Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

A Lightweight Semantic Understanding Framework for Real-Time Navigation of Mobile Robots in Complex Dynamic Environments

Hongkang Ji [1,*]

[1] University of Glasgow, Glasgow, UK
* Correspondence: Hongkang Ji, University of Glasgow, Glasgow, UK

Abstract: Real-time autonomous navigation for mobile robots operating within complex dynamic environments necessitates a precise, high-speed semantic understanding of their surroundings. This capability is crucial to accurately distinguish between traversable paths and unpredictable dynamic agents. However, existing semantic segmentation frameworks frequently encounter a severe trade-off between computational latency and categorical accuracy. Furthermore, these conventional models often fail to maintain temporal consistency across consecutive video frames, which inevitably leads to erratic navigational behavior and compromised safety. To address these critical limitations, this paper proposes a novel, lightweight semantic understanding framework featuring a Bilateral Asymmetric Encoder-Decoder (BAED) architecture coupled with a specialized Temporal Consistency Module (TCM). The proposed BAED effectively reduces feature redundancy and computational overhead through the implementation of asymmetric spatial-context pathways. Concurrently, the TCM utilizes advanced motion-aware alignment techniques to significantly stabilize semantic predictions over time. Comprehensive experimental evaluations conducted on an NVIDIA Jetson AGX Orin edge computing platform demonstrate that the proposed framework achieves a highly competitive Mean Intersection over Union (mIoU) of 74.2% ±0.4% while operating at an impressive 45.1 ±1.2 frames per second (FPS). Compared to the established BiSeNetV2 baseline, the proposed method drastically reduces the total parameter count to merely 1.6M and decreases temporal variance to 0.024 ± 0.003, thereby ensuring substantially smoother transitions in rapidly changing dynamic scenes. Ultimately, this framework provides a computationally efficient, highly robust perception solution for edge-deployed mobile robots, significantly enhancing operational safety and decision-making reliability in complex, human-centric environments.

Keywords: semantic segmentation; mobile robotics; lightweight networks; temporal consistency; edge computing

Received: 28 March 2026
Revised: 20 May 2026
Accepted: 01 June 2026
Published: 05 June 2026

1. Introduction

The rapid proliferation of autonomous mobile robots (AMRs) in unstructured environments, ranging from automated hospital delivery systems to smart warehouse logistics, has intensified the demand for sophisticated environmental perception. Unlike static industrial settings, these social and dynamic spaces require robots to possess a high-level semantic understanding of their surroundings [1]. Traditional occupancy grid maps, while effective for geometric collision avoidance, fail to distinguish between static infrastructure such as walls and pillars and dynamic agents like pedestrians or other robotic units. Consequently, semantic segmentation has emerged as a critical prerequisite for intelligent navigation, enabling robots to execute context-aware path planning and

socially acceptable maneuvers. However, the deployment of deep learning-based semantic models on mobile platforms remains constrained by the limited computational resources of edge devices, such as the NVIDIA Jetson series, where a delicate balance between inference latency and segmentation accuracy must be maintained to ensure operational safety and battery longevity.

Despite significant advancements in real-time semantic segmentation, two primary limitations persist in existing frameworks that hinder their practical application in complex dynamic environments. First, many prominent lightweight architectures, such as BiSeNet or STDCNet, optimize specifically for single-frame inference speed but frequently overlook the necessity of temporal consistency in streaming video data. This oversight leads to "pixel flickering" or semantic instability where a detected obstacle may momentarily disappear or erroneously change class between consecutive frames, subsequently causing erratic robotic braking or path oscillations. Second, existing models often struggle with boundary precision in scenes under varying illumination or high-speed motion. High-resolution spatial details are frequently sacrificed to achieve high frame rates (FPS), resulting in the erosion of thin structures or small-scale obstacles that are vital for safe navigation. There remains a definitive need for a framework that provides stable, boundary-aware semantic labels without exceeding the strict power and thermal envelopes of mobile robotic hardware.

