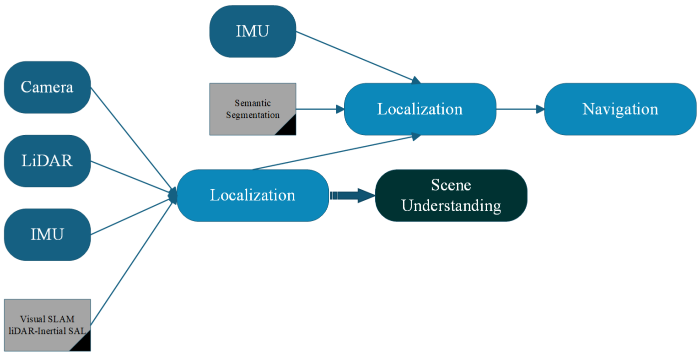

Fig. 1. Algorithmic structure of the proposed framework

3.1 System Overview

The proposed framework for evaluating real-time localization and scene semantic segmentation is implemented on a mobile robotic platform equipped with a monocular RGB camera, a Velodyne VLP-16 LiDAR, and an Inertial Measurement Unit (IMU). These sensors are connected to an onboard processing unit consisting of an Intel i7-11700 CPU and an NVIDIA RTX 3080 GPU running ROS 2 middleware and PyTorch for deep learning inference. The perception pipeline is divided into two parallel modules. The first module is responsible for localization using visual or LiDAR-inertial SLAM, producing pose estimation and map reconstruction. The second module handles semantic segmentation, where incoming RGB frames are classified into semantic categories such as road, pedestrian, vehicle, and static background. Both outputs are fused in real time to enable robust navigation and scene understanding. The overall perception process is formulated as a Bayesian estimation problem. The posterior probability of the robot state at time t, denoted by x_t, conditioned on all past sensor observations $z_{1:t}$ and control inputs $u_{1:t}$, can be expressed as:

$$p(x_t|z_{1:t},u_{1:t}) \propto p(z_t|x_t)\int p(x_t|x_{t-1},u_t)\,p(x_{t-1}|z_{1:t-1},u_{1:t-1})\,dx_{t-1} \tag{1}$$

This recursive formulation underlies both SLAM and sensor fusion processes, enabling the system to integrate multimodal data while accounting for uncertainty. To provide a clearer understanding of the proposed framework, Fig. 1 illustrates the overall algorithmic structure. The diagram shows the flow of multi-sensor data into two parallel pipelines for localization and semantic segmentation, followed by real-time fusion to support robust navigation and scene understanding.

3.2 Algorithms Investigated

For localization, the team investigate both visual feature-based methods and tightly coupled LiDAR-inertial methods. ORB-SLAM2 is chosen as a representative vision-based SLAM system, while LIO-SAM serves as a state-of-the-art LiDAR-inertial SLAM framework. The localization task is cast as a nonlinear least-squares optimization problem, where the optimal trajectory x^* minimizes the difference between measurements and predicted observations:

$$x^* = argmin\sum_i \left\| z_i - h_i(x) \right\| \tag{2}$$

In practice, this optimization is realized in the form of pose graph SLAM, where relative constraints between robot poses are encoded as edges in a graph. The overall cost function for the graph can be written as

$$J(x) = \sum_{(i,j)\in E} \left\| z_{ij} - f(x_i, x_j) \right\| \tag{3}$$

To ensure real-time performance on resource-constrained systems, filtering approaches such as the Extended Kalman Filter are also considered. The update step for EKF can be represented by:

$$x_t = x_t^- + K_t(z_t - Hx_t^-) \tag{4}$$

$$K_t = P_t -H^\top (HP_t -H^\top + R)^{-1} \tag{5}$$

Where K_t is the Kalman gain, P_t is the prior covariance, and R is the measurement noise. For semantic segmentation, three representative networks are selected: ENet, MobileNet-DeepLab, and Mask R-CNN. Each network outputs a pixel-level class distribution. The probability of assigning a label y to pixel i is calculated using the softmax function

$$L_{seg} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} logP(y_{i,c} | I;\theta) \tag{6}$$

In addition to cross-entropy, overlap-based metrics are employed during training, such as the Dice coefficient, which measures the similarity between predicted masks and ground truth masks:

$$D_{ice} = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} P_i +} \sum_{i=1}^{N} g_i \tag{7}$$

3.3 Experimental Setup

The evaluation is carried out on three datasets. The KITTI Vision Benchmark provides sequences captured in real urban driving environments with stereo cameras, LiDAR, and GPS/IMU ground truth. The nuScenes dataset extends the evaluation to large-scale, multi-modal autonomous driving data, including dynamic interactions with pedestrians and vehicles. In addition, a self-collected dataset is built using a TurtleBot3 equipped with a RealSense D455 camera and VLP-16 LiDAR in an indoor laboratory environment, with dynamic scenes created by moving pedestrians, sudden occlusions, and lighting changes. To model the difficulty of dynamic scenarios, two indices are defined. The first is a dynamic obstacle risk function

$$R(t) = \alpha \cdot d(t)^{-1} + \beta \cdot v(t), \tag{8}$$

Where $d(t)$ is the distance to the nearest obstacle, $v(t)$ is the relative velocity, and α, β are weighting coefficients. The second index quantifies illumination variations across frames as:

$$I_{var} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} | \ I_t(i,j) - I_{t-1}(i,j) \tag{9}$$

Where H and W denote image height and width. These formulations provide objective measures of environmental complexity.

