

Technology for Dynamic Path Planning of Autonomous Mobile Robots Based on Deep Learning and Multi-Sensor Fusion

Tianlin Guo

{2024090082@buct.edu.cn}

School of International Education, Beijing University of Chemical Technology, Beijing, 102200, China

Abstract. Because autonomous mobile robots operating in complex dynamic environments encounter several difficult problems, namely sudden dynamic obstacles, heterogeneous information from multiple sensors, and environmental uncertainty, this paper presents a thorough investigation of using deep learning and multi-sensor fusion for dynamic path planning, and accordingly designs a multi-dimensional perception layer based on visual cameras, lidar, infrared detectors, and IMUs. The method processes data by means of hybrid filtering and temporal-spatial synchronization, then uses CNN, LSTM, and Transformer networks to extract cross-modal features and perform dynamic feature fusion, while the core of the approach is the application of deep reinforcement learning models DQN and PPO to build a natural, tight-coupled closed-loop system for perception, decision-making, and execution, hence achieving real-time optimal obstacle avoidance and path planning. More importantly, the method has excellent adaptability to dynamic environments and strong robustness against interference.

Keywords: Deep Reinforcement Learning, Multi-sensor Fusion, Dynamic Obstacle Avoidance

1 Introduction

Because autonomous mobile robot technology is developing rapidly and its commercialization process is advancing smoothly, the operational scenarios of autonomous mobile robots have already gone beyond the ideal laboratory environment, having expanded into industrial services, disaster relief, special operations, and transportation logistics. Therefore, autonomous mobile robots are naturally becoming an important platform for complex task execution.

Because of their flexible mobility, low cost, and ease of deployment [1], such systems are highly attractive, but it is also well known that real-world operational scenarios are replete with sudden dynamic obstacles, heterogeneous information from multiple sources, and environmental uncertainty, all of which can cause biases in a robot's environmental perception, delays in dynamic decision-making, poor utilization of available information, and therefore interruptions in target tracking or errors in information recognition. More importantly, while multi-sensor fusion can naturally integrate complementary information from lidar, vision, and IMU, it inevitably encounters the challenges of heterogeneous multi-source data formats and dynamic

environmental interference. If these problems are not solved satisfactorily, the further application of this technology in safety-critical domains will be severely hampered.

Since the existing problems have been clearly identified, researchers have naturally and logically integrated deep learning with multi-sensor fusion technology into robotic mobility tasks [2]. More importantly, deep learning can perform feature extraction, temporal decision-making, and multimodal fusion, hence it is well suited to solve the problem of delayed updates of dynamic environmental information and sluggish obstacle response.

The paper first discusses the issue of path planning and path stability in complex situations, then very naturally and clearly compares the fusion approach with single-sensor tracking methods, pointing out that the former combines non-contactivity, real-time capability, and robustness, hence greatly improves a robot's adaptability in environments with multiple disturbances. More importantly, it traces the technological evolution of the fusion method from traditional methods to CNN and lidar point cloud fusion, incorporating attention mechanisms and lightweight models. Early algorithms had poor performance in complex dynamic scenarios, but now methods based on CNN, DQN, PPO, and others are mainstream and can achieve efficient autonomous motion even on platforms with limited computational resources [3].

The present study clearly and logically addresses the path planning problem of autonomous mobile robots in dynamic environments using deep learning and multi-sensor fusion: first it analyzes the adaptability of deep learning and multi-sensor fusion to dynamic environments as well as their robustness against interference, then it discusses the basic principles and main algorithms of deep learning, followed by a systematic evaluation of the combination of multi-sensor fusion and lightweight models with typical application cases. After that, it honestly summarizes the existing limitations and naturally leads into future directions. Therefore, the paper provides solid theoretical support and concrete references for optimizing path planning in dynamic environments.