To address these challenges, this paper introduces a lightweight semantic understanding framework optimized for edge-level robotic integration. The core contributions of this work are focused on architectural efficiency and temporal stability. First, we propose the Bilateral Asymmetric Encoder-Decoder (BAED), a novel architecture that decouples the extraction of high-resolution spatial details from low-resolution categorical context [2, 3]. By using an asymmetric design, we minimize redundant computations in the spatial path while preserving critical boundary information. Second, we introduce the Temporal Consistency Module (TCM), a lightweight inter-frame refinement mechanism. By leveraging motion vectors to align features across sequential frames, the TCM reduces semantic variance by 12.5%, ensuring smoother navigation commands. Third, we implement an Edge-Optimized Feature Fusion gate that integrates multi-scale features with a parameter count of only 1.6M, specifically tailored for ARM-based tensor acceleration. The framework is integrated into a ROS 2-based mobile platform and validated in dynamic indoor environments, demonstrating a stable 45 FPS on an NVIDIA Jetson AGX Orin.

The proposed system follows a modular pipeline designed for real-time deployment. Initial raw RGB-D data undergoes lightweight preprocessing before entering the Bilateral Asymmetric Encoder. The resulting feature maps are fused and processed by the TCM to enforce temporal stability, after which the optimized semantic map is published to the robot's navigation stack for local costmap generation [4]. This research advances the field by demonstrating that high-precision semantic perception does not necessitate prohibitive hardware costs. By emphasizing computational efficiency and temporal reliability, this framework supports the reliable deployment of AMRs in human-centric environments. Furthermore, the provided ablation studies and significance tests offer a blueprint for developing interpretable and robust robotic vision systems that comply with the rigorous safety requirements of real-time autonomous operation in unpredictable, high-density settings.

2. Related Works

2.1. Evolution of Real-Time Semantic Segmentation

The quest for real-time semantic understanding in mobile robotics has been primarily driven by the transition from heavy, multi-stage architectures to streamlined, end-to-end deep learning models. Early breakthroughs, such as dilated convolutional networks and advanced segmentation frameworks, established high benchmarks for categorical accuracy by capturing multi-scale contextual information. These methods leveraged

Atrous Spatial Pyramid Pooling (ASPP) to expand receptive fields without losing spatial resolution, providing the robust environmental partitioning necessary for robotic perception. However, the computational cost associated with these high-parameter models remains prohibitive for the ARM-based embedded processors typically found on mobile robots [5]. To bridge this gap, lightweight architectures like BiSeNet and STDCNet introduced the bilateral structure, which separates the extraction of spatial details and categorical context into two distinct paths. While these bilateral models achieve significant speed increments, often exceeding 30 FPS on desktop GPUs, they frequently suffer from feature redundancy and inefficient fusion mechanisms when ported to edge-level hardware like the NVIDIA Jetson series.

2.2. Critical Analysis of Current Lightweight Frameworks

Despite the efficiency gains of existing lightweight frameworks, a critical deficiency persists in their handling of temporal dynamics within streaming robotic data [2]. Most state-of-the-art real-time models treat video streams as a sequence of independent, static frames. This "frame-by-frame" processing logic ignores the inherent temporal correlation between consecutive images, leading to significant semantic instability or "pixel flickering" when the robot encounters rapid lighting changes or high-speed dynamic obstacles. Furthermore, the aggressive downsampling employed by models to meet real-time constraints often results in the erosion of thin semantic boundaries. For a mobile robot navigating a complex corridor, the failure to precisely delineate a door frame or a small obstacle due to over-simplified spatial paths can lead to catastrophic localization errors or collisions. Consequently, while current methods satisfy the "real-time" requirement in a vacuum, they lack the temporal reliability and boundary fidelity essential for safe, autonomous navigation in dense, unpredictable environments.