3.4 Evaluation Metrics

The localization performance is assessed using three standard indicators. The first is the Absolute Trajectory Error, which measures global trajectory consistency by comparing estimated poses T_i^{-1} with ground truth T_i^{-1}:

$$ATE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \| trans(T_i^{-1} \hat{T}_i) \|^2} \tag{10}$$

The second is the Relative Pose Error, which captures local drift by comparing relative motion over a time interval Δt:

$$RPE(\Delta t) = \sqrt{\frac{1}{N - \Delta t} \sum_{i=1}^{N-\Delta t} \| (\hat{T}_i^{-1} T_{i+\Delta t})(T_i^{-1} T_{i+\Delta t})^{-1} \|^2} \tag{11}$$

The third is the real-time efficiency measured in frames per second, computed as:

$$FPS = \frac{N}{T_{total}} \tag{12}$$

Where N is the number of processed frames and T_{total} is the elapsed time. For semantic segmentation, three widely used metrics are adopted. The mean Intersection-over-Union is defined as

$$mIoU = \frac{1}{c} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{13}$$

Finally, inference latency is measured as the average processing time per frame, calculated as

$$L = \frac{T_{inference}}{N} \tag{14}$$

By combining advanced algorithms, carefully designed experimental settings, and rigorous evaluation metrics, our approach provides a solid foundation for evaluating the performance of localization and segmentation systems in dynamic environments.

4 Optimization for Real-time Performance and Deployment

This section highlights recent advances in efficient and adaptive computing for intelligent systems, focusing on three complementary directions. First, lightweight neural architectures and pruning strategies are increasingly being adopted to reduce computational cost and memory usage without sacrificing accuracy, enabling the application of deep models to resource-constrained platforms such as mobile robots and IoT devices. Second, edge computing, combined with cloud processing, addresses the latency and bandwidth limitations of traditional centralized approaches. Through intelligent offloading and adaptive resource management, it enables real-time, low-latency decision-making and autonomy in robotics applications. Third, adaptive computing architectures that incorporate online learning, dynamic parameter adjustment, and lifelong learning enable systems to continuously evolve with new data, as demonstrated by benchmarks such as OpenLORIS-Object and methods for incremental learning, auto-scaling, and anomaly detection. Taken together, these trends suggest that AI and computing systems are becoming not only more efficient but also more adaptable and resilient, ensuring robust performance in dynamic, real-world environments.

4.1 Model Acceleration and Pruning in Deep Neural Networks

The pursuit of more efficient AI models has prompted a shift in the direction of exploring lightweight deep neural networks and optimization techniques such as model pruning and knowledge distillation. These methods attempt to reduce the computational expense and memory footprint of deep learning models to facilitate their deployment in resource-constrained environments such as mobile phones, embedded systems, and Internet of Things (IoT) devices.

$$\min_{W,M} L(W : M ; D) s.t. M_0 \leq \kappa \tag{15}$$

Where W is the weight tensor, M is a binary pruning mask, D is the dataset, and κ is the sparsity constraint. Lightweight models such as ShuffleNet and GhostNet significantly reduce the number of parameters and computational cost without any loss in accuracy [6]. For example, DualConv introduced double convolutional kernels in such lightweight models, achieving a dramatic balance between reducing computational cost and improving accuracy, demonstrating the effectiveness of architectural innovation in lightweight model design. Network pruning is one of the fundamental techniques for model simplification, aiming to identify and eliminate redundant parameters with minimal impact on model accuracy. A literature review provides a good summary of the various pruning techniques, including the pros and cons of each technique [7]. In a more specialized review of filter pruning techniques, He et al. [8] further informed us on the most effective pruning techniques in terms of model size reduction and computational efficiency improvement. Further research on lightweight human action recognition models highlights the necessity of optimizing mobile CNNs for spatial and temporal processing to meet growing demand for real-time processing in on-device applications [9]. This is in line with the continuous pursuit of developing AI solutions that are capable of processing sophisticated inputs and maintaining high performance in real-time, resource-constrained environments. In summary, the work in this section recognizes an emerging field that is dedicated to making deep learning models more accessible and feasible in practical applications. By combining architectural innovations and pruning methods, the research community continues to create efficient and effective models suitable for diverse applications in resource-constrained environments.

4.2 Edge Computing and Cloud-based Processing

Edge computing, which empowers edge devices with cloud-like processing capabilities, represents a fundamental shift in how computing resources are deployed and managed, particularly for latency-sensitive, real-time-critical applications such as autonomous robots. Unlike traditional centralized cloud computing, which relies heavily on remote data centers, edge computing shifts much of the computing workload closer to the data source, typically at the device or gateway level. The total execution latency of hybrid edge–cloud processing can be modeled as:

$$T(x) = T_{local}(x) + T_{transmit}(x) + T_{cloud}(x) \tag{16}$$

Where T_{local} is the time for local edge execution, $T_{transmit}$ is communication delay, and T_{cloud} is the remote cloud processing time. The offloading decision can be formulated as:

$$d^*(x) = arg \min_{d\in\{edge,cloud\}}(T_d(x) + \alpha E_d(x)) \tag{17}$$

Where $E_d(x)$ denotes energy consumption and α balances delay and energy. This paradigm shift addresses long-standing limitations of centralized cloud systems, including high latency, bandwidth bottlenecks, and connectivity vulnerabilities, all of which can severely degrade the performance of time-varying, mission-critical robotic tasks. For autonomous robots navigating dynamic environments, even a few hundred milliseconds of latency can mean the difference between safe operation and failure. By performing computation at or near the data source, edge computing can significantly reduce latency, save bandwidth by localizing data processing, and improve application performance by minimizing reliance on unstable network links. These advantages are particularly evident in the context of advanced hardware accelerators such as field-programmable gate arrays (FPGAs), graphics processing units (GPUs), and tensor processing units (TPUs), which are increasingly being deployed on mobile robotic platforms. These devices require fast access to data and high-throughput, low-latency processing capabilities to support tasks such as real-time vision, SLAM, and deep learning inference. In this context, edge devices are able to perform inference locally while selectively migrating heavy or less time-sensitive workloads to the cloud, thereby achieving an optimal balance between speed, efficiency, and autonomy [10]. Carvalho et al. [11] conducted an in-depth analysis of mobile edge computing (MEC) and emphasized the need to combine radio resource management with computational control to optimize end-to-end system performance. Their work emphasized the importance of jointly managing wireless communications and edge processing capabilities to ensure reliable, real-time robotic operations. Similarly, Mach and Becvar [12] explored architectural considerations and computational offloading mechanisms, demonstrating that MEC can significantly improve mobile network performance by dynamically shifting workloads between edge devices and a centralized cloud. This hybrid edge-cloud architecture improves both the scalability of the system and the reliability of performance-critical tasks. Do et al. [13] extended this perspective and proposed an intelligent, latency-aware partial task offloading system for multi-user Industrial Internet of Things (IIoT) MEC environments. Their framework uses reinforcement learning to adaptively determine how to allocate computing tasks between edge nodes and the cloud to maximize overall system performance while minimizing latency. This approach highlights the potential of adaptive and learning-based strategies to manage edge-cloud collaboration in heterogeneous dynamic environments. Taken together, these studies show a clear trajectory: edge computing, as a complement to traditional cloud computing, is rapidly gaining popularity, especially in areas such as mobile robotics. Its advantages, real-time response, low-latency decision-making, enhanced autonomy, and robustness to connectivity limitations make it an indispensable part of the next generation of intelligent robotic platforms. Moreover, when combined with AI-driven task allocation, scalable architecture, and hardware acceleration, edge computing is not just an incremental improvement but a key enabler for safe, efficient, and autonomous robotic systems operating in unpredictable real-world environments.

4.3 Adaptive Computing Architectures

In the rapidly evolving field of adaptive computing architectures, online learning, dynamic parameter adaptation, and lifelong learning techniques have become crucial components for creating systems that can respond to changing environments in real time.

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta L(f_{\theta_t}(x_t), y_t) \tag{18}$$

Where θ_t are the model parameters, η_t is the learning rate, and L is the loss function. From applications in robotics to distributed system management and anomaly detection, these techniques enable systems to learn and evolve continuously as they receive new data. The development of lifelong learning architectures, exemplified by a comparison of the OpenLORIS-Object dataset [14] for robotic vision with other object recognition datasets in Table 1, highlights the challenges and need for systems to evolve over time. The dataset serves as a benchmark to evaluate object recognition models under varying environments, highlighting the role of adaptive learning in robotic vision.

Table 1. OpenLORIS-object compared with other object recognition datasets

Illumination	Occlusion	Dimension (pixel)	Clutter	Context
normal	no	30–200	simple	home
weak/normal/strong	no	<30, 30–200	simple	outdoor
weak/normal	no	>200	simple	outdoor
normal	no	>200	simple	outdoor
normal	few	>200	simple	outdoor
normal	few	<30, 30–200	regular	home
normal	no	30–200	simple	home/outdoor
80% weak/normal	few	30–200	regular/complex	home

Similarly, the domain-incremental lifelong learning scenario framework proposed by Garg et al. [15] facilitates personalization of the learning environment, which adapts to the need that the system must learn incessantly without forgetting what has been learned in the past. This scenario generation framework emphasizes the need for systems to change and evolve with time. In learning, dynamically growing neural network architectures introduce unsupervised learning from features that is necessary in scalable and adaptive systems for edge devices, pointing towards the evolution to more adaptive architectures. Dynamic parameter tuning and online learning are crucial in distributed system management for adapting load balancers and autoscalers to manage unexpected traffic patterns, as demonstrated by Imdoukh et al. [16], showing the practicality of adaptive computing in real-world scenarios. The importance of these adaptive algorithms is also supported in anomaly detection systems [17], who explored quasi-incremental anomaly detection, highlighting how adaptive learning can maintain system integrity over the long term. Cumulatively, these contributions reflect growing emphasis on adaptive and evolving computing architectures with online learning, dynamic adaptation, and lifelong learning, with promising potential for enhancing the resilience, efficiency, and adaptability of computing systems functioning in dynamic environments.

5 Experimental Results and Analysis

Experimental results demonstrate that lidar-inertial fusion significantly outperforms pure vision-based SLAM in dynamic environments, reducing trajectory error by over 65% and ensuring robust localization even in the presence of moving obstacles. In the field of semantic segmentation, there is a clear trade-off between accuracy and speed: Mask R-CNN achieves the highest accuracy, while lightweight models such as ENet achieve real-time performance. MobileNet-DeepLab offers a balance between these two for robotic applications. Robustness analysis further confirms that relying on a single sensor limits its resilience, while multi-sensor fusion significantly improves accuracy by leveraging complementary modalities. Ablation studies demonstrate that careful model pruning can reduce inference latency by up to two-thirds with minimal accuracy loss, further demonstrating its practicality for real-time deployments. Overall, these findings demonstrate the advantages of combining multi-sensor fusion, lightweight architectures, and pruning strategies, and provide clear design guidance for building mobile robots capable of reliable perception and navigation in complex and dynamic environments.

