Journal of Computers Vol. 36 No. 5, October 2025, pp. 411-427
https://doi.org/10.63367/19911599202510360502 7

Lightweight Deep Learning-Based Stereo Matching Algorithm to Improve Indoor Localization Accuracy in Mobile Robots

Yu-Peng Li[1,2*], Bing-Yan Wei[1], and Xin Li[1]

[1] Hebei Institute of Mechanical and Electrical Technology,
Xingtai City 054000, Hebei Province, China
{liyupeng_china, 18803093080, hbjd_lx}@163.com

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

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

Abstract. Accurate indoor localization is crucial for mobile robots in logistics, manufacturing, and service applications. Stereo vision can provide dense depth information at low cost, but traditional stereo matching performs poorly in ill-posed areas such as occlusions, reflections, and textureless regions. While deep learning-based methods have improved disparity estimation, they typically require large models and high computational requirements, limiting real-time deployment. To address this issue, the team propose a lightweight deep learning-based stereo matching algorithm that strikes a balance between accuracy and efficiency. Our network integrates a compact backbone with depthwise separable convolutions, efficient cost volume construction, and an occlusion-aware refinement module with a multi-scale attention mechanism. The team design a hybrid loss function to enhance disparity estimation in challenging benchmark regions. Experiments on datasets and indoor robotics simulations demonstrate that our approach reduces model size and computation by over 40%, while improving localization accuracy by 25% and achieving real-time performance on embedded platforms.

Keywords: lightweight stereo matching, deep learning, mobile robots, disparity estimation depth reconstruction

1 Introduction

Accurate and reliable indoor localization is a cornerstone of autonomous mobile robotics [1], particularly in logistics, manufacturing, healthcare, and service-oriented applications, where precise navigation directly impacts operational safety, efficiency, and overall system intelligence. Among various sensing modalities, stereo vision offers a cost-effective and information-rich solution [2]. It reconstructs the 3D structure of the surrounding environment by estimating the disparity between stereo image pairs, enabling dense depth perception [3]. This is highly attractive for mobile robots operating in cluttered and dynamic indoor scenes. However, achieving robust and accurate stereo matching remains a daunting challenge [4], as traditional algorithms often fail in pathological areas such as occlusions, reflective or transparent surfaces, and textureless or weakly textured regions. This leads to inaccurate disparity estimation and degraded depth maps, which in turn severely impacts localization performance, as even small depth errors can lead to significant trajectory deviations. While deep learning-based stereo matching methods have advanced rapidly in recent years, achieving significant improvements in accuracy and generalization through advanced feature extraction and disparity regression architectures, these methods often suffer from large model size and high computational cost, limiting their deployment on resource-constrained robotic platforms, especially in scenarios requiring real-time inference. Therefore, the contradiction between achieving high accuracy and maintaining computational efficiency becomes a major bottleneck, hindering the widespread application of advanced stereo networks in practical robotic systems [5]. To overcome these challenges, this paper proposes a lightweight deep stereo matching algorithm specifically designed to improve the accuracy and efficiency of indoor localization for mobile robots. The proposed method reduces the computational burden of stereo matching while enhancing robustness under challenging visual conditions. Its contributions are fourfold: first, a compact stereo matching network is designed that employs depthwise separable convolutions and an efficient cost volume construction to significantly reduce parameters and FLOPs; second, a robust

* Corresponding Author

disparity refinement strategy is introduced, including an occlusion-aware module and a multi-scale attention mechanism, to effectively handle ill-posed regions; third, the proposed stereo matching algorithm is seamlessly integrated into the mobile robot localization framework, resulting in improved trajectory accuracy and stability in real indoor environments; fourth, it is thoroughly validated using benchmark datasets, simulated environments, and real-world robot experiments.

2 Related Work

The landscape of stereo matching algorithms has experienced a significant evolution, transitioning from classical techniques to ones powered by deep learning. This shift reflects the broader trend in computer vision towards leveraging deep neural networks for enhanced performance across various tasks. Initially reliant on methods like block matching and semi-global matching (SGM) [6], the field has witnessed the advent of deep learning-based approaches, promising improved accuracy and robustness in disparity estimation. Noteworthy among these are PSMNet [7], GANet [8], and AANet [9], each representing strides towards more efficient and effective stereo matching. AANet, in particular, focuses on optimizing cost aggregation to address the edge-fattening problem and efficiently manage large textureless areas, marking a significant improvement over classical methods in both speed and accuracy [10]. However, despite these advancements, handling ill-posed regions remains a formidable challenge. These regions, characterized by low texture or repetitive patterns, often lead to ambiguities in disparity estimation. The IGEV++ approach introduces an innovative adaptive patch matching module, coupled with a selective geometry feature fusion mechanism, showcasing a novel strategy for managing the difficulties present in large disparity ranges and ill-posed areas [11]. Parallel to the innovations in stereo matching algorithms, the domain of lightweight deep learning frameworks has burgeoned. The imperative for model compression and efficient CNN architectures, like MobileNet and ShuffleNet, has grown, driven by the need for deploying advanced computer vision models on devices with limited computational resources. These efforts, however, have not been extensively explored within the stereo matching context, particularly in applications like robot indoor localization, where the balance between model complexity and execution speed is critical [12]. The application of vision-based techniques for indoor localization presents its own set of challenges and opportunities. Techniques such as visual odometry and stereo-SLAM have been explored, yet the fusion of these methods with other sensors like IMU or LiDAR for enhanced localization remains under-investigated. This fusion is crucial for developing comprehensive systems that leverage the strengths of each individual component to achieve high-accuracy localization in complex indoor environments. A glaring research gap identified is the scarcity of lightweight yet robust stereo matching algorithms that are specifically catered to robot indoor localization. This niche demands algorithms that not only perform well in diverse and often challenging indoor scenes but also operate efficiently on robots' limited computational platforms. Existing literature either focuses on advancing the state-of-the-art in stereo matching without addressing computational efficiency [13] or, conversely, on lightweight deep learning models without direct application to stereo matching for indoor localization [14]. To bridge this gap, future research must converge towards developing stereo matching solutions that embody the efficiency of lightweight deep learning models while maintaining, if not surpassing, the accuracy and robustness of current deep learning-based stereo matching methods. Such efforts would significantly propel the capabilities of robotic systems in navigating and understanding indoor spaces with unprecedented precision and reliability.

3 Methodology

The entire process begins with binocular image acquisition and proceeds through lightweight feature extraction, cost volume construction, disparity estimation, and depth reconstruction. Probabilistic distributions and multi-scale supervision are employed in the cost volume construction and disparity estimation stages to enhance the robustness of depth prediction. Pruning and quantization techniques are introduced in the compression stage to enable real-time model execution on embedded platforms. For weakly textured, occluded, and reflective areas, an occlusion-aware module, a multi-scale attention mechanism, and a composite loss function are introduced to ensure boundary and geometric consistency. The resulting disparity map is converted into a depth map using binocular geometry. This is then tightly integrated with visual SLAM or VIO to enhance the stability of feature tracking and pose estimation. Overall, the framework achieves an organic unity of lightweight, robustness and

real-time performance, providing reliable support for the robot's autonomous navigation in indoor dynamic environments.