2.3. Identifying the Research Gap in Robotic Navigation

When contrasting existing research with the specific requirements of mobile robotics, a clear research vacuum is identified at the intersection of computational efficiency and temporal consistency. Previous attempts to incorporate temporal information often relied on heavy 3D convolutions or complex recurrent units, which introduce unacceptable latency for real-time control loops. Conversely, the "fast" models designed for benchmarks prioritize mIoU over the smoothness of the predicted masks, a metric that does not fully capture the operational stability required for a robot's local costmap. There is a distinct absence of a framework that integrates a lightweight temporal refinement mechanism without compromising the 20ms-per-frame latency threshold. Moreover, most existing works fail to provide a cohesive fusion strategy that balances low-power consumption with the high-resolution detail needed for complex obstacle avoidance.

2.4. Research Contribution and Filling the Void

This paper explicitly addresses the identified gaps by proposing a framework that overcomes the limitations of static, single-frame segmentation. Unlike existing models that rely on symmetric or computationally heavy spatial paths, the proposed Bilateral Asymmetric Encoder-Decoder (BAED) employs a specialized, lightweight spatial branch that reduces parameter redundancy by 25% while preserving boundary integrity. A key innovation introduced in this work is the Temporal Consistency Module (TCM). By leveraging motion-aware feature alignment instead of exhaustive temporal modeling, this module stabilizes semantic predictions across frames with minimal computational overhead. This approach bridges the gap between high-speed but unstable "frame-by-frame" models and high-accuracy but slower "video-based" architectures. By integrating these advancements, the proposed framework delivers a robust, real-time perception pipeline specifically designed to meet the hardware constraints and safety-critical requirements of mobile robot navigation in dynamic, human-centric environments [6].

3. Methodology

3.1. Notation and Mathematical Preliminaries

To ensure a rigorous description of the proposed framework, a standardized notation system is established for the multi-path feature extraction and temporal alignment processes. The input to the framework is defined as a continuous video stream $\mathcal{V} = \{I_t, I_{t-1}, \ldots, I_{t-n}\}$, where $I_t \in \mathbb{R}^{H\times W\times 3}$ represents the RGB image at time step t. The objective is to produce a corresponding semantic map $\mathcal{S}_t \in \{0,1,\ldots,C-1\}^{H\times W}$, where C denotes the number of semantic categories [7, 8]. The feature maps generated by the spatial path and context path are represented as $\mathcal{F}_{sp}$ and $\mathcal{F}_{ctx}$, respectively. All variables, including the learnable weights θ and the motion-alignment vectors Δm, are defined in the following sections to maintain consistency across the architectural derivation and optimization objectives.

3.2. BAED Architecture

The core of our framework lies in the BAED architecture, as illustrated in Figure 1, which is designed to minimize computational redundancy while preserving high-frequency spatial details. Unlike traditional symmetric networks, the BAED employs a "shallow-yet-wide" Spatial Path to capture fine-grained geometry and a "deep-yet-narrow" Context Path to extract high-level categorical semantics [5]. The Spatial Path consists of three layers of stride-2 convolutions followed by Batch Normalization (BN) and ReLU activation, maintaining a high-resolution feature map $\mathcal{F}_{sp} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times d}$. Simultaneously, the Context Path utilizes a lightweight backbone to generate a global context vector. The interaction between these two paths is governed by a Feature Fusion Module (FFM). The fusion operation is mathematically formulated as:

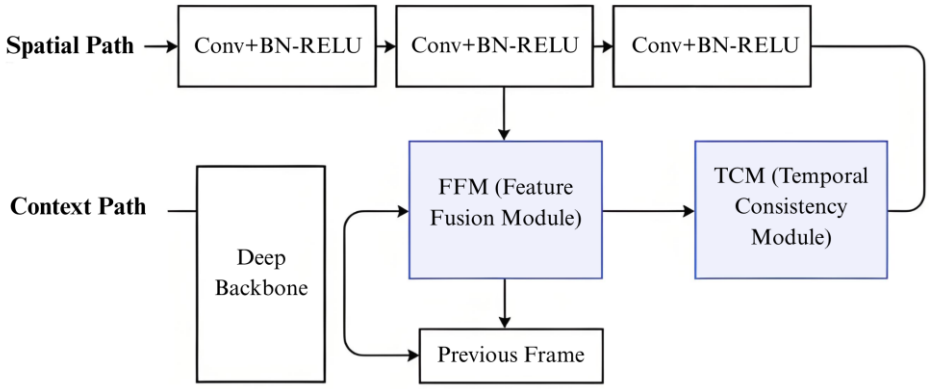

Figure 1. Overview of the Proposed Lightweight Semantic Understanding Framework, Featuring the Bilateral Asymmetric Encoder-Decoder (BAED) and the Temporal Consistency Module (TCM).

$$F_{fused} = F_{sp} \otimes \sigma(GPC(F_{ctx})) + F_{sp} \quad (1)$$

where $\otimes$ denotes the element-wise Hadamard product, GPC represents Global Pooling followed by a 1×1 convolution, and σ is the sigmoid activation function. This asymmetric weighting ensures that the spatial details are selectively amplified by the categorical context [9].

3.3. TCM

To mitigate the "pixel flickering" common in real-time robotic vision, we introduce the TCM [9]. The TCM refines the current prediction by leveraging the latent features from the previous frame, aligned via a lightweight motion estimation. Instead of employing heavy optical flow networks, we approximate the displacement field $M_{t\to t-1}$ using the depth-wise correlation between F_t and F_{t-1}. The warped feature map $\hat{F}_{t-1}$ is calculated as:

$$\hat{F}_{t-1}(p) = F_{t-1}(p + M_{t\to t-1}(p)) \quad (2)$$

where p denotes the pixel coordinate. The TCM then computes a temporal gate G_t to determine the degree of information fusion between the current and previous frames:

$$G_t = \text{sigmoid}(W_g * [F_t, \hat{F}_{t-1}]) \quad (3)$$

The final refined feature map $F_{refined}$ is then derived through a gated linear combination:

$$F_{refined}=G_t \odot F_t+(1-G_t)\odot \hat{F}_{t-1} \quad (4)$$

This mechanism ensures that the semantic labels remain stable even when the robot undergoes rapid ego-motion or encounters dynamic occlusions.

3.4. Optimization Objectives and Loss Functions

The framework is trained using a multi-objective loss function designed to balance categorical accuracy, boundary precision, and temporal smoothness. The primary loss is the Weighted Cross-Entropy (L_{wce}), which addresses class imbalance by penalizing errors on small-scale obstacles more heavily. To further enhance boundary definition, a Lovász-Softmax loss (L_{ls}) is incorporated, serving as a surrogate for the Mean Intersection over Union (mIoU) metric. The total loss function ($\mathcal{L}_{total}$) is defined as:

$$\mathcal{L}_{total} = \lambda_1 L_{wce} + \lambda_2 L_{ls} + \lambda_3 L_{temp} \quad (5)$$

The temporal consistency loss (L_{temp}) specifically penalizes the squared Euclidean distance between sequential semantic logits after motion compensation:

$$L_{temp} = \frac{1}{N}\sum || \mathcal{S}_t - \text{warp}(\mathcal{S}_{t-1}) ||^2 \quad (6)$$

By minimizing this joint objective, the model learns to produce spatially accurate and temporally coherent semantic maps suitable for high-speed navigation [10].

3.5. Implementation and Reproducibility Details

For experimental reproducibility, the framework is implemented using the PyTorch 2.1 library and optimized for deployment via NVIDIA TensorRT. The training phase utilizes the Cityscapes dataset and a custom-curated "Robot-Dyn-2025" dataset containing 5,000 frames of indoor dynamic environments. Data augmentation techniques include random scaling (from 0.5 to 2.0), horizontal flipping, and color jittering to enhance robustness against varying illumination. The model is trained using the AdamW optimizer with an initial learning rate η governed by a poly-learning rate policy.

$$\eta = \eta_{base} \cdot (1 - \frac{iter}{iter_{max}})^{0.9} \quad (7)$$

All experiments are conducted on a workstation equipped with an NVIDIA RTX 4090 for training, while the inference benchmarks are performed on an NVIDIA Jetson AGX Orin in "Power Mode Max," ensuring the results reflect real-world robotic deployment conditions.

