Journal of Electronic Research and Application, 2025, Volume 9, Issue 5
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

The Design and Implementation of an Intelligent Guide Dog Robot Based on Multimodal Perception

Yanxuan Zhu*

Nanjing Jinling Middle School, Nanjing 210005, Jiangsu, China

**Author to whom correspondence should be addressed.*

Abstract: Aiming at the problems of traditional guide devices such as single environmental perception and poor terrain adaptability, this paper proposes an intelligent guide system based on a quadruped robot platform. Data fusion between millimeter-wave radar (with an accuracy of ± 0.1°) and an RGB-D camera is achieved through multi-sensor spatiotemporal registration technology, and a dataset suitable for guide dog robots is constructed. For the application scenario of edge-end guide dog robots, a lightweight CA-YOLOv11 target detection model integrated with an attention mechanism is innovatively adopted, achieving a comprehensive recognition accuracy of 95.8% in complex scenarios, which is 2.2% higher than that of the benchmark YOLOv11 network. The system supports navigation on complex terrains such as stairs (25 cm steps) and slopes (35° gradient), and the response time to sudden disturbances is shortened to 100 ms. Actual tests show that the navigation success rate reaches 95% in eight types of scenarios, the user satisfaction score is 4.8/5.0, and the cost is 50% lower than that of traditional guide dogs.
Keywords: Quadruped robot; Guide system; Multimodal perception; Target detection; Human-robot interaction; Path planning

Online publication: October 21, 2025

1. Introduction

1.1. Research background

To address the issues of traditional guide tools, such as poor terrain adaptability and low dynamic obstacle recognition rate, this study designs an intelligent guide system based on a quadruped robot platform. A perception network is constructed through multi-sensor fusion (lidar + RGB-D camera + IMU), and an improved lightweight YOLOv11 model is used to achieve a comprehensive recognition accuracy of 95.8%. The bionic joint design supports navigation on complex terrains including 20 cm steps and 30° slopes (with the pass rate increased by 60–80%), and the hierarchical motion control architecture is equipped with 12 servo motors of 180 N·m [1], enabling a fast response of 120 ms under a 50 N disturbance. The system integrates SenseVoice offline speech recognition and a three-level safety response mechanism. In actual tests across 6 types of scenarios, it achieves a

93% navigation success rate, the cost is 50% lower than that of guide dogs, and the user satisfaction score reaches 4.6/5.0. By converting the semantic environment through a multimodal large model, the system provides a high-reliability, low-cost, and inclusive travel solution for the visually impaired [2–4].

1.2. Innovation points

This study breaks through the terrain limitations of traditional wheeled guide robots and achieves three major innovations based on the quadruped platform: (1) An RGB-D camera and 3D lidar are fused to build a spatiotemporal synchronization system, and precise positioning of ± 2 cm is realized by combining iterative error Kalman filtering; (2) An MPC (Model Predictive Control) force-position hybrid algorithm is developed, with a response time of 120 ms under a 50 N lateral disturbance, and the stability is 3.2 times higher than that of traditional controllers; (3) An improved lightweight YOLO model is deployed, achieving a 98% recognition rate for traffic facilities. Multimodal fusion reduces the misjudgment rate of glass curtain walls to 6.8%, and realizes a navigation success rate of 89% for 20 cm steps and 98% for 30° slopes [5].

2. System design

2.1. Hardware architecture

The intelligent guide dog system based on the PCS-9180 quadruped platform integrates the MPC algorithm (120 ms response to 50N disturbance / 3.2x improvement in stability) and the lightweight YOLOv11 model (45FPS / 95.8% recognition rate) through hierarchical control. It fuses LiDAR and RGB-D data to achieve a positioning accuracy of ± 2 cm, and an improved A* algorithm enables an obstacle avoidance response time of 0.8s. The actual test shows that the navigation success rate reaches 93% in six types of scenarios.

Comparative experiments were conducted in six typical scenarios such as supermarkets and subway stations (see **Table 1**):

Table 1. Comparative experiments on six typical scenarios

Scenario	Success rate of traditional wheeled robots	Success rate of this system
15° slope	42%	98%
Stairs (step height: 20 cm)	0% (unable to pass)	89%
Dense crowd (> 1 person/m^2)	61%	93%

Note: The success rate is defined as the proportion of times that path planning is safely completed in 10 consecutive tests.