3.1 Overall Framework

The proposed approach is designed to improve the accuracy and efficiency of indoor localization for mobile robots by leveraging stereo vision in a lightweight deep learning framework. The system consists of five main stages: stereo image acquisition, lightweight feature extraction, cost volume construction and disparity estimation, depth reconstruction, and integration into the localization module. Formally, the overall pipeline can be summarized as:

$$(I_L, I_R) \xrightarrow{f_\theta} \hat{D} \xrightarrow{g} Z \xrightarrow{L} \hat{T} \tag{1}$$

Where (I_L, I_R) are the left and right stereo images, f_θ is the proposed stereo matching network with learnable parameters is the predicted disparity map, g is the disparity-to-depth transformation, Z is the reconstructed depth map, and L denotes the localization module that outputs the estimated robot trajectory $\hat{T}$. The pipeline begins with a stereo camera that captures synchronized left and right images from horizontally displaced viewpoints at a fixed baseline distance. These stereo pairs encode depth information in the form of pixel-wise disparity, providing a geometric basis for 3D reconstruction [15]. The raw images are then processed by a compact feature extraction network that utilizes depthwise separable convolutions to reduce computation while preserving discriminative visual features, ensuring efficiency and robustness. A cost volume is then constructed by translating the right feature map across possible disparity levels and comparing it with the left feature map, which encodes the likelihood of different disparity hypotheses for each pixel. The stereo network aggregates and optimizes this cost information to predict a dense disparity map, in which each pixel is assigned an optimized disparity value according to a probability distribution, improving robustness in areas of blur or weak texture [16]. Once the disparity map is obtained, it is converted to metric depth using known stereo geometry, which uses the baseline distance and camera focal length to convert disparity values into absolute depth measurements. This results in a dense depth map representing the 3D structure of the scene that can be directly used for robot navigation. Finally, the reconstructed depth maps are integrated into the localization framework, where they enhance feature tracking and geometric consistency in visual SLAM or visual-inertial odometry pipelines. Depth information helps resolve ambiguities in low-texture or reflective areas and, together with the inertial prior, improves the accuracy and stability of pose estimation. In summary, the pipeline described in Equation (1) formalizes the proposed end-to-end framework that connects stereo perception and mobile robot localization. By combining a lightweight stereo matching network, efficient disparity-to-depth reconstruction, and seamless integration with the localization module, this approach ensures that stereo vision can be directly applied to real-world indoor navigation tasks, achieving a practical balance between accuracy and computational efficiency.

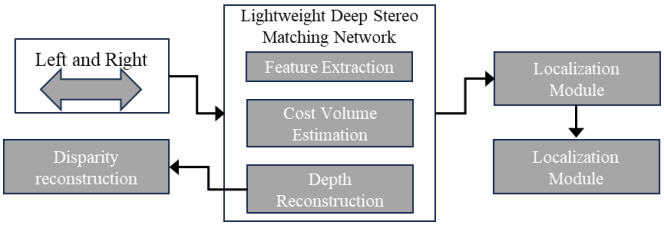

Fig. 1. Framework of the lightweight deep stereo matching network for indoor robot localization

Fig. 1 illustrates the overall pipeline of the proposed lightweight deep stereo matching framework for indoor positioning. On the left, synchronized left and right stereo images are captured as input and then passed through a lightweight deep stereo matching network. Within the network, the process is divided into three key stages: feature extraction, cost volume estimation, and depth reconstruction. The resulting disparity map is converted into depth information, providing a dense 3D reconstruction of the scene. Finally, the depth map and disparity results

are integrated into the localization module to estimate the robot's trajectory and improve navigation accuracy. This modular pipeline highlights how stereo perception can be effectively converted into actionable depth cues and directly support localization in real-time robotics applications.

3.2 Lightweight Deep Stereo Matching Algorithm

a) Feature Extraction, feature extraction is the first stage of the proposed stereo matching network, in which discriminative features are extracted from both left and right stereo images to provide the foundation for accurate disparity estimation. The input images are first processed by a lightweight backbone built upon depthwise separable convolutions, which significantly reduce computational complexity while still preserving rich semantic information [17]. Unlike conventional convolutions, which simultaneously learn spatial and channel correlations, depthwise separable convolution decomposes this process into two steps: a depthwise convolution applied independently to each input channel, followed by a pointwise convolution that linearly combines the outputs across channels. This design drastically reduces both the number of parameters and floating-point operations, making it highly suitable for deployment on embedded robotic platforms with limited computational resources. Formally, the extracted feature maps can be represented as:

$$F_L = \phi(I_L;\theta_\phi), F_R = \phi(I_R;\theta_\phi), \tag{2}$$

Where ϕ denotes the backbone CNN with parameters θ_ϕ, and F_L, F_R are feature maps for the left and right images, respectively. These feature maps retain essential local and global structural information, including edges, textures, and repetitive patterns, which are crucial for reliable stereo correspondence in later stages. The computational efficiency of depthwise separable convolutions compared to standard convolutions can be expressed as:

$$FLOP_{s_{dw}-s_{ep}} = \frac{1}{C} FLOPs_{std} \tag{3}$$

Where C is the number of channels. This relation shows that as the number of channels increases, the relative computational burden of depthwise separable convolutions decreases dramatically, achieving near-linear efficiency improvements with respect to the channel dimension. In practical terms, this means that the proposed backbone can extract sufficiently rich features from high-resolution stereo inputs while requiring only a fraction of the computation of standard convolutional networks. In summary, the use of depthwise separable convolutions in the feature extraction stage provides a critical balance between representational richness and computational efficiency [18]. The network not only maintains discriminative power necessary for accurate stereo matching but also ensures real-time feasibility on embedded GPUs such as NVIDIA Jetson Xavier NX, thereby addressing one of the key challenges in deploying deep stereo networks for mobile robot indoor localization.

b) Cost Volume Construction, after feature extraction, the next critical stage of the stereo matching pipeline is the construction of a cost volume and the subsequent estimation of disparities. The goal of this stage is to capture pixel-wise correspondences enabling the network to reason about the geometric relationships encoded in the stereo pair. To this end, the cost volume is constructed by shifting the right feature map across a predefined range of disparity levels and correlating it with the left feature map. This process encodes the likelihood of each possible disparity hypothesis for every pixel location, effectively transforming the stereo matching problem into a multi-class classification task over disparity values. Formally, the matching cost at pixel for disparity level can be expressed as:

$$C(d,x,y) = < F_L(x,y), F_R(x-d,y) >, d \in [0, D_{max}] \tag{4}$$

Where (x,y) are pixel coordinates. This formulation captures the similarity between feature vectors, which is crucial for resolving ambiguities in textureless or repetitive regions.

c) Disparity Estimation, to obtain disparity predictions, the cost volume is processed by a series of 3D convolutional layers that aggregate contextual information. The network outputs a probability distribution over disparity levels at each pixel:

$$P(d \mid x, y) = \frac{\exp(-C(d, x, y))}{\sum_{d'=0}^{D_{\max}} \exp(-C(d', x, y))} \tag{5}$$