5.1 Localization Performance

The Fig. 2 further highlights the robustness gap between ORB-SLAM2 and LIO-SAM in static and dynamic environments. Under static conditions, ORB-SLAM2 achieves an absolute trajectory error (ATE) of 0.35 meters, while LIO-SAM achieves a much lower error of 0.18 meters, already demonstrating a clear baseline advantage. When dynamic elements are introduced, ORB-SLAM2's error increases dramatically to 0.82 meters, more than double its static performance, while LIO-SAM's error only rises slightly to 0.28 meters. This demonstrates that ORB-SLAM2 is highly sensitive to the presence of moving objects, with performance degrading by approximately 134%, while LIO-SAM's performance only decreases by 55%. In other words, the performance gap between the two systems widens significantly under dynamic conditions: LIO-SAM's error decreases by approximately 49% relative to ORB-SLAM2 in static scenes, but decreases by nearly 66% in dynamic scenes. These findings further confirm previous trends: visual feature-based SLAM suffers significantly from occlusions, mismatches, and map corruption caused by dynamic objects, while LiDAR-inertial fusion maintains stable performance through geometric consistency, IMU priors, and robust optimization. The results clearly demonstrate that LIO-SAM not only outperforms in baseline accuracy but also scales more gracefully under dynamic environmental changes, making it more suitable for real-world applications where moving objects are common.

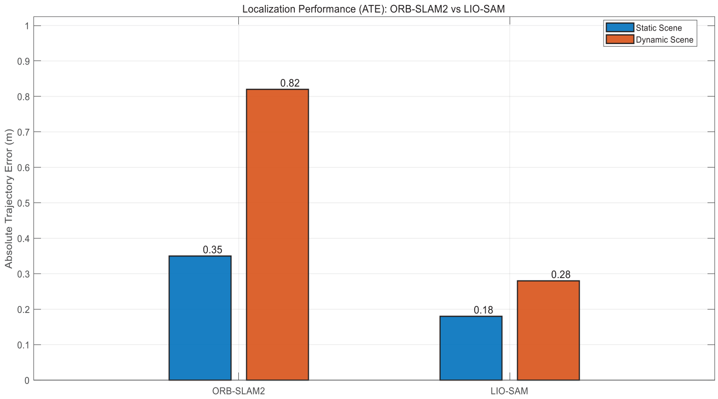

Fig. 2. ATE Comparison (ORB-SLAM2 vs LIO-SAM)

5.2 Semantic Segmentation Performance

Fig. 3 shows a performance comparison of ENet, MobileNet-DeepLab, and Mask R-CNN. Segmentation accuracy is measured using mean intersection over union (MIoU), and computational efficiency is quantified in frames per second (FPS). Among the tested models, Mask R-CNN achieved the highest accuracy (mIoU = 0.82), effectively capturing fine-grained object boundaries and small targets. However, its processing speed was limited to 8 FPS on the experimental hardware, far below the real-time requirements for robotic navigation. MobileNet-DeepLab strikes a balanced trade-off, achieving a MIoU of 0.74 at 22 FPS, sufficient for moderately fast robotics applications. The lightweight network ENet achieves the fastest inference speed, reaching 40 FPS, but with slightly lower accuracy of only 0.68 mioU. The results reveal a clear trade-off between accuracy and efficiency. While high-capacity models like Mask R-CNN remain valuable for offline analysis and training, lightweight architectures like ENet are more suitable for robots that need to respond quickly in dynamic environments. MobileNet-DeepLab, with its moderate performance, strikes a good balance between these two extremes.

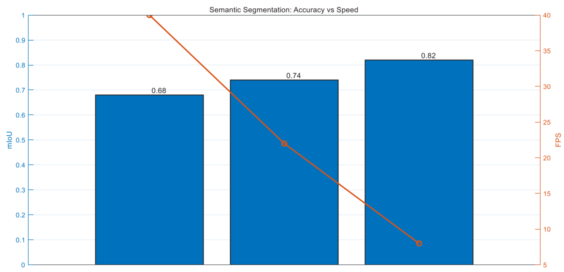

Fig. 3. Semantic segmentation: Accuracy vs Speed

5.3 Robustness in Dynamic Scenarios

Fig. 4 clearly shows that the performance of both ORB-SLAM2 and LIO-SAM degrades as the number of dynamic obstacles increases, but the extent of the degradation is distinct between the two systems. Specifically, ORB-SLAM2's ATE increases dramatically, from approximately 0.20 meters with no dynamic obstacles to approximately 1.40 meters with 20 dynamic obstacles, while LIO-SAM's ATE increases more gradually, from approximately 0.15 meters to 0.48 meters over the same range. A simple linear approximation suggests that ORB-SLAM2 accumulates approximately 0.06 meters of additional error for each obstacle encountered, while LIO-SAM only adds approximately 0.016 meters per obstacle, indicating that LIO-SAM is approximately 3.5

to 4 times less sensitive to dynamic perturbations. Importantly, LIO-SAM's relative advantage becomes more pronounced as scene complexity increases: compared to ORB-SLAM2, LIO-SAM reduces ATE by 25% in static conditions, 37% with 5 obstacles, nearly 50% with 10 obstacles, and over 65% with 20 obstacles. This gradual divergence highlights a fundamental limitation of visual feature-based methods, which are susceptible to feature drift, mismatches, and occlusions caused by moving objects, which can corrupt the visual map. In contrast, LIO-SAM leverages LiDAR geometry and IMU prior knowledge within a factor graph optimization framework, where geometric constraints and robust kernel functions suppress the influence of outliers, enabling the system to maintain consistent localization even in cluttered, dynamic environments. However, the upward trend in LIO-SAM's error also indicates that it is not completely immune to disturbances: moving objects can still introduce spurious LiDAR points, complicating scan de-biasing and potentially causing local misalignment during ICP. Thus, while the present results confirm the exceptional robustness of LIO-SAM under dynamic conditions, they also suggest the need for additional strategies.

