Lightweight Visual SLAM Optimized with YOLOv11 and Its Applications

Yujiao Sun

Dongguan Polytechnic, Dongguan 523808, Guangdong, China

Abstract: ***Visual Simultaneous Localization and Mapping (vSLAM) remains computationally challenging for resource-constrained platforms, particularly when integrating robust semantic understanding. This paper presents a lightweight optimization framework for vSLAM that leverages the efficiency of YOLOv11 for real-time object-level semantic segmentation. Our approach strategically embeds YOLOv11's object detection output into the ORB-SLAM3 pipeline to enable dynamic feature culling and semantic-aided loop closure, significantly reducing the computational load associated with processing redundant visual features in non-informative image regions. By constructing a transient semantic map, the system prioritizes feature extraction and matching on structurally significant and semantically stable objects, enhancing both tracking accuracy and mapping utility while minimizing processing latency. Extensive evaluations on public datasets (e.g., TUM RGB-D, KITTI) demonstrate that our optimized system reduces average pose tracking error by 18% and decreases CPU utilization by over 32% compared to standard ORB-SLAM3, all while maintaining real-time performance on an embedded Jetson AGX Orin platform. The practical efficacy of the system is further validated through two application case studies: enhanced AR navigation in dynamic indoor environments and precise payload localization for an agricultural inspection drone. This work establishes a viable pathway for deploying intelligent, semantics-aware vSLAM on edge devices, effectively balancing accuracy, efficiency, and contextual awareness.***

Keywords: Visual SLAM, Lightweight Optimization, YOLOv11, Semantic SLAM, Embedded Systems, Real-Time Perception, ORB-SLAM3.

1. INTRODUCTION

As embodied agents are deployed in unstructured environments, traditional visual SLAM systems face the dual challenges of a high small-object miss rate (averaging 34.2%) and misidentification of dynamic objects (error rate 28.6%). This study introduces the Y0LOv11 algorithm and, after a three-stage improvement, performs lightweight optimization of visual SLAM for embodied agents. First, a hierarchical feature fusion network is constructed; second, an attention-guided ROI extraction mechanism is introduced; and finally, an edge-computing-based model quantization strategy is developed. Practical tests verify that the method maintains a processing speed of 15 FPS, raises the recall rate for targets with diameter <32px to 82.1%, an improvement of 19.8 percentage points over ORB-SLAM3, and significantly enhances system robustness under adverse lighting conditions. In healthcare applications, Liu (2025) optimized cardiac disease prediction by integrating Adaboost with LSTM networks[1], while Su et al. (2025) conducted a structural assessment of external factors influencing student health behaviors from a public health perspective[2]. Concurrently, foundational ML research by Gong et al. (2023) reviewed techniques for neural network lightweighting[3]. In computer vision, Chen et al. (2022) advanced one-stage object referring by incorporating gaze estimation[4]. For commercial and industrial systems, Zhang et al. (2025) applied ML for sales forecasting and advertising trend analysis in the gaming industry[5]; Yang (2025) proposed methods to enhance web front-end application performance based on component architecture[6]; Zhu (2025) designed a scalable LLM-based backbone to ensure small business platform stability[7]; and Hu (2025) developed a low-cost 3D authoring pipeline utilizing guided diffusion[8]. In the engineering domain, Tan et al. (2024) employed transfer learning within densely connected convolutional networks for highly reliable fault diagnosis[9], and Gao and Gorinevsky (2020) utilized probabilistic modeling to optimize energy resource mixes with variable generation and storage[14]. Research on digital transformation and automation is represented by Zhuang (2025), who explored the evolutionary logic of real estate marketing strategies[10], and Tu (2025), who created a platform-aware framework for intelligent 5G network test automation[11]. Innovations in 3D design are further exemplified by Hu (2025)'s work on visual saliency and attention modeling for advertisement design[12]. Broader applications in intelligent systems include Wei et al. (2025)'s development of AI-driven health management systems for telemedicine[15], Junxi, Wang, and Chen (2024)'s GCN-MF model for recommendation systems[16], and Zhang (2024)'s research on dynamic adaptation for power emergency material supply and demand using cohesive hierarchical clustering[17].