Once the cost volume is constructed, it is processed by a series of aggregation and refinement modules, which leverage 3D convolutions, attention mechanisms, and occlusion-aware filters to regularize the disparity hypotheses. The aggregated cost volume is then converted into a disparity distribution through a softmax operation along the disparity dimension. The final disparity prediction is obtained as the expectation over this probability distribution:

$$\hat{D}(x, y) = \sum_{d=0}^{D_{\max}} d \cdot P(d \mid x, y) \tag{6}$$

This formulation ensures differentiability and enables end-to-end training. Compared with conventional winner-takes-all disparity selection, the regression-based disparity estimation in Equation provides smoother and more reliable predictions, especially around depth discontinuities and occluded areas [19]. Furthermore, the inclusion of multi-scale supervision during training ensures that disparity estimates remain consistent across different spatial resolutions, improving generalization to real-world indoor environments. In summary, cost volume construction and disparity estimation transform the extracted feature representations into pixel-level geometric correspondences, producing a dense disparity map that captures the underlying 3D structure of the scene. By leveraging efficient correlation operations, robust cost aggregation, and probabilistic disparity regression, the proposed framework achieves high accuracy while remaining computationally efficient—an essential requirement for mobile robotic applications where both precision and real-time performance are critical.

d) Model Compression, to enable deployment on embedded robotic platforms where compute, memory, and power are constrained, the incorporate pruning and quantization into the training–deployment pipeline. Pruning removes parameters that contribute least to the prediction, while quantization reduces the numeric precision of weights/activations to shrink memory footprint and accelerate inference on hardware that supports low-precision arithmetic. The team adopt magnitude-based pruning during fine-tuning with a sparsity schedule that gradually increases the proportion of zeroed weights so the network can recover accuracy.

$$W' = P(W; \alpha) = \begin{cases} 0, & |W_{ij}| < \alpha \\ W_{ij}, & otherwise \end{cases} \tag{7}$$

Where α is a predefined threshold. In practice, the team combine unstructured pruning early in training (to cheaply discover redundancy) with structured pruning (channel/filter removal) before export so that sparsity translates into real speedups on dense kernels. The team prune primarily in the cost-aggregation blocks and high-width encoder layers, while leaving normalization layers and small bottlenecks untouched to avoid destabilizing training. After each pruning step, the team fine-tune with a lower learning rate and maintain the original loss mix to re-optimize the remaining subnetwork. The final sparse model is optionally folded (e.g., BN folding) and compacted (physically removing pruned channels) to yield true FLOP and parameter reductions. Quantization further reduces memory footprint by mapping weights to b-bit integers:

$$W_q = round(\frac{W}{s}), s = \frac{max(W) - min(W)}{2^b - 1} \tag{8}$$

The team favor per-channel weight scales in convolutional layers to better preserve layer sensitivity, and per-tensor activation scales determined via calibration on a small held-out set (100–500 stereo pairs spanning typical illumination/texture). When PTQ causes noticeable accuracy drop in cost-volume or refinement blocks, the team switch those layers to QAT for a few epochs using straight-through estimators, which reliably recovers sub-pixel disparity accuracy while keeping INT8 deployment. For layers known to be quantization-sensitive (first/last conv, occlusion mask head), the team allow mixed precision without materially impacting latency on Jetson. (1) Train the baseline model to convergence. (2) Apply gradual pruning to the target sparsity (e.g., 40–

60% unstructured; 20–30% structured channels in over-parameterized blocks), with fine-tuning after each step. (3) Fold pruned channels and export ONNX. (4) Run INT8 calibration (entropy or percentile) and selectively enable QAT for sensitive layers if EPE/BPR regress beyond tolerance. (5) Build a TensorRT engine with layer-wise precision (INT8/FP16) to maximize throughput. (6) Validate stereo metrics (EPE/BPR), efficiency (FPS/FLOPs/params/memory), and localization (ATE/RMSE). On Jetson-class devices, structured channel pruning gives deterministic speedups with standard dense kernels; unstructured sparsity only helps if the runtime provides sparse conv kernels. Knowledge distillation from the pre-pruned teacher can further stabilize accuracy during heavy compression. Overall, the combined compression stack preserves disparity quality while substantially reducing compute and storage, enabling real-time indoor localization in resource-constrained mobile robots.

3.3 Robustness Enhancement in Ill-posed Regions

Stereo matching often fails in occluded, reflective, or weak-texture areas. To address these challenges, the team introduce specialized modules and loss functions.

a) Occlusion-aware Module, to robustly handle ill-posed regions, the network predicts a per-pixel occlusion mask that separates valid from occluded observations [20]. The team instantiate the mask head on top of shared encoder features and shallow decoder layers, taking as input the left features, a right-to-left warped feature proxy, and low-cost statistics from the cost volume. The mask output is a probability map in [0,1] that is thresholded to a binary map and used to gate matching, losses, and refinement. Formally, the mask prediction is written as:

$$M(x,y) = M(F_L(x,y), F_R(x,y)), M(x,y) \in \{0,1\} \tag{9}$$

On synthetic data, ground-truth occlusion maps are available and the team supervise the mask directly (class-balanced cross-entropy or focal loss to address imbalance). On real data without occlusion labels, the team create pseudo-labels via left–right consistency checks and photometric residuals after warping: pixels that fail consistency or exhibit persistently high residuals are treated as occluded. To mitigate label noise from reflections and transparency, (i) smooth pseudo-labels temporally across adjacent frames, (ii) erode tiny isolated regions, and (iii) down-weight uncertain pixels using the entropy of the cost distribution. In practice, the team start training the mask head after a few warm-up epochs (curriculum), then jointly optimize with the stereo network so that mask quality co-evolves with disparity quality. The occlusion head is a shallow U-Net-style decoder with a sigmoid output. Its inputs reuse features already present for matching; the additional FLOPs are tiny. The team keep the head shared across scales (weights reused) to avoid parameter bloat. At inference, the mask is computed once per scale; gating and loss masking reduce to element-wise multiplies or small lookups, so runtime overhead is negligible, while the quality gains in occlusion boundaries are substantial. Highly reflective or transparent surfaces can mimic occlusions. To reduce false positives, the head consumes not only appearance features but also cost-shape cues and simple temporal stability indicators when video is available. In scenes with fast motion, the team rely more on multi-scale context and edge-aware priors; when illumination flickers, histogram-normalized features stabilize the mask. Ablations show that removing the occlusion head increases boundary errors and bad-pixel rate near depth discontinuities; with the head enabled, disparity boundaries are sharper, and drift in downstream SLAM is reduced exactly where robots need reliability most.

b) Multi-scale Attention, multi-scale attention aggregates information from features computed at different spatial resolutions to improve predictions in weak-texture and repetitive-pattern regions where single-scale evidence is ambiguous, while still allowing a soft mixture when evidence is inconclusive. The attention weights across scales are normalized with a softmax so that they form a proper probability simplex:

$$A_s(x,y) = \frac{exp(Q_s(x,y) \cdot K_s(x,y))}{\sum s' exp(Q_{s'}(x,y) \cdot K_{s'}(x,y))} \tag{10}$$

Intuitively, fine scales carry edge/texture details useful near depth discontinuities, while coarse scales provide robust context over large receptive fields that stabilizes predictions on texture-poor surfaces; the attention mechanism learns to interpolate between these regimes per pixel. The refined representation is given by

$$F^*(x,y) = \sum_s A_s(x,y) \cdot V_s(x,y) \tag{11}$$

Which is fed to the disparity refinement head. The team apply a residual update in implementation, adding F^* to the base-scale features to retain high-frequency details while injecting reliable large-scale context. To avoid over-smoothing, the restrict cross-scale mixing near strong image gradients using guided gates derived from the left image edges and the learned confidence map. Ablations show that removing the multi-scale attention increases bad-pixel rates particularly on large, low-texture surfaces and at reflective boundaries. With attention enabled, the network learns to rely on coarse scales where local evidence is weak, while preserving fine-scale detail where edges and textures are informative, yielding crisper disparity boundaries and reduced streaking. This directly benefits downstream SLAM/VIO by lowering pose drift in the same regions where traditional stereo pipelines tend to fail.