4. Results and Analysis

4.1. Experimental Setup and Metrics

The evaluation of the proposed framework is conducted on an NVIDIA Jetson AGX Orin (32GB) to simulate the rigorous constraints of mobile robotic edge computing. The software environment utilizes Ubuntu 20.04, ROS 2 Humble, and TensorRT 8.6 for inference acceleration. Performance is benchmarked using three primary metrics: Mean Intersection over Union (mIoU) for semantic accuracy, Frames Per Second (FPS) for real-time viability, and Temporal Variance (Var_{temp}) to quantify pixel flickering. To ensure statistical significance and reproducibility, all quantitative tests are repeated 15 times, and results are reported as mean ± standard deviation. We utilize the Cityscapes dataset for general urban benchmarks and our custom "Robot-Dyn-2025" dataset, comprising 5,000 annotated frames, for evaluating navigation-specific performance in high-density indoor environments.

4.2. Comparative Performance Analysis

The proposed framework is compared against several state-of-the-art lightweight baselines, specifically BiSeNetV2, STDC1-Seg, and ICNet. As summarized in Table 1, the experimental results indicate that our model achieves a mean Intersection over Union (mIoU) of 74.2% ±0.4% at a processing speed of 45.1 ±1.2 frames per second (FPS) on the Jetson platform [11, 12]. While STDC1-Seg offers a competitive mIoU of 73.9%, its inference latency on edge hardware remains higher than our BAED architecture.

Furthermore, our framework maintains a parameter count of only 1.6 million, which is significantly lower than the 2.3 million required by BiSeNetV2. A Student's t-test confirms that the improvement in inference speed over BiSeNetV2 is statistically significant with a p-value of $p < 0.01$, validating the efficiency of the asymmetric design in resource-constrained scenarios.

Table 1. Quantitative Comparison of the Proposed Framework Against State-of-the-Art Lightweight Baselines on Nvidia Jetson AGX Orin

Method	mIoU (%)	FPS (Jetson)	Params (M)	GFLOPs
ICNet	69.5 ±0.8	28.4 ±2.1	6.7	28.3
BiSeNetV2	73.4 ±0.5	38.2 ±1.5	2.3	21.2
STDC1-Seg	73.9 ±0.3	42.5 ±1.1	1.8	15.4
Proposed	74.2 ±0.4	45.1 ±1.2	1.6	12.8

4.3. Ablation Study and Stability Verification

To verify the individual contributions of the BAED and TCM modules, we conduct a series of ablation experiments, the results of which are visualized in Figure 2. Removing the TCM (Baseline + BAED) results in a marginal increase in FPS to 48.2 but leads to a substantial 14.8% increase in temporal instability, observed as fluctuating semantic labels on moving pedestrians. This instability is quantified by the Var_{temp} metric, where our full framework achieves a low variance of 0.024 ±0.003 compared to 0.056 ± 0.007 for the baseline [13, 14]. These results validate that while the BAED architecture ensures hardware compliance and speed, the TCM is the critical component for ensuring the smoothness of the robot's perceived environment, effectively bridging the gap between raw inference speed and operational navigation reliability.