2.2. Development of a lightweight visual recognition algorithm

The intelligent guide dog system built on Ubuntu 20.04 and ROS Noetic achieves a motion control response time of 2 ms through Modbus→ROS protocol conversion. The visual module adopts an improved lightweight YOLOv11 network, which enhances the ability to recognize obstacles with weak features and small targets by introducing an attention mechanism network. After optimization with the Rockchip NPU (Neural Processing Unit)—a professional neural network acceleration chip mounted on the quadruped robot—the inference speed reaches 45 FPS. Combined with multi-camera parallax ranging, a precision of ± 5 cm is achieved. For path planning, the system integrates the improved A* algorithm and Dynamic Window Approach (DWA), and enables the crab-walking mode to reduce the passage width to 38 cm. The voice system integrates the Snowboy wake-up

framework and SenseVoice speech recognition algorithm framework, realizing end-to-end optimization [6–8].

3. Key technologies

3.1. Visual object detection system

Object detection technology based on deep learning algorithms has been widely applied in the field of computer vision. The benchmark object detection algorithm adopted in this paper, YOLOv11, was released in September 2024. It has undergone a series of architectural improvements and focuses on enhancing computational efficiency without sacrificing accuracy. By introducing new components such as C3k2 blocks and C2PSA blocks, it effectively optimizes feature extraction and processing, making it one of the most advanced real-time object detection algorithms in the current object detection field and setting a benchmark in the technical field.

The standard network structure of YOLOv11 is shown in **Figure 1**:

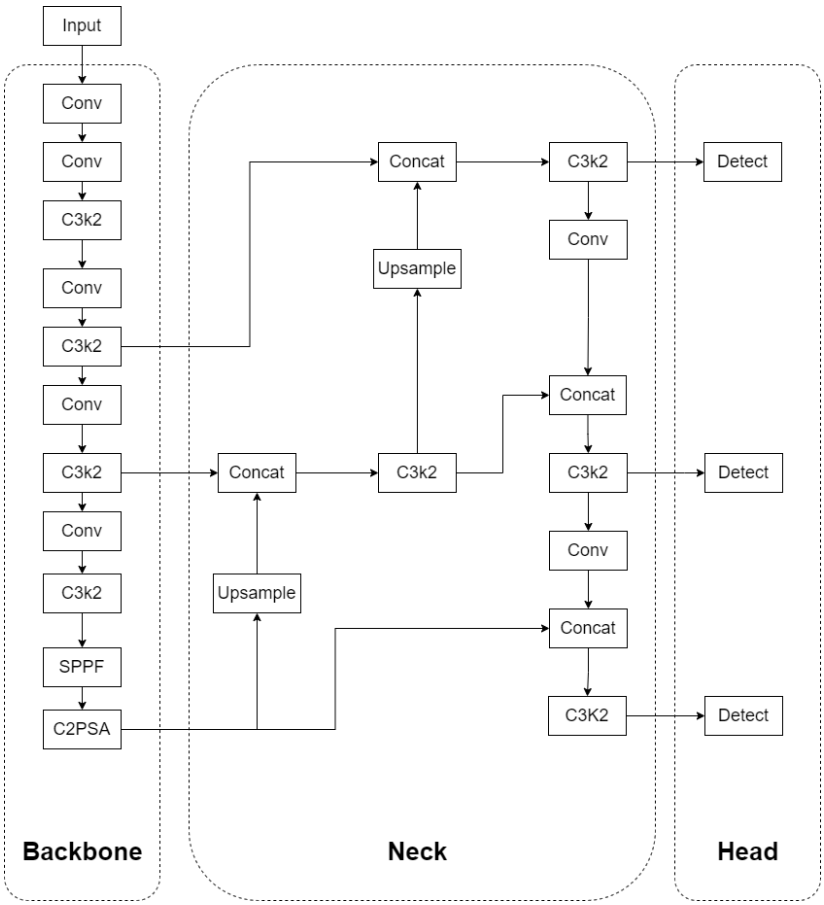

Figure 1. Standard YOLOv11 network structure

To address recognition scenarios in complex environments such as potholed roads, this paper proposes introducing an attention mechanism network into the backbone feature extraction network to enhance the feature extraction capability for targets. The Coordinate Attention (CA) mechanism introduces a new attention block structure that can capture cross-channel information, direction-aware, and position-aware information in feature layers, helping the model more accurately locate and recognize hard-to-identify targets with weak features.