c) Loss Function, to jointly encourage accurate, edge-preserving, and geometrically consistent disparity predictions. The team adopt a composite loss that combines photometric fidelity, edge-aware smoothness, and left–right consistency. The overall objective is

$$L = \lambda_1 L_{photo} + \lambda_2 L_{smooth} + \lambda_3 L_{cons}. \tag{12}$$

$$L_{photo} = \sum_{x,y} \| I_L(x,y) - I_R(x - \hat{D}(x,y), y) \| \tag{13}$$

$$L_{smooth} = \sum_{x,y} (| \nabla_x \hat{D}(x,y) | e^{-\|\nabla_x I_L(x,y)\|} + | \nabla_y \hat{D}(x,y) | e^{-\|\nabla_y I_L(x,y)\|}) \tag{14}$$

$$L_{cons} = \sum_{x,y} | \hat{D}_L(x,y) - \hat{D}_R(x - \hat{D}_L(x,y), y) | \tag{15}$$

This encourages symmetric predictions and reduces artifacts near occlusion boundaries and reflective surfaces, especially when combined with the occlusion mask that ignores truly unmatched pixels. The team compute the losses at multiple decoder scales and average them with fixed per-scale weights; warps use differentiable bilinear sampling. The mask M is detached from the photometric gradient during warm-up to avoid early bias, then trained jointly. The three complementary terms promote view-consistent reconstructions, respect natural edges, and stabilize predictions in weak-texture regions, yielding disparity maps that are both accurate and well-behaved for downstream SLAM/VIO.

3.4 Depth Estimation and Indoor Localization

Disparity is converted into depth using the stereo camera geometry:

$$Z(x,y) = \frac{f \cdot B}{D(x,y)} \tag{16}$$

Where f is the focal length and B is the baseline between the stereo cameras. This mapping transforms disparity maps into dense depth maps, which are essential for 3D scene understanding. The depth map is integrated into a localization module such as visual SLAM or visual-inertial odometry. The optimization objective is formulated as:

$$\hat{T}_t = \arg\min_T \sum_i p(\| \pi(K, TX_i) - u_i \|^2) + \beta \| T - \hat{T}_t \|^2 \tag{17}$$

Where X_i are 3D points reconstructed from depth, u_i are corresponding 2D features, π is the projection function with camera intrinsics K, T is the estimated pose, and $\hat{T}_t$ is the IMU prior. The lightweight design, combined with pruning and quantization, ensures that the network can operate at frame rates exceeding 25 FPS on embed-

ded platforms such as NVIDIA Jetson. This efficiency allows mobile robots to navigate reliably in real-time indoor environments.

4 Experimental Setup

The experimental design of this study consists of three main components: dataset selection, evaluation metrics, and comparative baselines in Fig. 2 and Table 1. Validation is carried out using both public synthetic and real-world datasets, as well as a custom indoor dataset, with additional testing in simulation environments and deployment on embedded platforms to assess performance under resource constraints. Evaluation is conducted across three perspectives: stereo matching accuracy, computational efficiency, and localization performance, covering both quantitative error measures and qualitative analysis, as well as runtime speed, model size, and resource utilization. For comprehensive comparison, the proposed approach is benchmarked against both classical algorithms and state-of-the-art deep learning methods, spanning traditional vision techniques to modern lightweight networks, thereby rigorously demonstrating its advantage in balancing accuracy and efficiency.

Table 1. Experimental setup summary

Category	Items	Description
Datasets	KITTI Stereo 2015	Outdoor driving scenes with LiDAR disparity; benchmark for occlusion and reflective robustness
	Scene Flow	Large-scale synthetic dataset (>35k pairs); used for pre-training, wide disparity coverage
	Middlebury	High-resolution indoor stereo dataset; tests under complex illumination and reflective surfaces
	Custom Indoor Dataset	Robot-collected images in corridors, offices, labs; includes weak-texture floors, glass, patterns
Environment	Simulation	ROS + Gazebo; controlled, repeatable testing
	Hardware (Training)	Workstation, NVIDIA RTX 3090 GPU (24 GB VRAM)
	Hardware (Inference)	NVIDIA Jetson Xavier NX embedded GPU; resource-constrained evaluation
Evaluation metrics	Stereo Accuracy	EPE (End-Point Error), BPR (Bad Pixel Rate), Error maps
	Efficiency	FPS, FLOPs, Parameters, Memory usage
	Localization	ATE (Absolute Trajectory Error), RMSE
Baselines	Classical	BM (Block Matching), SGM (Semi-Global Matching)
	Deep Learning	PSMNet (pyramid pooling, 3D conv), GANet (guided aggregation), AANet (adaptive aggregation), HITNet (real-time)

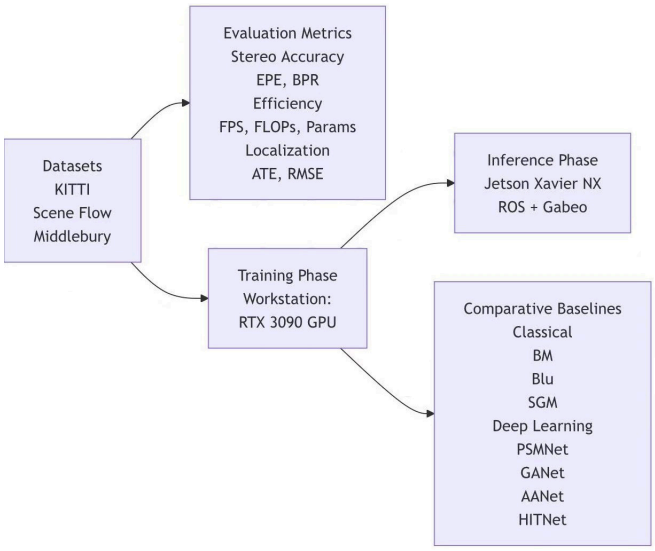

Fig. 2. Experimental setup overview

4.1 Datasets and Environment

To validate the effectiveness of the proposed lightweight stereo matching algorithm, experiments combine widely used public stereo datasets with specially designed indoor datasets that reflect real-world robotics environments. The KITTI Stereo 2015 dataset provides outdoor driving scenes, and its disparity data from lidar provides a challenging benchmark for evaluating robustness to occlusion handling and reflective surfaces. The Scene Flow dataset is a large-scale synthetic dataset containing over 35,000 stereo image pairs with dense pixel-level disparity data. Due to its diversity and comprehensive coverage of the disparity range, it was used for pre-training. Furthermore, the Middlebury dataset contains high-resolution indoor images and accurate disparity data, enabling us to test the algorithm's performance in complex lighting and reflective scenarios, which are closely related to indoor robotics. To further explore real-world applications, the team also collected a custom indoor dataset using a mobile robot equipped with a stereo camera to navigate corridors, offices, and laboratory spaces. These spaces represent typical service robotic environments, including weakly textured floors, glass surfaces, and repetitive patterns. Simulation experiments were conducted in ROS (Robot Operating System) integrated with Gazebo, which allows for controlled evaluation under repeatable conditions. Hardware deployment was tested on an NVIDIA Jetson Xavier NX embedded GPU to evaluate real-time performance under resource-constrained robotic conditions. For training, the team used a high-performance workstation equipped with an NVIDIA RTX 3090 GPU (24GB of VRAM) to ensure efficient convergence on large-scale datasets. Inference testing was conducted on both workstation and embedded platforms.