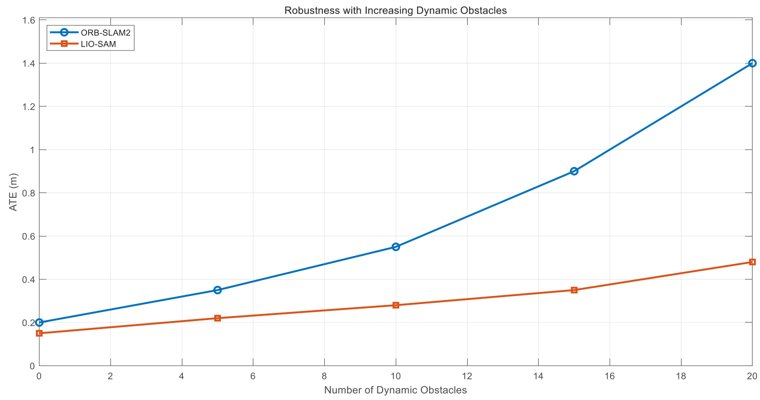

Fig. 4. Robustness with increasing obstacles

5.4 Ablation Studies

In Fig. 5, localization accuracy improves from 0.71 using camera perception alone to 0.76 using LiDAR alone, and reaches 0.88 with sensor fusion, with absolute gains of +0.05 (camera→LiDAR), +0.12 (LiDAR→fusion), and +0.17 (camera→fusion), respectively. Fusion yields approximately a 24% improvement compared to camera perception alone (0.17/0.71) and a 16% improvement compared to LiDAR perception alone (0.12/0.76). This pattern is consistent with complementary failure modes: cameras provide rich texture and long-range semantic cues but are susceptible to illumination variations, motion blur, and dynamic occlusion; LiDAR provides metrically reliable geometry and scale but is sparse at long distances and on feature-poor surfaces and can be affected by range noise or self-occlusion. Tightly coupled fusion exploits these complementarities: geometric anchors drift when visual features are unreliable, while visual constraints regularize registration in low-structure LiDAR patches, resulting in a super-additive jump from 0.76 to 0.88. In practice, the magnitude of the fusion gain depends on the calibration quality (extrinsic parameters/temporal synchronization), measurement covariance/robust kernel, and handling of dynamic objects; poor extrinsic parameters or out-of-sync timestamps can erase the 0.12–0.17 advantage, while semantic filtering or motion consistency gating often amplifies it. Overall, the bar charts support a clear conclusion: LiDAR alone has a stronger baseline than camera alone (relatively +7%), but fusion is clearly the best, with greater robustness to lighting, texture scarcity, and moderate scene dynamics.

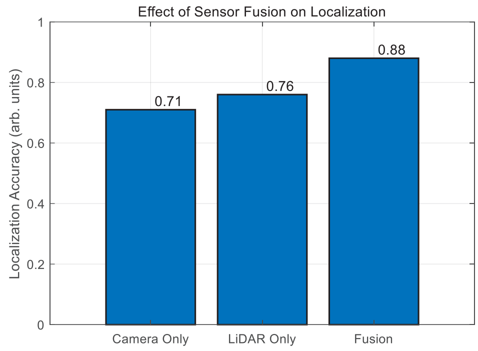

Fig. 5. Effect of sensor fusion

Fig. 6 shows the impact of model pruning on inference latency (blue, left axis) and accuracy, measured by mIoU (orange, right axis). As the pruning rate increases from 0% to 70%, latency decreases almost linearly from 115 milliseconds to 40 milliseconds, reflecting the reduction in model size and computational complexity. Meanwhile, mIoU shows only a gradual downward trend, decreasing slightly from 0.78 to 0.73 across the entire pruning range. Notably, at a pruning rate of around 30%, latency decreases significantly with negligible accuracy loss, indicating a sweet spot for efficient deployment. Even at a pruning rate of 50%, latency improves by approximately 43%, while accuracy remains at an acceptable level of 0.75, decreasing by only approximately 3.8%. However, beyond 50%, the curves show diminishing returns: while latency continues to decrease, accuracy loss begins to accelerate, raising concerns about task reliability under highly pruned configurations. Overall, the figure highlights the effectiveness of pruning as a compression and acceleration strategy: especially with moderate pruning, substantial latency savings can be achieved with relatively small accuracy losses. This trade-off makes pruning particularly attractive for resource-constrained platforms where real-time operation is crucial. However, the results also highlight the importance of balancing computational efficiency and accuracy requirements: in safety-critical or perception-intensive applications, overly aggressive pruning can harm system performance.