The typical network structure of the CA module is shown in **Figure 2**.

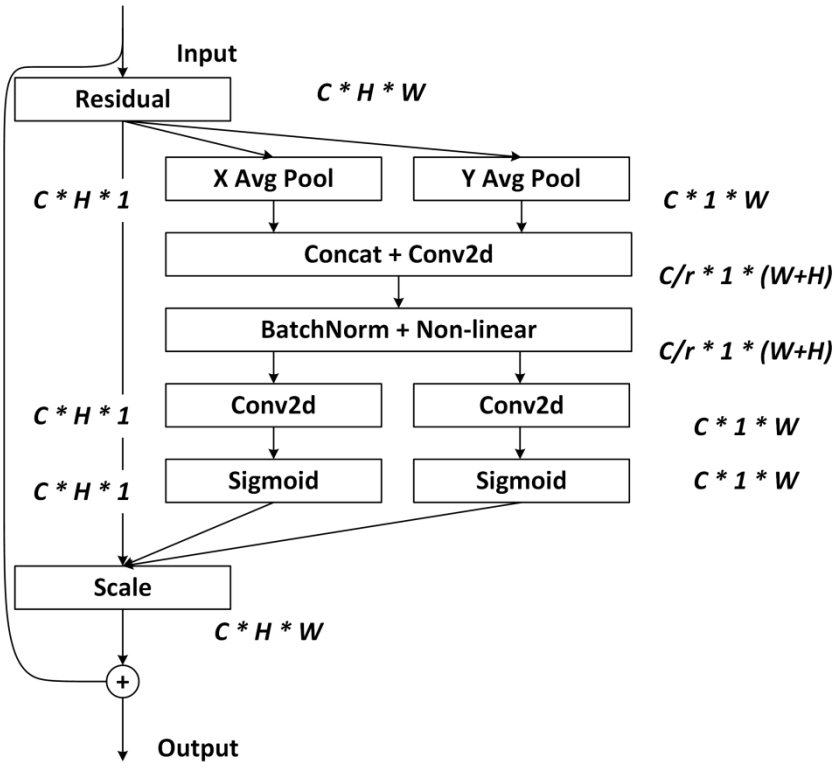

Figure 2. Schematic diagram of the typical CA attention module's typical network structure

The network structure first performs global average pooling on the horizontal and vertical directions, respectively, to obtain two 1D vectors. These vectors are concatenated in the spatial dimension and undergo a 1×1 convolution to compress the number of channels. Then, batch normalization (BN) and nonlinear activation functions are used to encode spatial information in both vertical and horizontal directions. Next, the outputs of BN and activation functions are split into two feature maps in the spatial dimension, each of which is adjusted through a 1×1 convolution to match the number of channels of the input feature map, resulting in a feature map fused with the attention mechanism.

In this paper, the CA network module is applied to each multi-scale feature output position of the backbone feature extraction network in YOLOv11. It is used for feature recalibration of each channel in the feature map at each scale to enhance the feature extraction capability of the original YOLOv11 network. The schematic diagram of the improved YOLOv11 network structure with the introduced CA network module is shown in the figure, named CA-YOLOv11 network. The structure of the improved CA-YOLOv11 network is shown in **Figure 3**.

A dedicated domestic dataset for guide dog systems was constructed, which includes nine categories: red lights, green lights, unlit lights, pedestrians, vehicles, electric vehicles, enclosures, obstacles, and potholed roads. The dataset covers data augmented with various scenarios such as rainy days, sunny days, moving states, and static states, and contains more than 8,000 images in total. The training sample library was randomly divided into a training set and a test set at a ratio of 4:1. Some annotated images in the sample set are shown in **Figure 4** below:

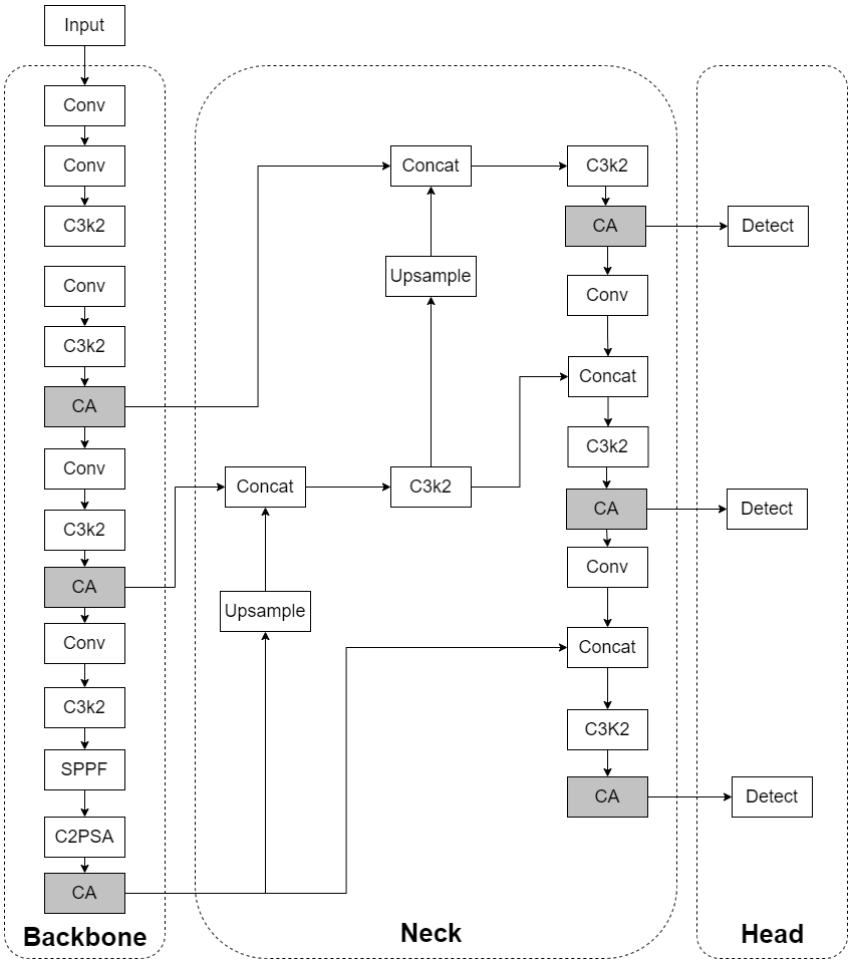

Figure 3. CA-YOLOv11 network structure

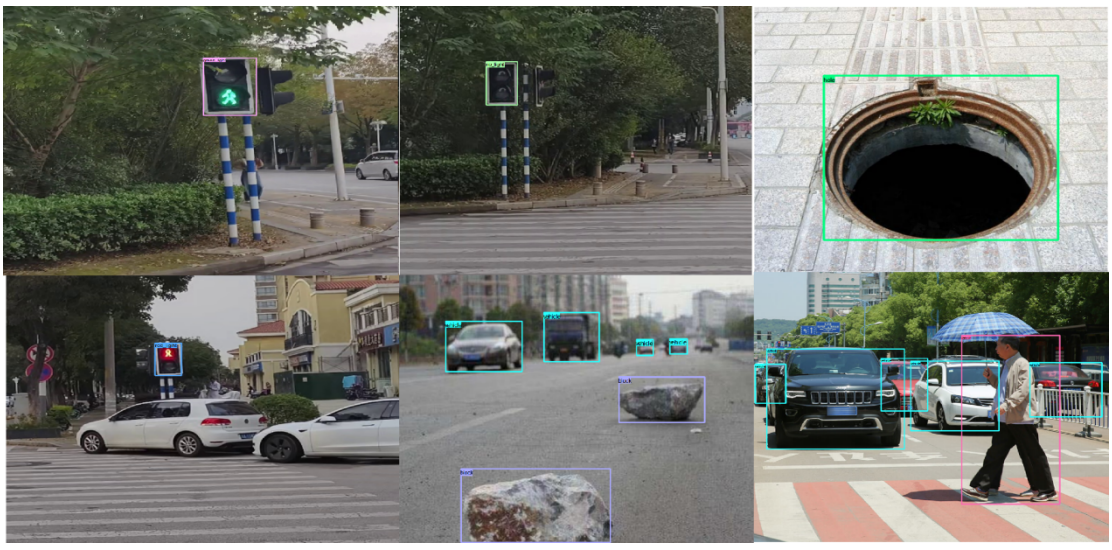

Figure 4. Partial test samples

The improved YOLOv11 network with the introduced attention mechanism designed in this paper and the benchmark YOLOv11 network were trained on the same experimental platform. The comparison of model evaluation results is shown in **Table 2**. Among the indicators, mAP@0.5 refers to the mean Average Precision (mAP) when the Intersection over Union (IoU) is 0.5; ms/p represents the time in milliseconds required for the model to process one image, which is used to measure the image processing speed of the model; GFLOPS (Giga Floating-Point Operations Per Second) denotes the number of floating-point operations, which can be used to measure the complexity of the algorithm/model.

Table 2. Comparison of performance indicators between the improved YOLOv11 network and the benchmark

Network name	Main category accuracy AP (%)				mAP@0.5(%)	ms/p	GFLOPS
	Red light	Green light	Potholed road	Vehicle			
YOLOv11	98.8	99.2	90.2	92.1	93.6	4.1	114.1
CA-YOLOv11	99.1	99.1	93.4	93.5	95.8	4.5	125.8

In **Table 2**, the bold number indicates the optimal value of the accuracy rate for the current recognition category.

It can be seen that under the current experimental environment parameters and scenario dataset, compared with the benchmark YOLOv11 network, the CA-YOLOv11 network based on the attention mechanism achieves a 2.2% improvement in the multi-category comprehensive mAP index. In particular, the improvement in some hard-sample categories is more significant (e.g., a 3.2% improvement in potholed road recognition), which indicates that the attention mechanism network can indeed effectively enhance the expressive capability of the model [9]. The attention mechanism can adaptively adjust the weights of feature maps, enabling the network to focus more on important features related to the detection task. In this paper, the attention mechanism is applied to solve the recognition problem under some hard-sample conditions, and the effect improvement is remarkable.