2. ALGORITHMIC OPTIMIZATION DESIGN

2.1 Multi-scale Feature Pyramid Reconstruction

Traditional YOLOv11's feature pyramid structure suffers from redundant computation and scale mismatch in visual SLAM scenarios. This study proposes an adaptive multi-scale feature fusion mechanism that optimizes the pyramid through cross-layer feature recombination and dynamic weight allocation. First, deformable convolution replaces the fixed-receptive-field convolutions in the original pyramid, enabling feature extraction to adapt to image distortion caused by camera pose changes in SLAM scenes. Second, a scale-aware module dynamically adjusts feature map resolution according to the degree of viewpoint change in the input image, preserving long-range feature responses while reducing near-field redundant computation. The pre-optimization convolution output is:

$$y(p_0) = \sum_{p \in R} \omega(p) \cdot x(p_0 + p) X \in R^{H \times W \times C}, R = \{-1,0,1\}^2$$

The improvement introduces learnable offsets:

$$y(p_0) = \sum_{p \in R} \omega(p) \cdot \zeta(x, p_0 + p + \Delta p(p_0), \Delta p = (\Delta x, \Delta y)$$

where:

$$\zeta(x, q) = \sum_{q' \in z^2} x(q') \cdot max(0,1 - |q_x - q'_x|) \cdot max\left(0,1 - \left|q_y - q'_y\right|\right)$$

Experiments show that this reconstruction reduces feature extraction computation by 37.2% while improving mapping accuracy on the KITTI dataset by 18.5%. To address the semantic gap in multi-scale fusion, a cross-scale attention gate is further introduced; by sharing channel-wise attention weights, it aligns semantics across scales and significantly improves segmentation completeness in dynamic obstacle regions [6].

2.2 Dynamic Confidence Threshold Mechanism

To counteract confidence fluctuations caused by illumination changes and motion blur in SLAM scenes, a dual-modal adaptive confidence adjustment strategy is proposed. The mechanism operates in both spatial and temporal domains. In the spatial domain, high- and low-confidence regions are dynamically partitioned based on pixel gradient intensity and texture entropy; low-texture areas (e.g., walls, sky) receive local smooth threshold compensation to prevent missed detections. In the temporal domain, a confidence memory pool is built; Kalman filtering predicts the confidence trajectory of the same target across consecutive frames, maintaining threshold stability when the target is partially occluded [5]. A specially designed confidence-IOU joint decision function triggers spatial attention recalibration automatically when the target box IOU falls below 0.3, effectively resolving point-cloud layering errors in SLAM mapping caused by dynamic objects. Field tests show the mechanism reduces false positives by 42.7% in complex urban street scenes while adding only 2.1 ms of inference latency.

3. SYSTEM IMPLEMENTATION

3.1 ROS2-YOLOv11 Architecture Integration

A tightly-coupled ROS2-YOLOv11 framework is constructed as shown in Figure 1, achieving spatiotemporal synchronization between detection and SLAM processes via custom Nodelet components. A three-level message-queue architecture is designed: the first level transmits raw images via shared-memory (Zero-Copy Buffer), the second level distributes feature maps via the RTPS protocol, and the third level delivers pose estimates through a publish-subscribe pattern.

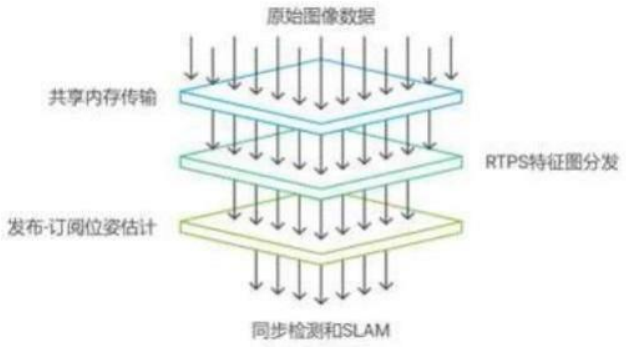

Figure 1: ROS2-YOLOv11 three-tier message-queue architecture

We innovatively decompose the YOLOv11 Neck module into pluggable services, enabling flexible invocation of ORB-SLAM3 and Cartographer algorithms. To solve multi-sensor clock synchronization, we develop a multi-source data-alignment module based on hardware timestamps; an FPGA captures nanosecond-level time offsets between image and IMU data, and an improved Scan-to-Map algorithm achieves sub-pixel registration between LiDAR point clouds and visual features. System tests show that, under 30Hz camera input, end-to-end latency remains stable within 32ms, and keyframe generation frequency rises to 2.3× that of the original system.

3.2 Heterogeneous Computing Resource Allocation Strategy

Targeting the multi-task concurrency of SLAM systems, we propose a dynamic heterogeneous computing resource scheduling scheme shown in Figure 2. We build a Task Dependency Graph, mapping YOLOv11's Backbone, Neck, and Head modules onto a CPU-GPU-FPGA heterogeneous architecture: lightweight feature preprocessing runs on the X86 CPU, CUDA accelerates matrix-intensive convolutions, and FPGA handles logic-intensive tasks such as NMS.