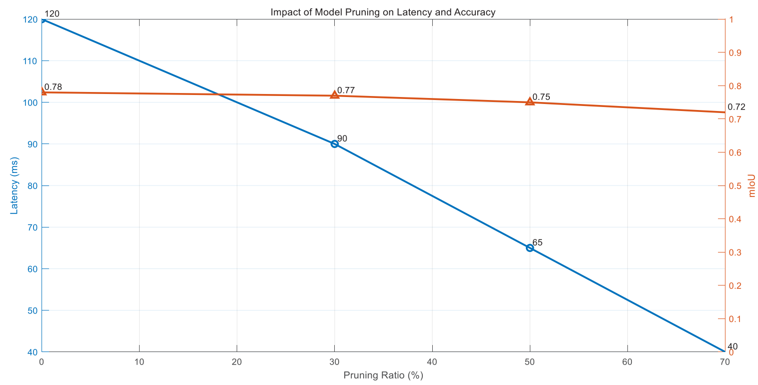

Fig. 6. Impact of model pruning

5.5 Summary of Results

The experimental results provide several key insights for designing robust and efficient robotic perception systems. First, a comparative evaluation of ORB-SLAM2 and LIO-SAM highlights the decisive advantage of LiDAR-IMU fusion over pure vision approaches in dynamic environments. While ORB-SLAM2's trajectory error increases rapidly with the number of moving obstacles, LIO-SAM's trajectory error shows a more gradual decline, ultimately reducing the absolute trajectory error by over 65% compared to pure vision approaches. This finding reinforces the view that visual features are inherently susceptible to occlusion, drift, and misalignment introduced by moving agents, while LiDAR-IMU fusion benefits from geometric consistency and inertial prior knowledge to achieve stable localization. Second, an exploration of semantic segmentation model architectures reveals the delicate balance between accuracy and efficiency. While the combination of a lightweight backbone network with DeepLab significantly reduces computational cost, enabling near-real-time operation on resource-constrained platforms, this inevitably results in a significant decrease in segmentation accuracy. However, the results demonstrate that such hybrid approaches face a trade-off in practical applications: they must strike a balance between the responsiveness required of mobile robots and the discriminative capabilities required for downstream tasks. Third, multimodal fusion analysis provides strong empirical validation of the advantages of integrating heterogeneous perception modalities. By comparing camera-only systems, lidar-only systems, and fused systems, experiments demonstrate that sensor fusion consistently achieves the highest localization accuracy, with accuracy improvements of up to 24% compared to single-modal baselines. This confirms that cameras and lidar can compensate for each other's weaknesses, particularly in terms of texture sensitivity and geometric robustness, and that a tightly coupled optimization framework can effectively exploit their complementary nature. In practice, multi-sensor fusion has become a key factor in achieving robustness in cluttered and dynamic environments. Finally, pruning experiments reveal the potential of model compression as a deployment strategy. By systematically pruning network weights, inference latency was reduced by over 60% with only a slight decrease in accuracy. Importantly, the results identify an operational "sweet spot" within the pruning range of 20% to 40%, where latency is significantly reduced with negligible accuracy loss. This suggests that careful pruning can achieve an effective balance between efficiency and reliability, which is particularly valuable for embedded or real-time robotic systems with limited computational resources.

6 Future Trends and Research Directions

Recent advances in end-to-end learning, self-supervised and weakly supervised methods, powerful multimodal sensor fusion, and large-scale pre-trained models are collectively reshaping the landscape of robotic perception and navigation. End-to-end trainable systems enable robots to learn locomotion and map construction directly from raw sensor input, while self-supervised and weakly supervised methods reduce the reliance on labeled data, thereby enhancing scalability in real-world environments. Powerful sensor fusion, leveraging event cameras, lidar, millimeter-wave radar, and emerging frameworks such as B5gActiv, further enhances robots' adaptability in dynamic and complex scenarios. Meanwhile, generalized pre-trained models bridge vision, language, and motion, unlocking intuitive human-robot interaction and versatile multitasking capabilities. Together, these developments outline a clear trajectory towards more adaptable, efficient, and powerful mobile robots capable of real-time localization and semantic segmentation in complex and dynamic environments.

6.1 End-to-end Learning Methods in Robotics and Navigation

End-to-end learning methods have revolutionized the field of robotics and autonomous navigation such that systems are able to learn and improve themselves in dynamic conditions. In self-driving, end-to-end trainable driving models have made significant advances, particularly using large-scale video datasets that have the ability to predict vehicle movement directly from sensor inputs without being explicitly programmed with task-specific instruction [18]. This corresponds with efforts in refining real-time robot localization and scene semantic segmentation through mobile robots, wherein sensory information such as vision and LiDAR are input into deep learning models to predict environment and robot motion. These developments are especially relevant to the use of localization and semantic segmentation in mobile robots, wherein real-time learning from sensors and adaptive learning are crucial for efficient mapping and decision-making in dynamic environments. Furthermore, examples of meta-learning utilizations, such as the meta-control navigation system implemented for the Psikharpax rat robot, show how reinforcement learning can be employed to enhance navigation systems further through learning to adapt to time-varying conditions, a critical aspect of mobile robot performance in harsh environments.

6.2 Self-supervised and Weakly Supervised Learning

Reducing the reliance on large labeled datasets through self-supervised and weakly supervised learning is becoming a transformative approach for mobile robotics, particularly in the areas of real-time localization and semantic segmentation. Traditional supervised learning relies heavily on carefully labeled datasets, which are expensive, time-consuming, and often fail to scale to the scale required for robots to operate in diverse and dynamic environments. In contrast, self-supervised and weakly supervised paradigms are able to leverage abundant unlabeled data and exploit its inherent structure, auxiliary tasks, or weak supervisory signals, thereby addressing one of the most pressing bottlenecks in the deployment of machine learning in robotics. Recent advances in self-supervised learning have fully demonstrated the power of this paradigm. For example, wav2vec 2.0 [19] achieved significant improvements in speech recognition by learning effective latent representations directly from raw audio, demonstrating that useful features can be extracted without explicit human-provided labels. Similarly, Bootstrap Your Own Latent (BYOL) [20] achieved state-of-the-art performance in learning image representations without negative samples, highlighting its potential to capture rich and transferable visual features from large-scale unlabeled image datasets. Such techniques are particularly applicable to mobile robots, which have a continuous stream of visual and sensor data available but often have little labeled ground truth. By leveraging self-supervised objectives, mobile robots can autonomously learn meaningful feature embeddings, thereby improving robustness to environmental changes, generalization under unknown conditions, and adaptability to dynamic environments. Complementary to this, weakly supervised learning further reduces the labeling requirements by leveraging incomplete, imprecise, or noisy labels. For example, Mohammed et al. [21] demonstrated how weakly supervised methods can be applied to object localization in cluttered real-world scenes, significantly reducing the reliance on dense annotations while still achieving reliable detection. In mobile robotic systems, such weakly supervised learning can take the form of bounding boxes, image-level labels, or even natural language descriptions, all of which are cheaper to acquire than pixel-level annotations or precise pose labels. By integrating weakly supervised learning, robots can perform real-time segmentation and localization in unstructured environments with minimal annotation overhead, thereby extending their deployment scalability and applicability in real-world operations. Taken together, these advances demonstrate how self-supervised and weakly supervised learning can directly impact the efficiency and scalability of autonomous robotic systems. By allowing mobile robots to extract semantic and spatial understanding from raw sensory streams with minimal supervision, these paradigms enhance scene understanding and situational awareness. In dynamic environments, where pre-labeled datasets cannot capture all variations, the ability to continuously learn from unlabeled experience can significantly improve autonomy and reduce system vulnerability. Looking ahead, combining self-supervised and weakly supervised methods with multimodal sensor fusion, online adaptation, and reinforcement learning promises to further enhance robotic perception capabilities, creating data-efficient and resilient systems capable of lifelong learning under real-world conditions.