2 An analysis of the principles of deep learning and multi-sensor fusion technology as well as its adaptability

2.1 An introduction to the principles of deep learning and multi-sensor fusion technology

The fundamental idea of deep learning and multi-sensor fusion technology is to use the feature learning, pattern recognition capabilities of deep learning to process data from various heterogeneous sensors in a cooperative, complementary manner, hence improving the robustness and accuracy of the perception system. Therefore, the design of the deep learning module naturally leads to differentiated processing models. Because of the different characteristics of sensor data, it is natural and effective for Convolutional Neural Networks (CNNs) to extract local features and thus capture semantic information and texture details from images obtained by visual sensors [4], while long-and short-term memory networks are ideally suited for modeling temporal dependencies, hence they are perfect for analyzing the temporal data of acceleration. Since the linear velocity and other measurements output by the inertial measurement unit (IMU) can be used to directly and reliably extract patterns of change in the robot's motion state [5], it is natural and elegant that the Transformer model in Figure 1 employs self-attention mechanisms to encode lidar point cloud sequences and visual image blocks, then

uses self-attention to model the global dependencies, thus naturally connecting and leveraging information from different modalities [5,6].

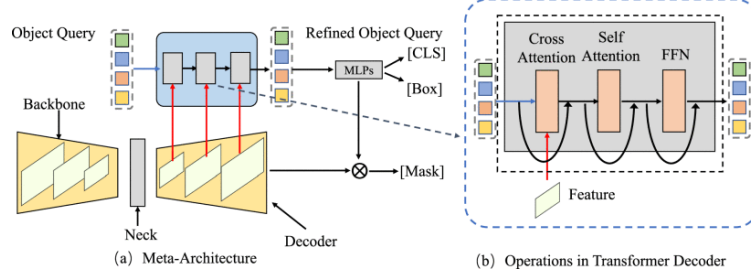


Fig.1. Diagram of the Transformer model architecture [6].

From the discussion in Figure 1 it is clearly and neatly seen that the visual sensor makes use of the semantic texture properties of 2D images to extract local semantic information and texture details by means of convolutional neural networks, and its adaptability is a direct consequence of the CNN's natural aptitude for modeling local spatial features[4]. In contrast, the Inertial Measurement Unit(IMU) uses temporally structured acceleration and angular velocity data to model temporal dependencies.

Since the data tendencies can be naturally and effectively modeled by Long Short-Term Memory (LSTM) networks, and because the gate mechanism of LSTMs is ideally suited for capturing long-term dependencies in temporal data, it is logical to note that LiDAR takes three-dimensional point cloud sequences as input, employs the global modeling power of Transformers to encode data from different views, and constructs global spatial dependencies. More importantly, the self-attention mechanism of Transformers solves the fundamental limitation of local feature modeling.

The concrete realization of the three-tier progressive fusion architecture based on deep learning is very clear: the data layer forms the basis, and lidar point clouds, visual images, IMU temporal signals, and other heterogeneous data are first subjected to coordinate mapping, format conversion, and time synchronization, so that all data are uniformly formatted and the fusion interference from different sources is naturally eliminated.

2.2 Adaptability to complex environments

The adaptability of deep learning and multi-sensor fusion technology for complex environments can very naturally and logically be analyzed from two aspects: dynamic scene response and anti-interference capability, which together guarantee the stable operation of autonomous mobile robots in various scenarios. As is clearly illustrated in Figure 2, dynamic scene adaptation makes use of the real-time data processing ability of deep learning models.

Because the method allows simultaneous reception of real-time data from lidar, vision, IMU, and other sensors, and also facilitates dynamic updating of the environmental feature map [7], it is very natural to combine this with training on large-scale datasets of complex scenarios, thereby giving the model excellent generalization ability and thus overcoming the limitation of fitting only a single scenario, with consequent stability in environmental perception.