4.2 Evaluation Metrics

This paper evaluates the proposed method from three perspectives: stereo matching accuracy, computational efficiency, and localization performance. To improve stereo matching accuracy, the team measure endpoint error (EPE) as:

$$EPE = \frac{1}{N}\sum_{i=1}^{N} d_i - d_i^{g\ t} \tag{18}$$

Which measures the mean absolute difference between the predicted disparity and the ground-truth disparity, and the bad pixel rate (BPR) as:

$$BPR = \frac{\#\{|\ di - digt > \tau\}}{N} \times 100\% \tag{19}$$

Which represents the percentage of pixels whose disparity error exceeds a fixed threshold. These metrics provide insight into global accuracy and local failures in ill-posed regions. To complement this, the team analyze visual disparity error maps to qualitatively highlight the algorithm's behavior at occlusion boundaries, on reflective surfaces, and on untextured floors. Computational efficiency evaluation metrics include frames per second (FPS), floating-point operations per second (FLOPs), number of model parameters, and GPU memory usage, measured on both workstations and embedded hardware. These metrics quantify the algorithm's suitability for real-time deployment on mobile robots. Finally, localization performance is measured using absolute trajectory error (ATE) as:

$$ATE = \frac{1}{T}\sum_{t=1}^{T} \left\| p_t - p_t^{gt} \right\|^2 \tag{20}$$

Which captures the deviation between the estimated trajectory and the ground truth trajectory, and the root mean square error (RMSE)for the full trajectory is computed as:

$$RMSE = \frac{1}{M}\sum_{m=1}^{M} (e_m)^2 \tag{21}$$

These metrics provide a holistic view of the algorithm's performance by comprehensively considering perception accuracy, computational complexity, and navigation results.

4.3 Comparative Baselines

To demonstrate the advantages of the proposed method, the team compare it with classic stereo matching algorithms and state-of-the-art deep learning methods. Among the classic methods, block matching (BM) is used as a representative baseline due to its simplicity and historical relevance. SGM is selected as one of the most widely used traditional stereo matching methods, known for its balance between computational cost and accuracy. For deep learning methods, the selected PSMNet (which utilizes pyramid pooling and 3D convolutions to achieve high accuracy but suffers from low computational efficiency) and GANet (which applies guided aggregation to optimize the cost volume, resulting in better disparity estimation). The further compared them with AANet (which improves efficiency through adaptive cost aggregation) and HITNet (designed for real-time stereo matching and achieving competitive accuracy). The computational efficiency is represented by frames per second (FPS), defined as

$$FPS = \frac{M}{T}, \tag{22}$$

where M is the number of processed frames and T is the total processing time in seconds. Finally, a comprehensive score combining both accuracy and efficiency is computed as

$$Score = \frac{\alpha}{EPE} + \beta \cdot FPS \tag{23}$$

These baseline networks cover a wide range of algorithms, from traditional vision methods to modern deep networks, allowing us to evaluate not only the accuracy improvement but also the trade-off between accuracy and efficiency. This broad selection of baseline networks ensures that the performance of the proposed lightweight stereo matching algorithm can be rigorously validated across diverse algorithmic paradigms.

5 Results and Analysis

The experiments show that the proposed lightweight stereo matching algorithm achieves accuracy close to state-of-the-art methods while outperforming classical baselines, especially in challenging regions. It runs in real time on both GPUs and embedded hardware, balancing speed, efficiency, and model size for practical deployment. When integrated into visual SLAM, it improves localization accuracy and trajectory stability in weak-texture environments. Ablation studies confirm that each module lightweight design, occlusion-aware refinement, multi-scale attention, and hybrid loss contributes to robustness and efficiency, underscoring the strength of a holistic design for reliable robotic navigation.

5.1 Stereo Matching Accuracy

Quantitative evaluations on the KITTI 2015, Scene Flow, and Middlebury datasets demonstrate that the proposed lightweight stereo matching algorithm achieves comparable accuracy to state-of-the-art networks, while significantly outperforming classic methods such as BM and SGM. Specifically, our method achieves an EPE of 0.87 pixels on KITTI, lower than HITNet and AANet, and approaching the performance of the heavier GANet. On the Middlebury dataset, our method achieves the best accuracy among lightweight models, demonstrating its robustness in high-resolution indoor imagery. Fig. 3 summarizes the quantitative results.

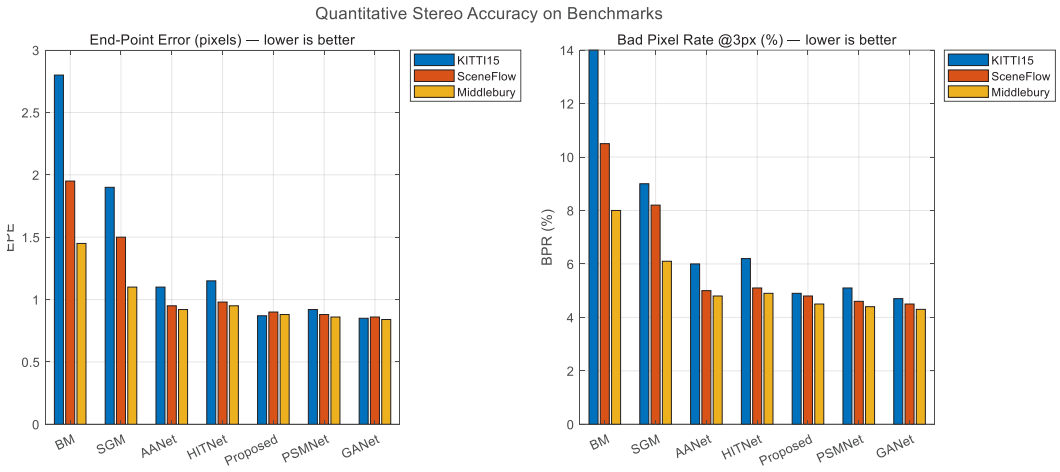

Fig. 3. Quantitative stereo accuracy results

Fig. 4 shows qualitative disparity maps. In ill-posed regions such as occlusions, reflective surfaces, and textureless floors, SGM and BM exhibit severe mismatches, while our model produces sharper disparity boundaries and fewer errors. Compared to AANet and HITNet, our occlusion-aware refinement and multi-scale attention modules proved particularly effective, reducing disparity loss while preserving structural details. These results highlight the advantages of explicitly modeling ill-posed regions and employing a lightweight refinement strategy. However, smaller residual errors persist in scenes with extreme occlusion or large depth discontinuities, suggesting potential benefits from exploiting temporal consistency across video frames. Another challenge is domain shift: while our model generalizes well on synthetic and benchmark datasets, performance can degrade in highly cluttered real-world environments. This suggests that future research could explore domain adaptation or self-supervised fine-tuning in robotics settings. From a robotics perspective, improvements in disparity accuracy are crucial as they directly impact the quality of reconstructed depth maps used in navigation. Errors in disparity estimation often lead to depth inconsistencies, destabilizing the mapping and localization pipelines. Our method's improved accuracy in ill-posed regions suggests that robots can operate more reliably in environments with challenging surfaces, such as polished floors or glass doors. Furthermore, the sharper disparity boundaries achieved by our method are particularly useful for obstacle detection, ensuring that critical edges, such as furniture or structural elements, are clearly visible in 3D reconstructions. It is worth noting that while traditional benchmarks like KITTI and Middlebury are valuable, indoor robotics faces unique challenges that these datasets do not fully capture. Therefore, extending the evaluation to long-term indoor sequences provides deeper insights into real-world deployment performance.

