Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 22, No. 1, 2026

Research on an Air-Ground Community Governance System with Wheel-Leg Composite Chassis and UAV Visual-Aided Collaboration

Siyu Wang, Jiayi Hao, Shiyi Yu, Bo Cui

North China University of Science and Technology, Tangshan 063210, China
* Corresponding author: 18714689@qq.com

Abstract: To address three prominent challenges prevailing in aging residential districts, including omnipresent three-dimensional sensing blind zones, insufficient traversal capability over unstructured terrains and lagged emergency disposal responses, the proposed "Zhikan-Yunxun" system equips the robot with a wheel-legged composite chassis to achieve stable traversal over steps and diverse ground obstacles. An unmanned aerial vehicle (UAV) is pre-stored inside the roof-mounted cabin and kept in standby mode; upon the occurrence of visual occlusion, the UAV takes off on demand and leverages the lightweight LAF-YOLOv10 detection algorithm to compensate for ground-level sensing blind zones from the aerial viewpoint. The UAV and unmanned ground vehicle (UGV) perform coordinated operation relying on UWB/IMU/vision hierarchical fusion localization and MOCHA opportunistic communication framework, which sustains centimeter-level navigation precision under complete GNSS signal deprivation. Furthermore, a SCARA manipulator is deployed to conduct flexible manipulation tasks such as litter retrieval and on-site emergency support. Benefiting from the integrated air-ground coordination architecture, raw sensing data can be directly converted into executable community governance interventions.

Keywords: Air-ground collaborative robot; wheel-legged composite chassis; LAF-YOLOv10; multi-sensor fusion localization; MOCHA opportunistic communication; smart community governance; digital twin.

1. Introduction

Against the backdrop of digital-intelligent transformation in China’s urban governance, the low-altitude economy has been designated a national strategic emerging industry. As stated in the *White Paper on China’s Low-Altitude Economy (2025)*, the domestic market scale of the low-altitude economy surpassed 600 billion RMB in 2025. The official issuance of the *Interim Measures for the Administration of Unmanned Aircraft Flight* in 2026 delivers solid policy backing for the large-scale deployment of unmanned aerial vehicles (UAVs) and intelligent robots within residential communities. Nevertheless, such policy incentives cannot fully address the persistent governance challenges of aging residential districts. Taking the 4,827 residential compounds in Tangshan City as a statistical sample, aging communities account for 38% of the total housing stock. Occlusion from high-rise buildings and dense vegetation generates extensive surveillance blind zones. Perennial hazards including illegally parked electric two-wheelers, high-rise littering, and blocked fire escape passages cannot be continuously regulated through conventional manual patrols and fixed video surveillance systems.

Three prominent bottlenecks impede current community governance practices. First, high-rise architectures and dense vegetation jointly form three-dimensional visual blind zones, making real-time detection and traceability of concealed hazards unfeasible. Second, unstructured road surfaces such as potholed pavements, snow-covered roads, and stepped terrains restrict the traversal performance of wheeled inspection robots; urban canyon effects induced by dense buildings block GNSS satellite signals, thereby depriving inspection platforms of mobility and positioning functions. Third, conventional surveillance equipment merely supports video data acquisition without on-site intervention modules. Emergencies such as sudden fires and unauthorized vehicle occupation require manual on-site disposal, which frequently misses the optimal response window and escalates minor hidden risks into severe safety incidents.

Numerous studies concerning air-ground collaborative robotic systems have been conducted worldwide, yet existing technical solutions suffer from prominent limitations. Cladera et al. proposed the MOCHA opportunistic communication framework, which only supports information interaction among multiple robots without physical actuation modules. The aerial detection model LAF-YOLOv10 proposed by Farooqui incurs excessive computational overhead under multi-task parallel scheduling and fails to adapt to scenarios featuring tiny scattered debris in residential areas. Cheng integrated UWB and visual-inertial algorithms to achieve centimeter-level relative docking under GNSS-denied conditions, whereas the overall computational burden remains prohibitively high. Dominguez merely carried out experiments on static marker registration alignment. Domestic lightweight algorithms such as EDNet and MambaNeXt solely concentrate on image feature extraction, failing to construct an end-to-end pipeline integrating hazard detection, autonomous navigation, and on-site disposal. Conventional robotic equipment generally lacks the capability to operate over complex terrains and execute emergency intervention tasks, resulting in an incomplete closed-loop workflow for community governance.