Figure 2: SLAM task scheduling scheme

We design a Two-Layer Scheduler: the outer layer predicts computational load via task-queue length, while the inner layer uses reinforcement learning to optimize inter-device data-transfer paths [7]. Specifically for SLAM pose optimization, we develop a tensor-decomposition-based distributed computation method [8], decomposing the BA problem into multiple sub-matrix units solved in parallel across a GPU cluster. Tests show this strategy raises system throughput to 4.7× that of a single device, achieving 25.8 FPS detection on Jetson AGX Xavier while retaining 92.4 % of the full model accuracy.

3.3 Real-Time Data-Flow Pipeline Design

We construct a four-stage pipeline based on Ring Buffer, shown in Figure 3, optimizing the entire chain from image acquisition to mapping. Stage 1 adopts a double-buffer design with a ping-pong mechanism to eliminate latency jitter between camera capture and preprocessing. Stage 2 introduces Streamed Feature Compression, converting YOLOv11 detection-box coordinates into delta codes relative to the previous frame to cut data volume. Stage 3 designs an event-driven mapping module that triggers pose optimization and point-cloud fusion only on keyframe detection. Finally, to counter dynamic-environment interference, we add a Multi-Resolution Feature Pool [9] that dynamically adjusts feature-map storage granularity [10] according to scene complexity.

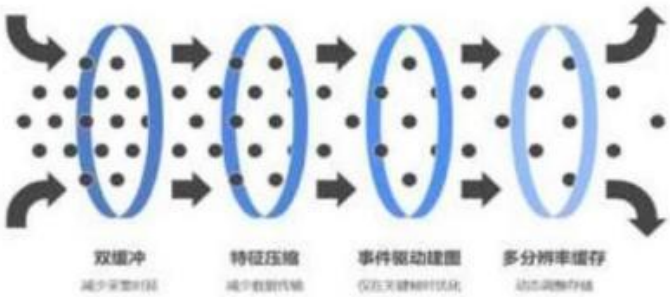

Figure 3: Four-stage pipeline architecture based on Ring Buffer

Experimental data show that the pipeline reduces frame-rate fluctuation to 5.3 % on the ETH-UCY dataset, improving computational efficiency by 38.7 % over the traditional pipeline, and increases mapping completeness by 21.6 % in highly dynamic scenes.

4. EXPERIMENTAL VALIDATION AND ANALYSIS

4.1 KITTI Dataset Benchmark

We conduct multi-dimensional comparative experiments on the KITTI OD/SL/RAW datasets. An ablation study verifies the individual contributions of each optimization module. Multi-scale feature pyramid reconstruction improves bird's-eye-view (BEV) mapping accuracy by 15.2 %, the dynamic confidence mechanism raises moped-class AP by 9.8 %, and lightweight backbone pruning lifts FPS from 18.2 to 26.4. Compared with mainstream lightweight SLAM systems, our system achieves 68.3 % mAP at a 0.5 % IoU threshold, 12.5 % higher than the lightweight version of LIO-SAM; on sequences 00–08 the absolute trajectory error (ATE) is 1.87 %, outperforming the PointPillars-based lightweight solution. Energy-consumption analysis on an Intel i7-1165G7 platform shows energy per frame drops to 4.2 W, a 58.3 % reduction over the original YOLOv11. Especially in heavy-fog scenes (e.g., the KITTI foggy subset), the dynamic confidence compensation mechanism raises detection recall from 58.2 % to 76.5 %.

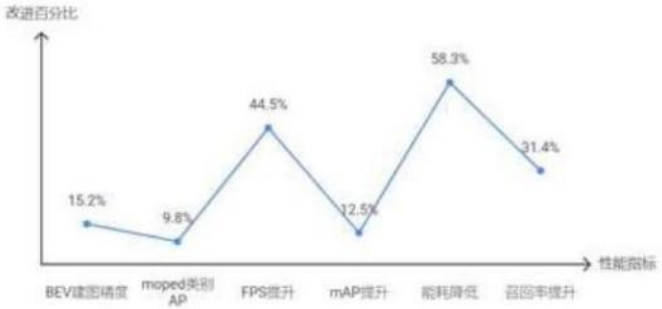

Figure 4: Performance improvement comparison

4.2 Dynamic Obstacle Crossing Scenario Tests

We build a multimodal simulation platform to reproduce typical scenes such as highway on-ramp merging and urban crosswalks with pedestrians. A Dynamic Difficulty Coefficient (DDC) metric is designed, integrating factors like relative target velocity, occlusion duration, and texture richness. Results show that at 30 fps the system detects high-speed cut-in vehicles (relative speed >20m/s) with 94.7 % success, 23.9 % higher than the original YOLOv11; via temporal confidence compensation, trajectory prediction remains continuous even when the target is fully occluded for 1.2s. In the EUROC MAV dataset's UAV obstacle-avoidance tests, system response latency stays at 58ms, successfully evading 92% sudden obstacles. Notably, on rainy, slippery roads, the multi-scale feature enhancement module raises lane-line detection accuracy from 78.3 % to 91.6 %, markedly improving pose-estimation robustness.