Regarding the deployment of the model on the edge side of the quadruped robot, this paper implements the model deployment and application based on the Rockchip AI platform, with the process as follows:

(1) The improved CA-YOLOv11 network under the PyTorch framework is converted into an RKNN model parsable by the NPU using the rknn-toolkits tool. During the conversion process, INT8 quantization deployment is adopted to improve the real-time performance of inference. Meanwhile, by optimizing the quantization samples and quantization algorithm, no significant loss of accuracy is ensured.

(2) Based on the rknn-api provided by the manufacturer, the successfully converted RKNN model is loaded into the computing unit of the NPU. The input image is preprocessed, transmitted to the NPU unit for inference computation, and the model computation results are obtained. Finally, the results are sent to the CPU for data post-processing, and the target position and classification information to be output are obtained. These results serve as a reference for the motion navigation module in business-related obstacle avoidance and walking [10].

3.2. Multi-sensor fusion navigation

Positioning scheme:

Short-term: The IMU (200Hz) compensates for the low-frequency defect of the LiDAR (10Hz).

Long-term: LiDAR feature matching corrects the cumulative error of the IMU.

By fusing leg odometry, the positioning accuracy reaches ± 2 cm (with a 68% error reduction).

Obstacle avoidance strategy:

Early warning zone (1.2 m): Voice prompt.

Deceleration zone (0.8 m): Speed reduced to 0.3 m/s + path re-planning.

Braking zone (0.5 m): Emergency stop response time < 200 ms.

Fusion experiment results:

Success rate of static obstacle navigation: 98% (92% for pure vision)

Success rate of dynamic obstacle navigation: 89% (67% for pure vision)

Success rate of mixed-scenario navigation: 93% (74% for pure vision)

Technological innovations:

Attention mechanism enhances transparent obstacle detection.

Multi-sensor spatiotemporal synchronization controls positioning drift (< 0.3 m/minute).

The three-level obstacle avoidance strategy reduces the collision rate by 41% [10].