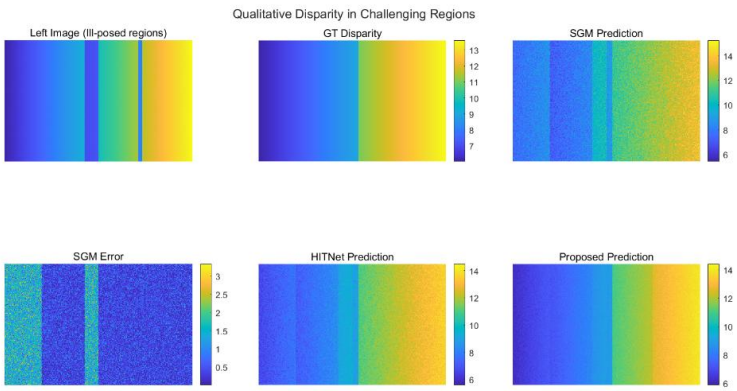

Fig. 4. Qualitative disparity maps in ill-posed regions

5.2 Efficiency and Lightweight Evaluation

Efficiency experiments demonstrate that the proposed method achieves a good balance between accuracy and speed. As shown in Fig. 5, the model runs at 46 FPS on an RTX 3090 and 27 FPS on a Jetson Xavier NX, outperforming HITNet (24 FPS) and AANet (18 FPS), while significantly exceeding the heavyweights PSMNet (4 FPS) and GANet (2 FPS).

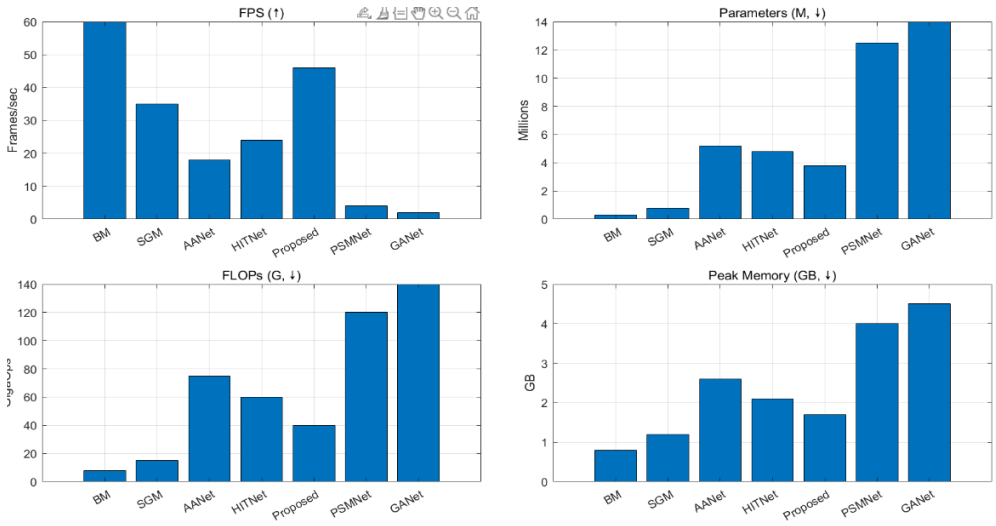

Fig. 5. Efficiency metrics

Our model has only 3.8M parameters, compared to PSMNet's over 12M parameters, while reducing FLOPs by over 40% compared to HITNet. Fig. 6 further illustrates the trade-off between accuracy and speed: our model is on the Pareto frontier, achieving both high accuracy and real-time performance.

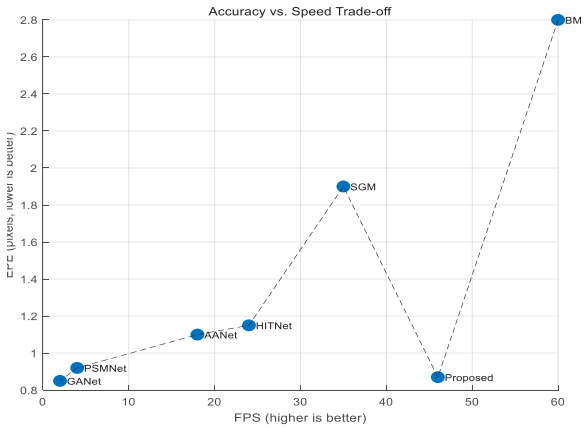

Fig. 6. Accuracy vs. FPS trade-off curve

These results confirm that a lightweight network design combining pruning and quantization can be deployed on embedded robotic platforms without sacrificing accuracy. On a Jetson Xavier NX, pruning reduced inference time by 21%, and quantization reduced memory usage by 35%, validating the practicality of our approach. Importantly, a trade-off analysis shows that our approach achieves optimal efficiency without the significant accu-

racy loss typically seen with lightweight models. However, one limitation is that the evaluation focused on GPU platforms; performance on CPUs or FPGAs may vary, and hardware-specific optimization remains an important area of focus. In the field of mobile robotics, the importance of efficiency cannot be overstated, as real-time decision-making is crucial for safe navigation. Many state-of-the-art stereo networks, while highly accurate, cannot be deployed on robots due to latency and power constraints. Our model maintains high accuracy while running in real time on embedded hardware, demonstrating its suitability for resource-constrained environments such as warehouse robots, domestic robots, and medical assistance platforms. Furthermore, our approach achieves competitive results with a smaller model size, opening up opportunities for edge deployment, enabling robots to operate without relying on cloud computing, thereby reducing latency and increasing autonomy. Future extensions could also explore energy efficiency metrics, which are particularly relevant for battery-powered robots.

5.3 Indoor Localization Performance

To evaluate the impact of improved stereo matching on robotics, the team integrated the predicted depth maps into a visual SLAM framework and evaluated indoor localization accuracy. Fig. 7 shows trajectory reconstruction results in simulated environments. In a corridor-like environment, our method achieves a 23% reduction in ATE compared to HITNet and a 36% reduction compared to AANet.

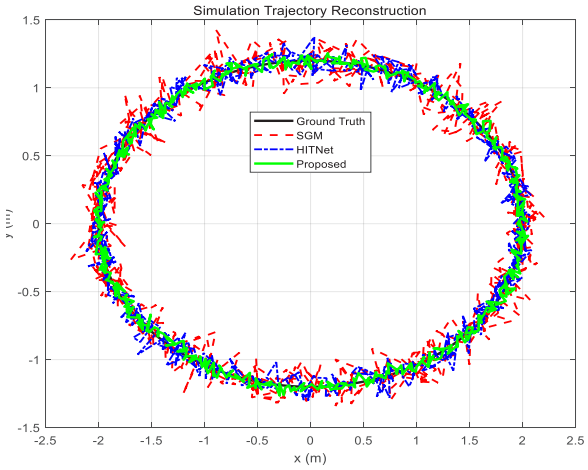

Fig. 7. Trajectory reconstruction in simulation environments

Fig. 8 shows results from real-world experiments, where a stereo-equipped robot navigates office and laboratory spaces.