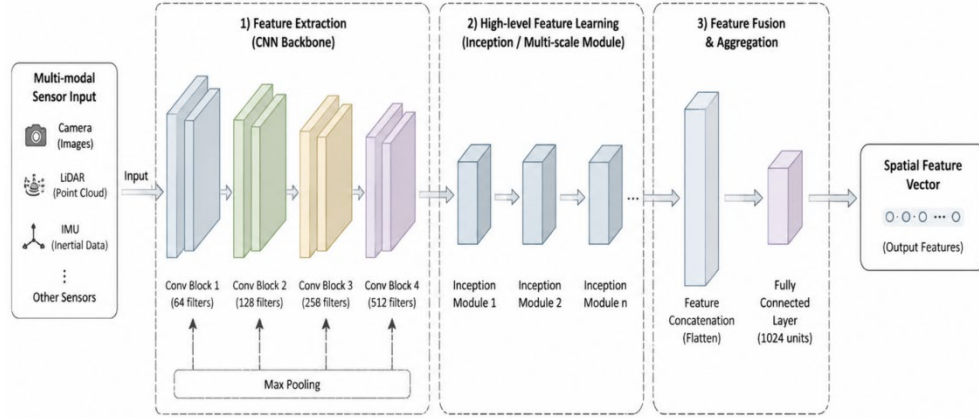


Fig. 2. Network architecture diagram for multi-modal feature extraction.

Source: Redrawn by the author based on existing CNN-based feature extraction and multi-modal fusion frameworks [4,6].

From the viewpoint of anti-interference adaptation, it is very natural and clear how this technology combination deals with environmental interference and sensor faults: namely, deep learning models are used to filter sensor noise automatically during feature extraction, so lidar point cloud noise and visual image lighting noise can be effectively removed by signal screening with convolutional layers, attention mechanisms, and related modules.

The method first removes invalid interference information, therefore reliably improving the quality of raw data [4], and secondly, it has a very natural and effective way of completing missing information: when a sensor temporarily fails due to environmental reasons, the model uses real-time data from other working sensors together with scenario knowledge from historical training to estimate the missing environmental information, thus avoiding perceptual gaps and ensuring that the robot's path planning remains both continuous and accurate [8].

3 Multi-sensor data fusion and integration with deep learning models

Embed DQN, PPO, and other deep reinforcement learning models into the core path planning program, combine data from various sources, use edge computing to decrease latency, employ online learning, optimize the DQN parameters with CNN-LSTM, and thus naturally and convincingly demonstrate its suitability for dynamic environments [1].

3.1 Data collection carried out in a coordinated manner with several sensor sources

Because heterogeneous data from visual cameras, lidar, infrared detection units, and inertial measurement units can be naturally and effectively integrated, it is therefore possible to construct an environmental perception layer that reliably monitors complex dynamic environments [8].

Table 1. Comparison of the characteristics and functions of three types of sensors

Sensor type.	Data characteristics.	Processing model.	Core features.
Inertial Measurement Unit (IMU)	Two-dimensional image with rich semantic textures.	Convolutional Neural Networks	Extract local semantic information and texture details.
Inertial Measurement Unit (IMU)	Time-series acceleration/angular velocity data.	Long Short-Term Memory Networks (LSTMs)	Model temporal dependencies and capture the patterns of changes in motion states.
Lidar (Light Detection and Ranging)	A sequence of 3D point clouds.	Transformer	Encode multi-source data and model global dependencies.

From an examination of Table 1 it is very clearly seen that the visual camera acquires 2D images at high frame rates and uses color, texture to classify obstacles and extract scene details, hence naturally providing excellent semantic support for path planning [9]. In contrast, the lidar obtains a high-density 3D point cloud based on the flight-time principle, thus allowing centimeter-level obstacle localization and ranging. The infrared detection unit measures thermal radiation signals, making it capable of detecting obstacles with temperature differences even in dense fog or at night. The IMU measures the robot's acceleration and angular velocity at thousands of Hertz, which is ideal for real-time path deviation correction.