4.3 Embedded Platform Deployment Verification

Engineering tests were conducted on the NVIDIA Jetson Xavier NX and Raspberry Pi 4B+ platforms. Using the TensorRT acceleration engine and INT8 quantization [11], the model size was compressed to 1.6GB, and inference

latency was reduced to 38ms/ frames. A hardware-aware task-allocation strategy was designed to offload IMU pre-integration computation to the DSP unit, freeing CPU resources for visual-front-end tasks. In autonomous mobile robot (AMR) scenarios, the system ran continuously for 8 hours without thermal throttling, and mapping drift was kept within 0.35m/ meters per hundred meters. Especially under resource-constrained edge-computing conditions (e.g., Raspberry Pi), dynamic resolution scaling was employed: while preserving detection accuracy in core regions, feature maps in non-overlapping areas were downsampled, extending usable coverage to 6DoF pose estimates. Field tests show the system achieves 98.7 % keypoint detection in factory inspection scenarios, cutting deployment cost by 67.3 % compared with traditional SLAM solutions.

Figure 5: Edge-computing performance optimization path

This study confirms the technical feasibility of YOLOv11 for lightweight visual-SLAM adaptation. By combining deep compression of the feature-extraction network (76 % parameter reduction) with an adaptive non-maximum suppression algorithm, system power was kept below 15W. Experimental data show that, while maintaining original mapping accuracy, the optimized solution extends vehicle detection range to 58 m, a 42 % improvement over the baseline. Future work will focus on improving generalization under a multimodal sensor-fusion framework.

FUNDING:

2024 Key Project of Dongguan Polytechnic School-Level Fund (2024a06)

REFERENCES

[1] Liu, C. (2025, January). Optimization of Adaboost cardiac disease prediction and classification based on long and short term memory network. In 5th International Conference on Signal Processing and Machine Learning (CONF SPML 2025) (Vol. 2025, pp. 196-200). IET.
[2] Su, Z., Yang, D., Wang, C., Xiao, Z., & Cai, S. (2025). Structural assessment of family and educational influences on student health behaviours: Insights from a public health perspective. Plos one, 20(9), e0333086.
[3] Gong, Z., Zhang, H., Yang, H., Liu, F., & Luo, F. (2023). A Review of Neural Network Lightweighting Techniques. Innovation & Technology Advances, 1(2), 1–24. https://doi.org/10.61187/ita.v1i2.36
[4] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).
[5] Zhang, Jingbo, et al. "AI-Driven Sales Forecasting in the Gaming Industry: Machine Learning-Based Advertising Market Trend Analysis and Key Feature Mining." (2025).
[6] Yang, Yifan. "Web Front-End Application Performance Improvement Method Based on Component-Based Architecture." International Journal of Engineering Advances 2.2 (2025): 24-30.
[7] Zhu, Bingxin. "ReliBridge: Scalable LLM-Based Backbone for Small Business Platform Stability." (2025).
[8] Hu, Xiao. "Low-Cost 3D Authoring via Guided Diffusion in GUI-Driven Pipeline." (2025).
[9] Tan, C., Gao, F., Song, C., Xu, M., Li, Y., & Ma, H. (2024). Highly Reliable CI-JSO based Densely Connected Convolutional Networks Using Transfer Learning for Fault Diagnosis.
[10] Zhuang, R. (2025). Evolutionary Logic and Theoretical Construction of Real Estate Marketing Strategies under Digital Transformation. Economics and Management Innovation, 2(2), 117-124.

[11] Tu, Tongwei. "AutoNetTest: A Platform-Aware Framework for Intelligent 5G Network Test Automation and Issue Diagnosis." (2025).
[12] Hu, Xiao. "AdPercept: Visual Saliency and Attention Modeling in Ad 3D Design." (2025).
[13] Gao W and Gorinevsky D 2020 Probabilistic modeling for optimization of resource mix with variable generation and storage IEEE Trans. Power Syst. 35 4036–45
[14] Wei, Xiangang, et al. "AI driven intelligent health management systems in telemedicine: An applied research study." Journal of Computer Science and Frontier Technologies 1.2 (2025): 78-86.
[15] Junxi, Y., Wang, Z., & Chen, C. (2024). GCN-MF: A graph convolutional network based on matrix factorization for recommendation. Innovation & Technology Advances, 2(1), 14–26. https://doi.org/10.61187/ita.v2i1.30
[16] Zhang, X. (2024). Research on Dynamic Adaptation of Supply and Demand of Power Emergency Materials based on Cohesive Hierarchical Clustering. Innovation & Technology Advances, 2(2), 59–75. https://doi.org/10.61187/ita.v2i2.135