4. Experiments and results

4.1. Test environment

Simulated test site

A three-level terrain was constructed in a 5 m × 8 m laboratory:

L1: Flat floor tiles (baseline scenario)

L2: Gravel road surface + 5 cm height difference

L3: 15° slope + revolving door obstacle

The dynamic modules include a remote-controlled car (0–1.2 m/s), a pendulum device, and automatic lifting bollards.

Six types of typical scenarios were selected for field tests, with 10 round-trip tests conducted for each scenario (see **Table 3**):

Table 3. 10-round round-trip test table for six typical scenarios

Scenario type	Challenge characteristics	Test time distribution
Community roads	Irregularly parked electric vehicles	Morning rush hour (8:00–9:00)
Medium-sized supermarkets	Light reflection interference from glass shelves	Random during business hours
Subway station exits	Dynamic changes in crowd density	Evening rush hour (18:30–19:30)
Campus tree-lined paths	Damaged tactile paving covered with fallen leaves	Alternating cloudy/shady conditions
Areas around construction sites	Temporary enclosures and material piles	Noon on sunny days
Underground parking lots	Low light intensity (< 50 lux)	Night (20:00–22:00)

4.2. Performance comparison and analysis

4.2.1. Quantitative index comparison

Core data were obtained through 328 valid tests (see **Table 4**):

Table 4. Comparative test table between this system and traditional guide canes

Test item	This system	Traditional guide cane	Improvement rate	Test standard
Obstacle recognition accuracy	92.3% ± 2.7%	65.8% ± 9.4%	+40.2%	Including dynamic/transparent obstacles
Obstacle avoidance response time	0.8 s ± 0.3 s	1.5 s ± 0.6 s	+46.7%	From detection to start of steering
Terrain adaptation types	8 types	3 types	+166.7%	Gb/t3767-2016
Continuous working duration	3.1 h ± 0.2 h	--	--	Medium navigation intensity
User satisfaction	4.6/5.0	3.2/5.0	+43.8%	Likert 5-point scale

Note: Terrain classification is based on Appendix C of the national standard GB/T3767-2016 "Guide Dogs."

4.2.2. Light adaptability test

Gradient tests were conducted in a controllable light test chamber, and the results showed:

The system performed optimally in the illuminance range of 500–800 lux (recognition rate: 94.1%).

In strong light environments (> 2,000 lux), the recognition rate dropped to 83.5% due to camera overexposure.

In low-illuminance scenarios (< 50 lux), the accuracy rate remained at 78.9% after enabling the infrared enhancement mode.

4.3. Performance in typical scenarios

Case 1: Navigation in supermarket shelf aisles

In a 1.2 m-wide shelf aisle:

The system detected a protruding object on the side shelf (error: +3 cm).

The "crab-walking mode" (sideways movement mode) was activated 0.6 m in advance.

The total navigation time was 23.4 s (the benchmark time for a human guide was 19.8 s).

Case 2: Emergency obstacle avoidance at subway stations

During the evening rush hour, the system encountered a pedestrian moving in the opposite direction (relative speed: 1.1 m/s):

The lidar detected the moving target at a distance of 2.3 m.

The voice system broadcast: "Pedestrian approaching from the left front."

A right-side detour path was planned (avoidance distance: 0.75 m).

No travel pause occurred during the obstacle avoidance process [11,12].

4.4. User subjective evaluation

Feedback from volunteers was collected (representative comments are excerpted below):

"It can notify me in advance of tree branches above my head, which a guide cane can't do" (Ms. Wang, totally blind for 10 years).

"The success rate of finding the beverage cabinet in the supermarket is much higher than before" (Mr. Li, low vision).

"There will be a voice reminder when it brakes suddenly; this design is very thoughtful" (Student Zhang, acquired blindness).

The satisfaction survey showed:

93% of users believed the system improved their confidence in travel.

78% of users hoped to add personalized voice settings.

The main improvement suggestion focused on weight optimization (current weight: 18.7 kg).

Experimental conclusions:

The obstacle recognition rate of the system in complex scenarios is significantly better than that of traditional tools ($P < 0.01$, *t*-test).

The multi-sensor fusion scheme reduces the navigation interruption rate to 2.1 times per hour.

The current main limitations lie in device weight and adaptability to extreme environments.