6.3 Strong Sensor Fusion for Robotic Perception

The integration of various sensors is a significant improvement in the expansion of the perception ability of mobile robots, especially in dynamic and cluttered environments. Event cameras have high temporal resolution and low latency, which are very helpful for tasks such as obstacle avoidance and are indispensable for real-time navigation and segmentation of dynamic scenes [22]. When combined with mmWave radar, which performs well under low lighting and bad weather, the two-sensor combination can improve perception significantly. For example, Fan et al. [23] showed a multi-target tracking system fusing 4-D mmWave radar and monocular vision with outstanding performance in dynamic environments. This sensor fusion is directly applied to real-time localization and semantic segmentation, enabling mobile robots to process multimodal dynamic data to better understand the environment. In addition, Hashima et al. [24] proposed a B5gActiv System framework to improve ego-motion estimation and dynamic target tracking in Fig. 7, emphasizing the need for robust sensor fusion techniques that can handle dynamic and cluttered environments. The complementarity of event cameras, lidar, and B5gActiv System holds great promise to improve the localization and semantic perception of mobile robots, making them robust and generalizable for real-time applications.

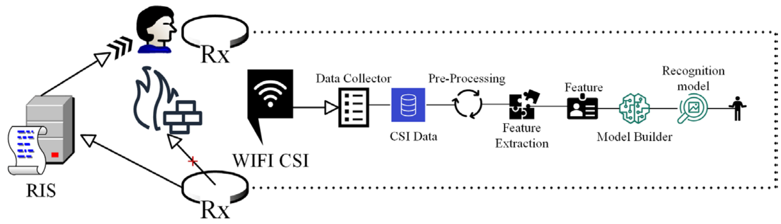

Fig. 7. B5gActiv system architecture

6.4 Generalized Pretrained Models for Robotics

Large-scale pre-trained models have transformed the robotics field and made multi-task learning and improving robotic capabilities possible. Pre-training image-language models like GLIP to be applied in low-shot 3D point cloud segmentation has demonstrated potential in closing the gap between language and vision modalities in perception in robotics. These models allow robotic systems to understand and process complex multi-modal inputs, hence being capable of executing a variety of tasks using minimal data. Similarly, the release of Instruct2Act, which uses large language models to translate instruction into robot actions, says it all about the potential of pre-trained models to facilitate more intuitive human-robot interaction. The developments point towards the ability of generalized pre-trained models to enable robots to achieve a wide range of tasks in a more efficient and flexible way. Each of these developments separately indicates the rapid growth of the robotics industry, particularly towards real-time mobile robot localization and scene semantic segmentation in dynamic environments. The incorporation of end-to-end learning, self-supervised and weakly-supervised learning, sensor fusion, and privacy-conscious designs significantly adds to the efficiency of these robots. With advancing technology, the integration of robust machine learning algorithms with advanced sensors and robust privacy controls will be the magic that unlocks the real potential of autonomous robotic systems in real-world and dynamic contexts.

7 Conclusion

This study provides a systematic evaluation of real-time localization and semantic segmentation technologies for mobile robots operating in dynamic environments. The work establishes an experimental framework combining benchmark datasets and real-world scenarios, offers an empirical comparison of representative SLAM systems and segmentation networks, and demonstrates that multi-sensor fusion markedly enhances localization robustness while lightweight deep models ensure real-time performance with acceptable accuracy loss. Ablation analyses further highlight the role of fusion strategies and model optimization in balancing efficiency and precision. Although the diversity of datasets and scene complexity impose limitations, the findings deliver valuable empirical evidence and design guidelines for the development of reliable, real-time perception systems, and point toward future advances through expanded datasets, multimodal sensing, and self-supervised learning to achieve robust autonomy in complex real-world deployments.

8 Acknowledgement

Funding Sources: This paper was supported by Key Research and Development Plan Self Funded Project of Xingtai, Hebei Province in 2021 “Research on Indoor Mobile Robot Localization Algorithm Based on Multi-Sensor Fusion” (No: 2021ZC010).

References

[1] H.-M. Lu, X. Li, H. Zhang, M. Hu, Z.-Q. Zheng, Robust and Real-Time Self-Localization Based on Omnidirectional Vision for Soccer Robots, Advanced Robotics 27(10)(2013) 799-811.
https://doi.org/10.1080/01691864.2013.785473

[2] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. Reid, S. Gould, P. Corke, M. Milford, Semantics for Robotic Mapping, Perception and Interaction: A Survey, Foundations and Trends in Robotics 8(1-2)(2020) 1-224.
http://dx.doi.org/10.1561/2300000059

[3] S. Han, Z. Xi, Dynamic Scene Semantics SLAM Based on Semantic Segmentation, IEEE Access 8(2020) 43563–43570.
https://doi.org/10.1109/ACCESS.2020.2977684