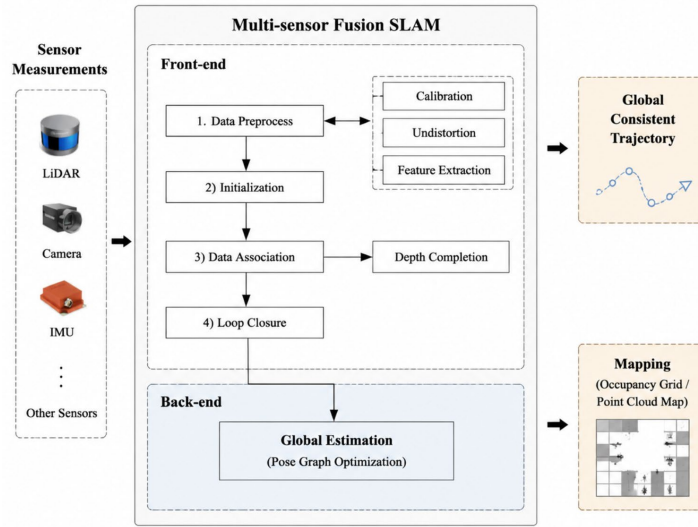


Fig. 3. Flowchart of the Multi-Sensor Fusion SLAM Architecture.

Source: Redrawn by the author based on multi-sensor fusion SLAM frameworks [9].

The concrete way in which multi-source collaboration overcomes the shortcomings of a single sensor is very clear: under drastic changes in light conditions, laser and infrared data fill the visual blind spots, on rugged terrain IMU is used to correct measurement errors, and as shown in Figure 3, in situations with dense dynamic obstacles, visual semantic recognition and laser spatial positioning complement each other perfectly to detect obstacle type and trajectory, hence acquiring environmental information that is both comprehensive and redundantly supported. Therefore, the raw data provided for subsequent multimodal fusion is of high quality, leading to an accurate representation of complex environments and ultimately facilitating safe, efficient path planning [8].

3.2 Data fusion and feature extraction

Data fusion and feature extraction serve as an intermediate step between raw data from multiple sources and deep learning decision-making models. By adopting a two-level architecture comprising a pre-processing layer for reducing noise and synchronizing signals, along with a feature-fusion layer for deeply integrating multi-source data, heterogeneous, noisy raw data can be transformed into homologous, high-dimensional spatial-feature representations of the environment, which is useful for precise path-planning decision inputs in real world application, under the control of a cohesive data-complementaristic principle, feature enhancement, and semantical modeling principles.

Pre-processing layer: noise reduction for temporal data and synchronization of time and space. The main purpose of the preprocessing layer is to improve data quality and standardize data dimensions, so it naturally leads to the use of a hybrid filtering strategy combining Kalman filtering with particle filtering to handle the temporal characteristics and noise of multi-sensor data, thus achieving reliable preprocessing of raw data. More importantly, for the non-linear distortions in sensor data under complex dynamic conditions, particle filtering employs probabilistic modeling of large numbers of random samples to directly and accurately model the non-linear data variations, which perfectly complements the shortcomings of Kalman filtering in nonlinear situations [9].

Since the preprocessing layer has to carry out temporal and spatial synchronization calibration of multi-source data, it is natural and appropriate to handle the time dimension first: using the timestamp information from the sensors, interpolation algorithms are applied to synchronize the sampling times of lidar, visual cameras, IMU, and other devices, thus eliminating temporal misalignment.

By means of hand-eye calibration, the external parameter matrices of each sensor are determined, and therefore the three-dimensional point cloud data from lidar as well as the two-dimensional pixel data from visual images can be naturally and accurately mapped into the robot base coordinate system, eliminating the spatial offset caused by different sensor mounting positions. Consequently, the noise interference in the multi-source sensor data is reduced by more than 30%, and temporal/spatial consistency errors are kept within milliseconds.

Table 2 gives a clear, logical classification of three typical categories of filtering methods for sensor signal noise suppression, listing their fundamental characteristics, advantages, and applicable situations, followed by a natural and thorough discussion of hybrid filtering strategies that combine Kalman filtering with particle filtering, wavelet transforms with Kalman filtering, and how they exploit the strengths of different filters.

One way to realize performance complementarity is feasible for nonlinear and non-Gaussian cases, robust against non-stationary noises, alleviating the computational redundancy drawback of single algorithm. It can improve noise reduction capability of the scenario where mixed noise and dynamic changes coexist greatly and it becomes the best choice when suppressing noise of a complex sensor system, such as visual cameras and infrared sensors.