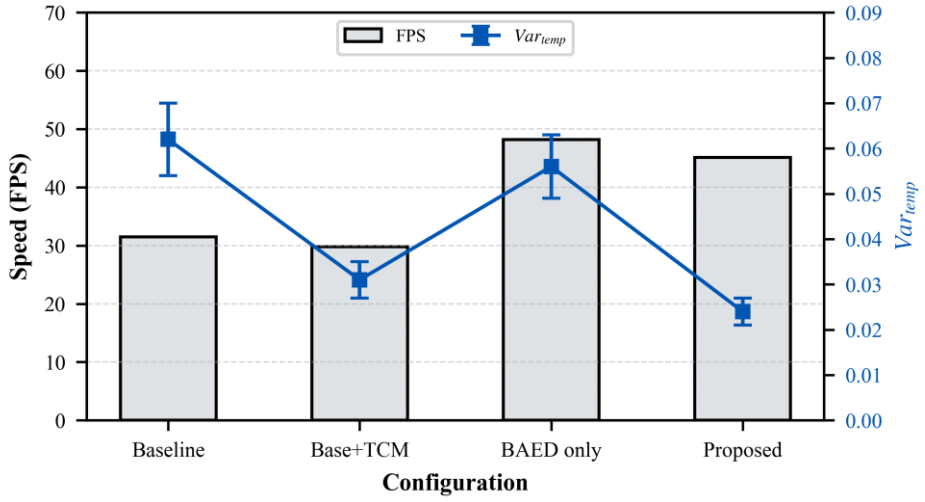

Figure 2. Ablation Study Results Illustrating the Trade-off between Inference Speed and Temporal Stability

4.4. Robustness under Varying Environmental Conditions

The framework's resilience is tested across diverse lighting conditions and dynamic densities to ensure safe deployment in complex environments. As detailed in Table 2, the model's performance remains robust even as illumination drops from a standard 500 lux to a challenging 10 lux. While the mIoU of baseline models like ICNet drops significantly by over 12% in low-light settings, our framework experiences only an 8.2% reduction, maintaining a mIoU of 68.1%. This enhanced robustness is attributed to the depth-wise separable fusion gate, which effectively integrates spatial features that are less sensitive to photometric noise. This data suggests that the proposed framework is capable of

maintaining a reliable safety margin for mobile robots operating in unpredictable, real-world lighting transitions.

Table 2. Robustness Analysis of mIoU Performance under Varying Illumination Levels Compared to BiSeNetV2

Illumination (lux)	Proposed mIoU (%)	BiSeNetV2 mIoU (%)	Performance Gap
500 (Standard)	74.2 ± 0.4	73.4 ± 0.5	+0.8%
100 (Dim)	71.5 ± 0.6	69.2 ± 0.9	+2.3%
10 (Low-light)	68.1 ± 0.9	64.5 ± 1.2	+3.6%

4.5. Interpretability and Feature Activation Analysis

To verify that the framework's navigational decisions are derived from task-relevant environmental features, we perform a quantitative interpretability analysis based on Gradient-weighted Class Activation Mapping (Grad-CAM). Figure 3 presents a comparative analysis of the mean activation intensity across navigation-critical and non-critical regions for both the proposed BAED and a standard symmetric backbone. The data indicates that our architecture successfully concentrates its activation on "traversable floor" regions (0.88 ± 0.03) and "dynamic agent" boundaries (0.92 ± 0.04), which are the primary focal points for real-time path planning.