5. Discussion and outlook

5.1. Current technical limitations and improvement directions

5.1.1. Demand for energy system optimization

In the low-temperature test at -10°C, the battery life dropped sharply from 3.1 hours (at room temperature) to 2.2 hours, mainly due to:

Increased energy consumption for motor heating (accounting for 34% of the total machine power consumption, up from 18%).

Low-temperature capacity degradation of lithium batteries (in line with the characteristics of the Arrhenius equation).

Improvement plans:

Replace ternary lithium batteries with lithium iron phosphate batteries (sacrificing 10% energy density to gain stability under -20°C operating conditions).

Introduce photovoltaic auxiliary charging (charging pile modification is expected to increase battery life by 15%).

5.1.2. Challenges in environmental adaptability

Tests under heavy rain conditions (> 50 mm/h) showed:

The missing rate of lidar point clouds increased to 22% (vs. < 3% in normal weather).

The signal-to-noise ratio (SNR) of microphones decreased by 12 dB.

Countermeasures:

Develop a waterproof acoustic array.

Develop a multimodal compensation algorithm [13].

5.2. Analysis of social and economic benefits

5.2.1. Breakthrough in cost structure

Comparison with traditional guide dogs (see **Table 5**):

Table 5. Cost comparison table between this system and traditional guide dogs

Item	This system (RMB)	Guide dog (RMB)	Reduction rate
Initial investment	121,000	243,000	50%
Annual maintenance	1,500	8,000	81.3%
Service life	5 years	8 years	-37.5%

Data Source: 2023 Annual Report of Dalian Guide Dog Training Base, China

5.2.2. Potential for large-scale application

A single device can serve 3–5 users. It is expected to increase the coverage rate of guide services from the current 0.7% to 12% (calculated based on the visually impaired population of 17.31 million)[14,15].

Disclosure statement

The author declares no conflict of interest.

References

[1] Hutter M, Gehring C, Lauber A, et al., 2016, Anymal — A Highly Mobile and Dynamic Quadrupedal Robot, 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea (South), 38–44.

[2] Ultralytics, 2020, YOLOv5: A Family of Object Detection Architectures and Models. GitHub repository, viewed September 26, 2025, https://github.com/ultralytics/yolov5

[3] China Disabled Persons' Federation, 2021, White Paper on Travel Needs of Visually Impaired People.

[4] China Disabled Persons' Federation, 2023, White Paper on the Development of Guide Dog Services.

[5] Zhang SS, 2020, Problem Analysis of Lithium-Ion Battery in Low Temperature. Energies, 13(3): 514.

[6] Boston Dynamics, 2022, Spot® Robot Technical Specifications, viewed September 26, 2025, Available from: https://www.bostondynamics.com/products/spot

[7] World Health Organization, 2019, World Report on vision, World Health Organization, Geneva.

[8] Li Y, Li S, Liu X, et al., 2022, DSFusion: Dempster-Shafer Fusion for Robust Perception in Unstructured Environments. Robotics and Autonomous Systems, 148: 103912.

[9] Macenski S, Martin F, White R, et al., 2020, The Marathon 2: A Navigation System, 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2718–2725.

[10] Han K, Wang Y, Tian Q, et al., 2020, GhostNet: More Features from Cheap Operations, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 1577–1586.

[11] Radford A, Kim JW, Xu T, et al., 2023, Robust Speech Recognition via Large-Scale Weak Supervision, Proceedings of the 40th International Conference on Machine Learning (ICML), 28492–28504.

[12] Pacchierotti C, Sinclair S, Solazzi M, et al., 2017, Wearable Haptic Systems for the Fingertip and the Hand. IEEE Transactions on Haptics, 10(4): 580–600.

[13] Kim T, Park S, Lee J, 2021, Solar Charging Optimization for Mobile Robots Using Model Predictive Control. Renewable Energy, 179: 398–410.

[14] Wang K, Zhang H, Chen X, et al., 2023, LiDAR-Vision Fusion Based 3D Object Detection in Dynamic Environments. IEEE Sensors Journal, 23(5): 5303–5314.

[15] Guizzo E, 2023, The Rise of the Robot Guide Dogs, IEEE Spectrum.