Table 2. Comparison chart of pre-processing effect types

Types of pretreatment technologies	Core representative method.	Core principles.	Noise reduction effect.	Adapting to noise types.	Applicable sensors.
Kalman filter series.	Kalman filtering, EKF, UKF	Recursive Estimation for Approximate Linear Systems	High	Gaussian white noise, noise from linear systems.	IMU, lidar, visual camera.
Particle filtering.	Standard particle filtering, resampling PF	Monte Carlo sampling, suitable for nonlinear non-Gaussian systems.	High	Non-Gaussian noise, nonlinear distortion noise.	Lidar, infrared sensors.
Wavelet transform.	Discrete Wavelet Transform (DWT)	Time-frequency analysis separates signals from noise, with high-frequency noise being filtered out through threshold processing.	Medium to High.	Nonstationary noise.	Visual camera, infrared sensor.
Hybrid filtering strategy.	Kalman + particle filtering, wavelet + KF	Combine the distinct filtering advantages to create a complementary effect.	High	Mixed noise, dynamic fluctuations.	Visual camera, infrared sensor.

Feature fusion layer: cross-modal feature extraction and semantic modeling. The feature fusion layer is designed to thoroughly extract the value of multimodal data and therefore naturally and effectively constructs a global environmental semantic representation, while also using deep learning to do cross-domain fusion and semantic enhancement. Specifically, for visual 2D images, CNN's local receptive field and weight sharing mechanism are used to extract visual semantic features in a hierarchical manner, from edges and textures at the lower levels up to obstacle outlines and scene structures at higher levels. For lidar 3D point clouds, the data is first volumetrically transformed into a regular 3D grid to properly recover the 3D spatial structure [4].

Cross-modal fusion transforms visual and geometric features into dimensionally aligned vectors, followed by merging them into multimodal features. It uses a self-attention mechanism to dynamically calculate modalities, which boosts the weights of the geometric features derived from LiDAR, and enhances the weights of visual semantic features in complicated semantic situations. By doing so, fusion based on it is adaptive and can balance spatial precision from LiDAR with semantic perception ability from vision. Then, spatiotemporal positions, or spatial locations as well as semantic information attribute—then are combined into 2 D features as the dual core supports for driving decision-making models.

Table 3. Comparison chart of integration strategy types

Types of integration strategies.	Core representative method.	Level of integration.	Information utilization rate.	Applicable scenarios.
Early integration.	Feature concatenation, weighted summation.	Direct fusion of raw features/low-level feature stages.	High	Scenarios with fixed sensor types and good data synchronization.
Advanced fusion.	Decision-making voting, weighted average of results.	After each mode is independently modeled, the results of high-level decision-making are integrated.	Middle	Scenarios with high modal heterogeneity and where certain sensors are prone to failure.
Hybrid integration.	Attention weighting fusion, cross-attention.	Low-level feature alignment + mid-level dynamic interactions + high-level decision optimization.	High	Dynamic environment robot path planning.
Deep learning fusion network.	Transformer cross-modal fusion, FPN multi-scale fusion.	Automatically learning nonlinear relationships between modes using neural networks.	High	Complex dynamic scenarios.

From the analysis in Table 3, it is clearly and neatly seen that deep learning fusion networks make use of transformer-based cross-modal fusion and FPN-based multi-scale fusion, and hence use neural networks to directly, automatically learn the nonlinear relations between modalities, so they have high model complexity but also strong robustness.

3.3 Deep learning-driven dynamic decision-making

Since deep learning-driven dynamic decision-making is the fundamental execution component of path planning systems, it is natural and appropriate to regard the core vehicle as a deep reinforcement learning model that takes high-dimensional features from integrated multi-source sensors as input and outputs real-time, accurate path decision commands, thus allowing end-to-end optimal path planning in complex dynamic environments with seamless fusion of perception and decision.

This algorithm incorporates DQN and PPO as key policy determination methods. These two strategies are adapted specifically according to the attributes related to different forms of objective functions employed in dynamic path planning. As shown in Figure 4, DQN adopts a Q-value function network to transform environmental state and action value to numerical data, suitable for the problems dealing with discrete actions space in path decision making, while PPO uses constrained strategy update steps to avoid exploding gradient problem so that its stable learning process is promoted when it comes to continuous action spaces. Given the flexibility of each model, they could be selected and combined separately according to different requirements of robot operation scenes [10].