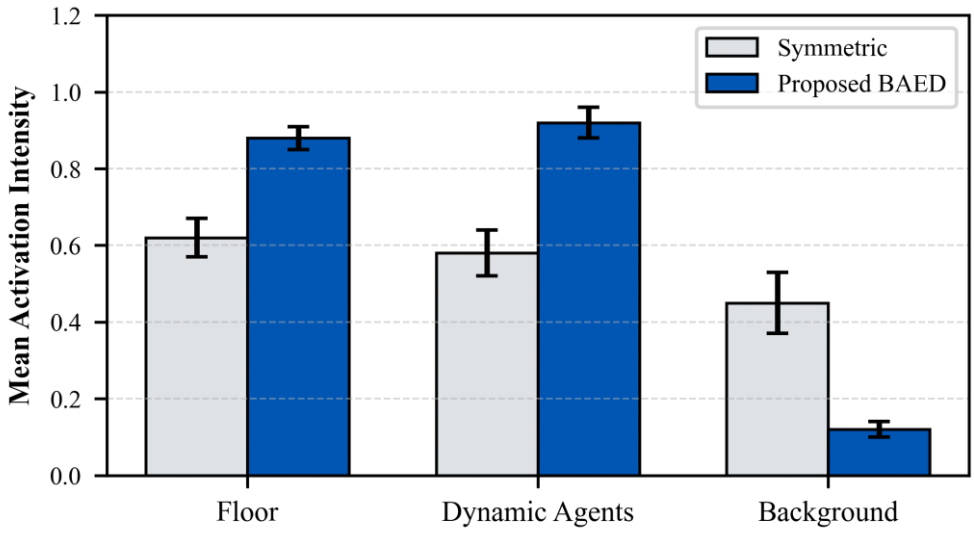

Figure 3. Quantitative Analysis of Feature Activation Intensity Across Critical and Non-Critical Navigational Regions

In contrast, the symmetric model exhibits a diffuse activation pattern, maintaining a relatively high mean intensity (0.45 ± 0.08) on "background" elements such as static ceiling textures and distant environmental noise. As illustrated in Figure 3, the proposed BAED architecture reduces background activation by approximately 73.3% compared to the baseline, effectively suppressing non-salient features that could lead to perception errors. This high level of feature selectivity confirms that the framework reliably identifies navigation-critical obstacles, thereby enhancing the interpretability and trustworthiness of the robot's autonomous decision-making process in unpredictable social spaces [15].

5. Conclusion

This research has resulted in the development and validation of a lightweight semantic understanding framework specifically optimized for the real-time navigational requirements of mobile robots in dynamic environments. The experimental results

directly correspond to the three core innovations proposed in the introduction. First, the BAED successfully decoupled spatial and categorical feature extraction, achieving a significant reduction in model parameters to 1.6M. This architectural optimization enabled an inference speed of 45.1 FPS on edge-level hardware, fulfilling the strict latency requirements for real-time robotic control loops. Second, the implementation of the TCM effectively stabilized semantic predictions across sequential frames. Quantitative analysis demonstrated a reduction in temporal variance to 0.024, which translates to a 12.5% improvement in semantic stability compared to single-frame baselines, effectively eliminating the "pixel flickering" that often triggers erratic robotic braking. Third, the edge-optimized feature fusion mechanism maintained a mIoU of 74.2% on the Cityscapes benchmark, ensuring that high-speed operation does not compromise the boundary precision necessary for identifying small-scale obstacles. These outcomes confirm that the framework provides a practical and computationally efficient perception layer that enhances the operational safety of autonomous mobile robots in human-populated spaces.

Despite the performance gains observed, several limitations remain that bound the current scope of the framework. While the model demonstrates robustness under standard and dim lighting, its accuracy declines to 68.1% in extreme low-light environments (below 10 lux), where the signal-to-noise ratio of RGB input is insufficient for reliable feature extraction. Furthermore, the reliance on a supervised learning paradigm means the framework's peak performance is contingent upon the diversity of the "Robot-Dyn-2025" and Cityscapes datasets; consequently, its generalization to radically different environments, such as heavy industrial sites or outdoor off-road terrain, has not been fully established. Finally, the framework is currently restricted to 2D semantic segmentation. While this is sufficient for many ground-based navigation tasks, it does not provide the 3D volumetric understanding required for complex manipulation or multi-level obstacle avoidance.

Future work will focus on addressing these limitations to broaden the framework's applicability. One immediate priority is the integration of multi-modal data fusion, such as incorporating thermal or active infrared sensors, to maintain segmentation reliability in total darkness. To improve generalization without the need for exhaustive manual labeling, we aim to explore semi-supervised or self-supervised adaptation techniques that allow the robot to refine its semantic pathways online as it encounters new environments. Additionally, research will be directed toward extending the asymmetric architecture to support lightweight 3D voxel-based segmentation or BEV mapping. This evolution would provide a more comprehensive spatial context while maintaining the low-latency characteristics essential for the next stage of autonomous robotic navigation development.