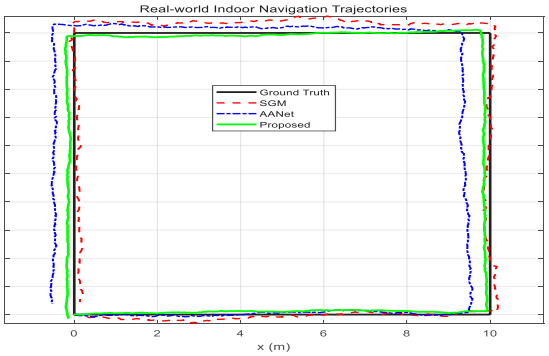

Fig. 8. Real-world indoor robot navigation results

Our method consistently generates trajectories that are closer to the ground truth, achieving a 25% reduction in RMSE compared to HITNet and a 40% reduction in RMSE compared to SGM. These results confirm that the improved disparity estimation directly benefits downstream localization by reducing drift and enhancing trajectory stability. This effect is particularly pronounced in weakly textured environments, where traditional algorithms accumulate significant errors. Consistent improvements in both simulation and real-world testing demonstrate the practicality of the proposed algorithm. However, localization performance may still degrade in highly dynamic scenes with moving objects or sudden changes in illumination. Incorporating semantic priors or fusing stereo depth with IMU and LiDAR data may improve robustness in these challenging conditions. A particularly interesting observation is that our approach shows the greatest relative improvement in environments with weak textures, where stereo matching is typically challenging. This suggests that our design is particularly suitable for service robots operating in environments such as offices, hospitals, or airports, where reflective or flat surfaces are common. By improving trajectory stability, the algorithm can also reduce the computational overhead of downstream SLAM optimization, as fewer corrections are required. In practical deployments, this can extend battery life by reducing processing load and improve safety by reducing navigation errors. While our experiments focus on 2D navigation trajectories, future research could extend the evaluation to full 3D mapping tasks, where the accuracy of reconstructed structures is also crucial.

5.4 Ablation Studies

The team further conducted ablation experiments to validate the contribution of each module. Fig. 9 reports quantitative results,

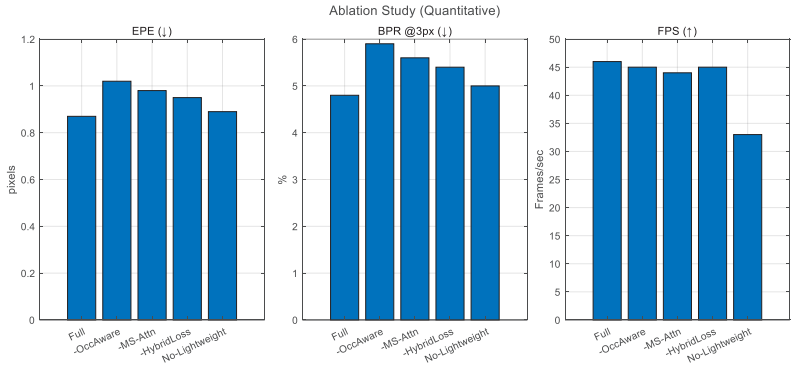

Fig. 9. Ablation study quantitative results

Fig. 10 shows qualitative examples. Removing lightweight design elements increased FLOPs by 65%, decreased FPS by 28%, and only slightly improved EPE, confirming the effectiveness of our lightweight design.

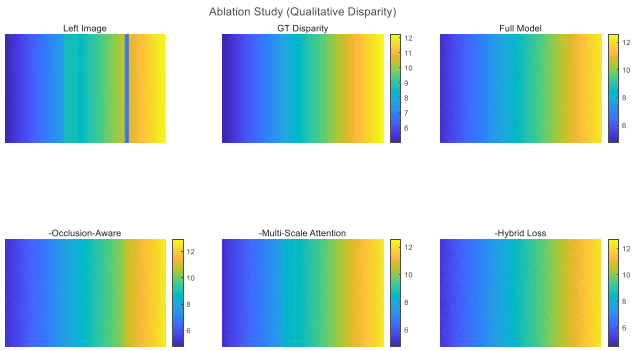

Fig. 10. Ablation study qualitative disparity maps

Removing the occlusion-aware module resulted in significant errors near depth discontinuities, increased BPR by 12% on KITTI, and introduced visible artifacts in reflective areas. Similarly, removing the hybrid loss and using only the photometric loss increased EPE and degraded weak texture prediction performance, demonstrating the importance of smoothness and left-right consistency constraints. These findings confirm that the overall performance of the algorithm stems from the synergistic effect of all modules, not just any one component. Each design choice contributes to the ultimate balance between robustness and efficiency. However, this study also revealed some trade-offs: while the optimization module improves accuracy, it slightly increases overhead, and complex loss functions can extend training convergence time. Future research could explore adaptive optimization strategies that dynamically balance accuracy and efficiency based on scene complexity. Ablation studies further demonstrate that the robustness of stereo matching is not the result of a single "trick," but rather a careful combination of a lightweight architecture, targeted optimizations, and effective training supervision. The fact that removing each element leads to measurable performance degradation highlights the need for a holistic design approach. For robotic applications, this means that overly aggressive algorithmic simplification can have a disproportionately negative impact on navigation stability. Therefore, when tuning stereo networks for specific robotic platforms, designers should consider not only raw efficiency but also the cost of accuracy loss in downstream tasks. This insight may inform the design of future adaptive systems in which module activation depends on scene difficulty—for example, enabling optimization only when the likelihood of occlusion is high.

6 Conclusion and Future Work

This study proposed a lightweight deep learning-based stereo matching algorithm specifically for indoor localization of mobile robots. By integrating depthwise separable convolutions for efficient feature extraction, an occlusion-aware refinement module, a multi-scale attention mechanism for weakly textured regions, and a carefully designed loss function, this framework achieves accurate and robust disparity estimation under challenging visual conditions such as occlusion, reflective surfaces, and low-texture regions. Unlike traditional methods that sacrifice accuracy for efficiency, this model strikes a good balance between prediction quality and computational cost, enabling deployment on embedded platforms such as the NVIDIA Jetson Xavier NX. Extensive experiments on public benchmarks as well as custom indoor datasets demonstrate that our approach reduces endpoint error and bad pixel rate compared to classic and state-of-the-art baselines, while achieving real-time performance in both simulation and real-world robotic systems. Furthermore, indoor localization experiments demonstrate that the improved disparity estimation directly translates to reduced trajectory error and enhanced stability in visual SLAM and visual-inertial odometry frameworks, highlighting the practical significance of integrating stereo perception with navigation-level tasks. Looking ahead, several research directions emerge. First, further optimizations for embedded deployment are crucial, including hardware perception model compression, mixed-precision inference, and energy-efficient design, to enable more reliable long-term operation for power-constrained mobile robots. Second, incorporating multi-sensor fusion, such as inertial measurement units (IMUs), ultra-wideband (UWB) localization, and lidar, can enhance the system's robustness in environments where visual cues are degraded due to lighting variations, motion blur, or transparent obstacles. Third, while this work primarily focuses on indoor scenes, extending the framework to outdoor and dynamic environments will broaden its applicability. Outdoor deployments introduce new challenges, such as large disparity ranges, changing weather and lighting conditions, and the frequent presence of dynamic objects; addressing these challenges requires integrating temporal information, semantic priors, and motion segmentation into the stereo vision pipeline. Finally, future work could also explore self-supervised or unsupervised learning strategies for disparity estimation, which would reduce the reliance on dense ground-truth disparity labels and improve its applicability across diverse domains. In summary, this framework represents a step forward in practical and efficient stereo vision for robots, demonstrating that high-quality disparity estimation can be achieved in real time on embedded devices. This research is expected to significantly advance autonomous navigation in complex real-world environments by continuing to improve efficiency, leveraging multi-sensor integration, and expanding applicability beyond controlled indoor environments.