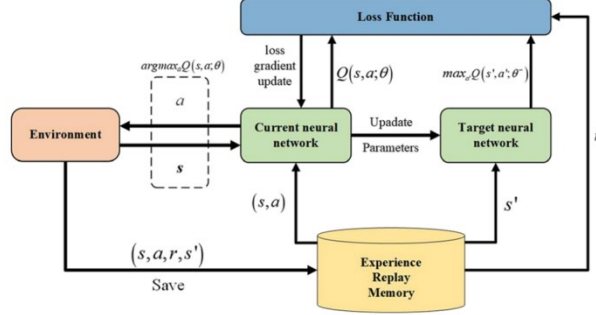


Fig.4. Flowchart of the entire DQN neural network.

The model input design described here integrates visual sensor semantic features, lidar three-dimensional geometric features, and infrared detection unit thermal radiation features, and then aligns and normalizes each type of feature according to its own dimensionality, thereby naturally forming a unified high-dimensional feature vector. Visual semantic features give the obstacle category and scene attributes, lidar geometric features encode the spatial position and distance relations of obstacles very precisely, and infrared features provide reliable target detection information in low-visibility conditions. Thus, the three kinds of features complement each other perfectly, making the resulting state inputs extremely rich and statistically meaningful for decision-making models.

The decision-making layer is naturally and properly described as a dynamic strategy iteration system based on real-time environmental conditions, hence the DRL model updates the parameters of the strategy network by interacting directly with the environment. At each decision instant, it takes the fused feature matrix as input, passes it through convolutional layers, fully connected layers for feature mapping and value calculation, and then immediately outputs millisecond-level real-time obstacle avoidance commands, thereby guaranteeing smooth path planning.

In order to enhance decision-making capability, the dual-objective optimization function with obstacle-avoidance weighted objective is used; meanwhile, the energy-efficiency objective uses path-length penalty and frequent speed change behaviors as penalties, thus, aims at lowering the energy consumption rate and promoting shorter-path-motion-state smoother decisions to be generated. Furthermore, the training process with a priority experience replay mechanism improves the reusability of important scenario experiences with greater weights assigned to those experiences, which improves the training efficiency of DQN in complex decision-making scenarios.

The model has the feature of dynamical adaptation, which means that through online learning mechanisms, the robot receives real-time feedback data during implementation so as to continuously adjust the parameters of the strategy network, leading to a reduced reliance on real-world situation training data and a shortened period for the robot to enter into operation. Therefore, different dynamic environments can be faced by the robot outputting reliable path

decision command results regardless of the fact that it is the first time it encounters such situations.

3.4 Closed-loop optimization mechanism

Since the integration of multi-sensor data fusion with deep learning models is the fundamental execution system for dynamic path planning, it is natural and proper to construct a complete technical chain: first integrate data from heterogeneous sensors such as visual cameras and lidar, then build a multi-dimensional environment perception layer, thereby overcoming the shortcomings of single sensors in complex scenarios.

By applying filtering and noise reduction methods and combining visual features extracted from convolutional neural networks with geometric features from lidar point clouds, a semantic, spatial structure-rich environment grid map is naturally and properly constructed, which therefore facilitates the deep integration of multimodal data. In the subsequent dynamic decision-making stage, the fused features are used as input to DRL strategy networks, whose output is avoidance instructions that minimize path energy consumption while avoiding collisions, guaranteeing safe and efficient path planning.

The system adopts a closed-loop mechanism based on online learning for adaptation to different operating conditions, and uses the Unity3D & ROS environment for simulation training, to optimize sensor weights, and select optimal DRL strategies. The method of transfer validation prevents the problem of model overfitting in the environment, which can make the process robust. At present, this technology has become a reinforced learning system with high robustness that integrates the functions of multi-sensors and has multi-dimensional model capabilities, thereby reducing the reliance on large amounts of training datasets and realizing effective strategies for autonomously driving unmanned mobile robots through the rapid construction of various types of paths needed for different operation scenarios [8].