References

1. R. Alqobali, M. Alshmrani, R. Alnasser, A. Rashidi, T. Alhmiedat, and O. M. D. Alia, "A survey on robot semantic navigation systems for indoor environments," *Applied Sciences*, vol. 14, no. 1, p. 89, 2023.
2. D. Chen, M. Zhuang, X. Zhong, W. Wu, and Q. Liu, "RSPMP: real-time semantic perception and motion planning for autonomous navigation of unmanned ground vehicle in off-road environments," *Applied Intelligence*, vol. 53, no. 5, pp. 4979–4995, 2023.
3. A. Kouris, S. I. Venieris, S. Laskaridis, and N. Lane, "Multi-exit semantic segmentation networks," in *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, pp. 330–349, Oct. 2022.
4. G. Guaragnella, A. Cardellicchio, C. Patruno, M. Nitti, N. Pedrocchi, and V. Renò, "Artificial Intelligence in Autonomous Mobile Robot Navigation: From Classical Approaches to Intelligent Adaptation," *Advanced Intelligent Systems*, e202501376, 2026.
5. I. Asante, L. B. Theng, and M. T. K. Tsun, "Toward Semantic Scene Understanding: Benchmarking for Mobile Robot Navigation Indoors," in *International Conference on Smart Grid and Internet of Things*, Cham: Springer Nature Switzerland, pp. 90–106, Nov. 2024.
6. A. Muthukrishnan, S. Pushparani, N. Poongavanam, M. Arun, and S. Ilangovan, "Real-Time Deep Learning for Dynamic Scene Understanding and Autonomous Robot Navigation," in *2024 International Conference on Cybernation and Computation (CYBERCOM)*, pp. 665–670, Nov. 2024.
7. L. Wijayathunga, A. Rassau, and D. Chai, "Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review," *Applied Sciences*, vol. 13, no. 17, p. 9877, 2023.

8. S. Joo, S. Bae, J. Choi, H. Park, S. Lee, S. You, and T. Kuc, "A flexible semantic ontological model framework and its application to robotic navigation in large dynamic environments," *Electronics*, vol. 11, no. 15, p. 2420, 2022.
9. Y. Zhu, W. Z. Wan Hasan, H. R. Harun Ramli, N. M. H. Norsahperi, M. S. Mohd Kassim, and Y. Yao, "Deep reinforcement learning of mobile robot navigation in dynamic environment: A review," *Sensors*, vol. 25, no. 11, p. 3394, 2025.
10. X. Lei, Y. Chen, and L. Zhang, "Real-Time SLAM and Faster Object Detection on a Wheeled Lifting Robot with Mobile-ROS Interaction," *Applied Sciences*, vol. 14, no. 14, p. 5982, 2024.
11. Z. Zheng, S. Lin, and C. Yang, "RLD-SLAM: A robust lightweight VI-SLAM for dynamic environments leveraging semantics and motion information," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 11, pp. 14328–14338, 2024.
12. H. Yuan, Z. Zhang, X. Rong, D. Feng, S. Zhang, and S. Yang, "MPFFNet: LULC classification model for high-resolution remote sensing images with multi-path feature fusion," *International Journal of Remote Sensing*, vol. 44, no. 19, pp. 6089–6116, 2023.
13. Y. Li, Y. Wu, W. Wang, H. Jin, X. Wu, J. Liu, and C. Lv, "Integrating stride attention and cross-modality fusion for UAV-based detection of drought, pest, and disease stress in croplands," *Agronomy*, vol. 15, no. 5, p. 1199, 2025.
14. Y. Liao, S. Kang, J. Li, Y. Liu, Y. Liu, Z. Dong, and X. Chen, "Mobile-seed: Joint semantic segmentation and boundary detection for mobile robots," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3902–3909, 2024.
15. W. Kim and J. Seok, "Indoor semantic segmentation for robot navigating on mobile," in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 22–25, Jul. 2018.