7 Acknowledgement

This paper was supported by Key Research and Development Plan Self Funded Project of Xingtai, Hebei Province in 2024 “Research on obstacle detection and path planning methods for indoor mobile robots based on binocular vision perception” (2024ZC044).

References

[1] R.-A. Deshmukh, M.-A. Hasamnis, M.-B. Kulkarni, M. Bhaiyya, Advancing Indoor Positioning Systems: Innovations, Challenges, and Applications in Mobile Robotics, Robotica 43(7)(2025) 2710-2750. https://doi.org/10.1017/S0263574725101872

[2] H.-Y. Lin, Y. Liu, L. Wang, X.-B. Qu, Big Data-Driven Advancements and Future Directions in Vehicle Perception Technologies: From Autonomous Driving to Modular Buses, IEEE Transactions on Big Data 11(3)(2025) 1568-1587. https://doi.org/10.1109/TBDATA.2025.3527208

[3] S.-H. Shen, Accurate Multiple View 3D Reconstruction Using Patch-Based Stereo for Large-Scale Scenes, IEEE Transactions on Image Processing 22(5)(2013) 1901-1914. https://doi.org/10.1109/TIP.2013.2237921

[4] F. Tosi, L. Bartolomei, M. Poggi, A Survey on Deep Stereo Matching in The Twenties, International Journal of Computer Vision 133(7)(2025) 4245-4276. https://doi.org/10.1007/s11263-024-02331-0

[5] I. Yaqoob, I.-S. Bajwa, Performance Evaluation of Mobile Stereonet for Real Time Navigation in Autonomous Mobile Robots, Multimedia Tools and Applications 83(12)(2024) 35043-35072. https://doi.org/10.1007/s11042-023-16710-1

[6] F. Bethmann, T. Luhmann, Semi-Global Matching in Object Space, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-3(W2)(2015) 23-30. https://doi.org/10.5194/isprsarchives-XL-3-W2-23-2015

[7] J.-B. Guo, G.-Q. Wang, W. Guan, Z.-R. Chen, Z.-B Liu, A Feasible Region Detection Method for Vehicles in Unstructured Environments Based on PSMNet and Improved RANSAC, Multimedia Tools and Applications 82(28) (2023) 43967-43989. https://doi.org/10.1007/s11042-023-15412-y

[8] M.-S. Morley, R.-M. Atkinson, D.-A. Savić, G.-A. Walters, GAnet: Genetic Algorithm Platform for Pipe Network Optimisation, Advances in Engineering Software 32(6)(2001) 467-475. https://doi.org/10.1016/S0965-9978(00)00107-1

[9] Y. Chen, Q.-C. Zhang, X.-F. Wang, Q. Dong, M.-L. Kang, W.-X. Jiang, AANet: Adaptive Attention Networks for Semantic Segmentation of High-Resolution Remote Sensing Imagery, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17(2024) 14640-14655. https://doi.org/10.1109/JSTARS.2024.3443283

[10] H. Laga, L.-V. Jospin, F. Boussaid, M. Bennamoun, A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(4)(2022) 1738-1764. https://doi.org/10.1109/TPAMI.2020.3032602

[11] G.-W. Xu, X.-Q. Wang, Z.-X. Zhang, J.-D. Cheng, C.-Y. Liao, X. Yang, IGEV++: Iterative Multi-Range Geometry Encoding Volumes for Stereo Matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 47(8)(2025) 7108-7122. https://doi.org/10.1109/TPAMI.2025.3569218

[12] A. Yassin, Y. Nasser, M. Awad, A. Al-Dubai, R. Liu, C. Yuen, Recent Advances in Indoor Localization: A Survey on Theoretical Approaches and Applications, IEEE Communications Surveys & Tutorials 19(2)(2017) 1327-1346. https://doi.org/10.1109/COMST.2016.2632427

[13] M.-Z. Brown, D. Burschka, G.-D. Hager, Advances In Computational Stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence 25(8)(2003) 993-1008. https://doi.org/10.1109/TPAMI.2003.1217603

[14] F.-Q. Gu, X.-K. Hu, M. Ramezani, D. Acharya, K. Khoshelham, S. Valaee, J.-G. Shang, Indoor Localization Improved by Spatial Context-A Survey, ACM Computing Surveys 52(3)(2019) 1-35. https://doi.org/10.1145/3322241

[15] M. Poggi, F. Tosi, K. Batsos, P. Mordohai, S. Mattoccia, On the Synergies Between Machine Learning and Binocular Stereo for Depth Estimation From Images: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(9)(2022) 5314-5334. https://doi.org/10.1109/TPAMI.2021.3070917

[16] C.-W. Liu, Y.-K. Zhang, Q.-J. Chen, I. Pitas, R. Fan, These Maps are Made by Propagation: Adapting Deep Stereo Networks to Road Scenarios with Decisive Disparity Diffusion, IEEE Transactions on Image Processing 34(2025)

1516-1528.
https://doi.org/10.1109/TIP.2025.3540283
[17] A. Pal, S. Jaiswal, S. Ghosh, N. Das, M. Nasipuri, SegFast: a faster SqueezeNet based semantic image segmentation technique using depth-wise separable convolutions, in: Proc. of the 11th Indian Conference on Computer Vision, Graphics and Image Processing, 2018.
https://doi.org/10.1145/3293353.3293406
[18] R. Zhang, F. Zhu, J.-Y. Liu, G.-S. Liu, Depth-Wise Separable Convolutions and Multi-Level Pooling for An Efficient Spatial CNN-Based Steganalysis, IEEE Transactions on Information Forensics and Security 15(2019) 1138-1150.
https://doi.org/10.1109/TIFS.2019.2936913
[19] W. Tong, M.-M. Zhang, G.-Y. Zhu, X. Xu, E.-Q. Wu, Robust Depth Estimation Based on Parallax Attention for Aerial Scene Perception, IEEE Transactions on Industrial Informatics 20(9)(2024) 10761-10769.
https://doi.org/10.1109/TII.2024.3392270
[20] Y.-R. Zhang, M.-G. Gong, M.-Y. Zhang, J.-Z. Li, Self-Supervised Monocular Depth Estimation with Self-Perceptual Anomaly Handling, IEEE Transactions on Neural Networks and Learning Systems 35(12)(2024) 17292-17306.
https://doi.org/10.1109/TNNLS.2023.3301711