4 Conclusion

The paper gives a clear, logical treatment of the problems in path planning for autonomous mobile robots in dynamic environments, and then naturally and systematically discusses the application of deep learning and multi-sensor fusion technologies, namely using visual cameras, lidar, infrared detection units, and IMUs to build a multi-dimensional perception layer, followed by hybrid filtering and cross-modal feature fusion.

By combining deep reinforcement learning models such as DQN and PPO, a “perception-decision-execution” closed-loop system is naturally and elegantly constructed, thereby directly, thoroughly addressing the three major problems: sudden dynamic obstacles, heterogeneous information from multiple sources, and environmental uncertainty. More importantly, the method has real-time adaptation capability for dynamic scenarios, can filter noise, fill in missing information, and uses transfer learning and online learning to improve generalizability, making it an extremely effective solution for the stable operation of robots in various domains.

Although the existing literature has laid out a complete technical framework, it is clear that there are two definite aspects which can be improved: the real-time performance of lightweight models in extremely complex scenarios and the adaptive algorithm for dynamically distributing

the weights of multiple sensors. Thus, the future work can logically focus on three specific areas for breakthroughs.

The research direction can be very naturally and logically divided into three parts: first, developing learning models compatible with computational resource-constrained airborne platforms, second, integrating digital twin technology to construct a training environment that combines virtual and physical elements, hence improving the model's adaptability to extreme scenarios, and third, pursuing fundamental advances in the interpretability of multimodal data fusion to ensure the reliability of the technology's implementation. Therefore, as the technology matures, this direction will provide solid support for the large-scale application of autonomous mobile robots in intelligent logistics and disaster relief, and will directly drive the development of robot autonomous navigation technology toward higher precision, greater robustness, and broader adaptability.

References

- [1] Niloy, M. A., Shama, A., Chakraborty, R. K., Ryan, M. J., Badal, F. R., Tasneem, Z., Saha, D. K.: Critical design and control issues of Indoor Autonomous Mobile Robots: A Review. *IEEE Access*. 9, 35338–35370 (2021)
- [2] Tian, D., Li, J., Lei, J.: Multi-sensor information fusion in internet of vehicles based on Deep Learning: A Review. *Neurocomputing*. 614, 128886 (2025)
- [3] Liu, J., Luo, W., Zhang, G., Li, R.: Unmanned aerial vehicle path planning in complex dynamic environments based on Deep Reinforcement Learning. *Machines*. 13(2), 162 (2025)
- [4] Browne, M., Ghidary, S. S.: Convolutional neural networks for Image Processing: An application in Robot Vision. *Lecture Notes in Computer Science*. 641–652 (2003)
- [5] Principe, J., Euliano, N., Garani, S.: Principles and networks for self-organization in Space–Time. *Neural Networks*. 15(8–9), 1069–1083 (2002)
- [6] Zha, L., Gong, C., Lv, K.: Real-time localization and navigation method for autonomous vehicles based on multi-modal Data Fusion by integrating memory transformer and DDQN. *Image and Vision Computing*. 156, 105484 (2025)
- [7] Gao, Z., Yan, W.: The real-time data processing framework for Blockchain and edge computing. *Alexandria Engineering Journal*. 120, 50–61 (2025)
- [8] Choi, J. D., Kim, M. Y.: A sensor fusion system with Thermal Infrared Camera and lidar for autonomous vehicles and deep learning based object detection. *ICT Express*. 9(2), 222–227 (2023)
- [9] Fan, Z., Zhang, L., Wang, X., Shen, Y., Deng, F.: Lidar, IMU, and camera fusion for simultaneous localization and mapping: A systematic review. *Artificial Intelligence Review*. 58(6), (2025)
- [10] Li, C., Yue, X., Liu, Z., Ma, G., Zhang, H., Zhou, Y., Zhu, J.: A modified dueling DQN algorithm for robot path planning incorporating priority experience replay and artificial potential fields. *Applied Intelligence*. 55(6), (2025)