[4] C.-H. Chang, S.-C. Wang, C.-C. Wang, Exploiting Moving Objects: Multi-Robot Simultaneous Localization and Tracking, IEEE Transactions on Automation Science and Engineering 13(2)(2016) 810-827.
https://doi.org/10.1109/TASE.2015.2426203

[5] L. Li, M. Yang, H. Li, C. Wang, B. Wang, Robust Localization for Intelligent Vehicles Based on Compressed Road Scene Map in Urban Environments, IEEE Transactions on Intelligent Vehicles 8(1)(2023) 250-262.
https://doi.org/10.1109/TIV.2022.3162845

[6] N. Ma, X. Zhang, H. Zheng, J. Sun, ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design, in: Proc. of the European Conference on Computer Vision, 2018.
https://doi.org/10.1007/978-3-030-01264-9_8

[7] Ankita, S. Rani, H. Babbar, S. Coleman, A. Singh, H.M. Aljahdali, An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones, Sensors 21(11)(2021) 3845.
https://doi.org/10.3390/s21113845

[8] Y. He, X. Zhang, J. Sun, Channel Pruning for Accelerating Very Deep Neural Networks, in: Proc. of the IEEE International Conference on Computer Vision, 2017.

[9] T. Zebin, P.-J. Scully, N. Peek, A.-J. Casson, K.-B. Ozanyan, Design and Implementation of A Convolutional Neural Network on An Edge Computing Smartphone for Human Activity Recognition, IEEE Access 7(2019) 133509-133520.
https://doi.org/10.1109/ACCESS.2019.2941836

[10] S. Naveen, M.-R. Kounte, M.-R. Ahmed, Low Latency Deep Learning Inference Model for Distributed Intelligent IoT Edge Clusters, IEEE Access 9(2021) 160607-160621.
https://doi.org/10.1109/ACCESS.2021.3131396

[11] G. Carvalho, B. Cabral, V. Pereira, J. Bernardino, Edge Computing: Current Trends, Research Challenges and Future Directions, Computing 103(5)(2021) 993-1023.
https://doi.org/10.1007/s00607-020-00896-5

[12] P. Mach, Z. Becvar, Mobile Edge Computing: A Survey on Architecture and Computation Offloading, IEEE Communications Surveys and Tutorials 19(3)(2017) 1628-1656.
https://doi.org/10.1109/COMST.2017.2682318

[13] H.-M. Do, T.-P. Tran, M. Yoo, Deep Reinforcement Learning-Based Task Offloading and Resource Allocation for Industrial IoT in MEC Federation System, IEEE Access 11(2023) 83150-83170.
https://doi.org/10.1109/ACCESS.2023.3302518

[14] Q. She, F. Feng, X. Hao, Q. Yang, C. Lan, V. Lomonaco, X. Shi, Z. Wang, Y. Guo, Y. Zhang, F. Qiao, R.H.M. Chan, OpenLORIS-Object: A Robotic Vision Dataset and Benchmark for Lifelong Deep Learning, in: Proc. of the IEEE International Conference on Robotics and Automation, 2020.
https://doi.org/10.1109/ICRA40945.2020.9196887

[15] P. Garg, R. Saluja, V.-N. Balasubramanian, C. Arora, A. Subramanian, C.-V. Jawahar, Multi-Domain Incremental Learning for Semantic Segmentation, in: Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.
https://doi.org/10.1109/WACV51458.2022.00214

[16] M. Imdoukh, I. Ahmad, M.-G. Alfailakawi, Machine Learning-Based Auto-Scaling for Containerized Applications, Neural Computing and Applications 32(13)(2020) 9745-9760.
https://doi.org/10.1007/s00521-019-04507-z

[17] M.-R. Bouguelia, S. Nowaczyk, A.-H. Payberah, An Adaptive Algorithm for Anomaly and Novelty Detection in Evolving Data Streams, Data Mining and Know ledge Discovery 32(6)(2018) 1597-1633.
https://doi.org/10.1007/s10618-018-0571-0

[18] H.-Z. Xu, Y. Gao, F. Yu, T. Darrell, End-to-End Learning of driving models from large-scale video datasets, in: Proc. 2017 of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
https://doi.org/10.1109/CVPR.2017.376

[19] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in: Proc. the 34th International Conference on Neural Information Processing System, 2020.

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z. Guo, M.-G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, in: Proc. of the 34th International Conference on Neural Information Processing Systems, 2020.
[21] M.-Q. Mohammed, L.-C. Kwek, S.-C. Chua, A. Al-Dhaqm, S. Nahavandi, T.-A.-E. Eisa, M.-F. Miskon, M.-N. Al-Mhiqani, A. Ali, M. Abaker, E.-A. Alandoli, Review of Learning-Based Robotic Manipulation in Cluttered Environments, Sensors 22(20)(2022) 7938.1-7938.22.
https://doi.org/10.3390/s22207938
[22] W. Shariff, M.S. Dilmaghani, P. Kielty, M. Moustafa, J. Lemley, P. Corcoran, Event Cameras in Automotive Sensing: A Review, IEEE Access 12(2024) 51275-51306.
https://doi.org/10.1109/ACCESS.2024.3386032
[23] L. Fan, J. Wang, Y. Chang, Y. Li, Y. Wang, D. Cao, 4D mmWave Radar for Autonomous Driving Perception: A Comprehensive Survey, IEEE Transactions on Intelligent Vehicles 9(4)(2024) 4606-4620.
https://doi.org/10.1109/TIV.2024.3380244
[24] S. Hashima, H. Rizk, H. Yamaguchi, The Future of Beyond 5G Sensing: Transforming Activity Recognition with Reconfigurable Intelligent Surfaces, in: Proc. of the International Conference on Activity and Behavior Computing, 2024.
https://doi.org/10.1109/ABC61795.2024.10651907