To tackle the above scenario-specific and technical drawbacks, this paper proposes an intelligent air-ground collaborative community governance system named “Zhikan-Yunxun”. Equipped with a wheel-legged composite chassis for traversing unstructured terrains, the system integrates an auto-retractable UAV to compensate for ground visual blind zones via overhead aerial observation. Lightweight vision algorithms are deployed to identify various micro hazards within communities. Combined utilization of the MOCHA communication framework and multi-sensor fusion localization enables coordinated navigation inside building clusters. A flexible SCARA manipulator is configured to perform physical operations including litter clearance and incipient fire suppression. Moreover, large language models and a digital twin platform are integrated to realize intelligent task scheduling and full-domain situational awareness monitoring.

The proposed system exhibits three core technical and practical merits. Firstly, the dual-layer air-ground perception network eliminates three-dimensional surveillance dead zones, transforming passive routine inspection into active hazard early warning and reducing the occurrence rate of safety accidents. Secondly, the coordination between obstacle-surmounting mechanisms and multi-sensor fusion positioning breaks dual restrictions imposed by rough terrains and satellite signal occlusion, covering the "last hundred meters" of full-coverage community inspection. Thirdly, a complete disposal chain ranging from hazard identification to on-site intervention is established, which shortens emergency response latency and upgrades community governance from simple passive monitoring to proactive on-site intervention.

2. Research Background and Theoretical Foundation

2.1. Macro Trends: Boom of Low-Altitude Economy and Digital Transformation of Community Governance

China's urbanization has entered a stage of high-quality development, and urban governance is undergoing an upgrade from refined management to digital-intelligent governance. According to the *White Paper on China's Low-Altitude Economy (2025)* issued by the Ministry of Industry and Information Technology (MIIT), the market scale of China's low-altitude economy exceeded 600 billion RMB by the end of 2025, and a trillion-yuan market is expected to take shape by 2030[1]. The full implementation of the *Interim Measures for the Administration of Unmanned Aircraft Flight* in 2026 officially incorporated the low-altitude economy into the list of national strategic emerging industries[2].

Driven by technological dividends, communities, as the smallest units of urban governance, are confronted with prominent safety and efficiency challenges. Taking Tangshan City as a research sample, aging residential districts account for 38% of its total 4,827 residential compounds. Restricted by narrow building spacing, dense vegetation occlusion and extensive surveillance blind zones, such communities suffer frequent hidden dangers including blocked fire escape passages and high-rise littering[10]. The conventional governance framework combining manual patrols and fixed video surveillance can hardly satisfy practical demands, which necessitates the introduction of an innovative air-ground collaborative governance system.

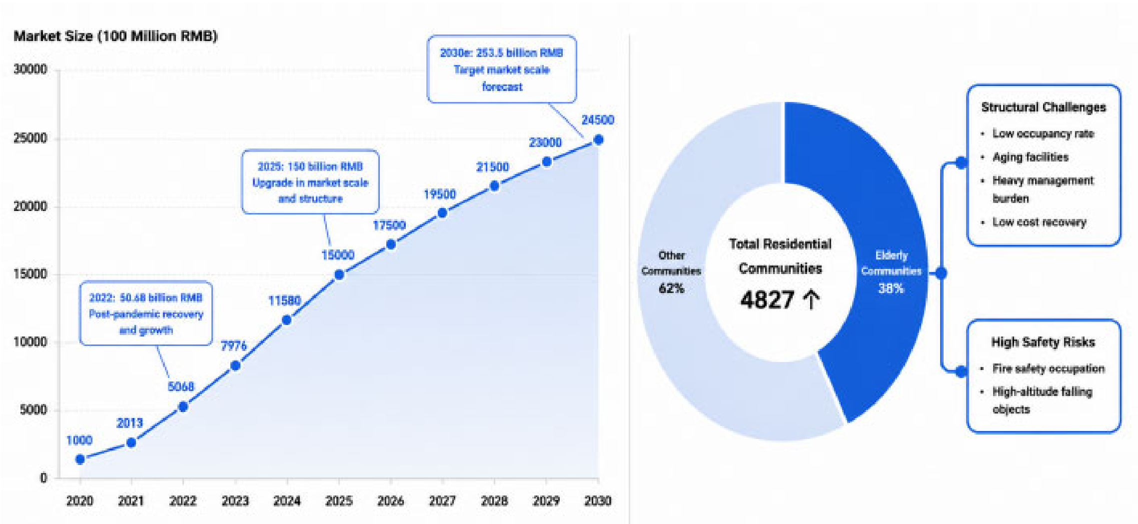
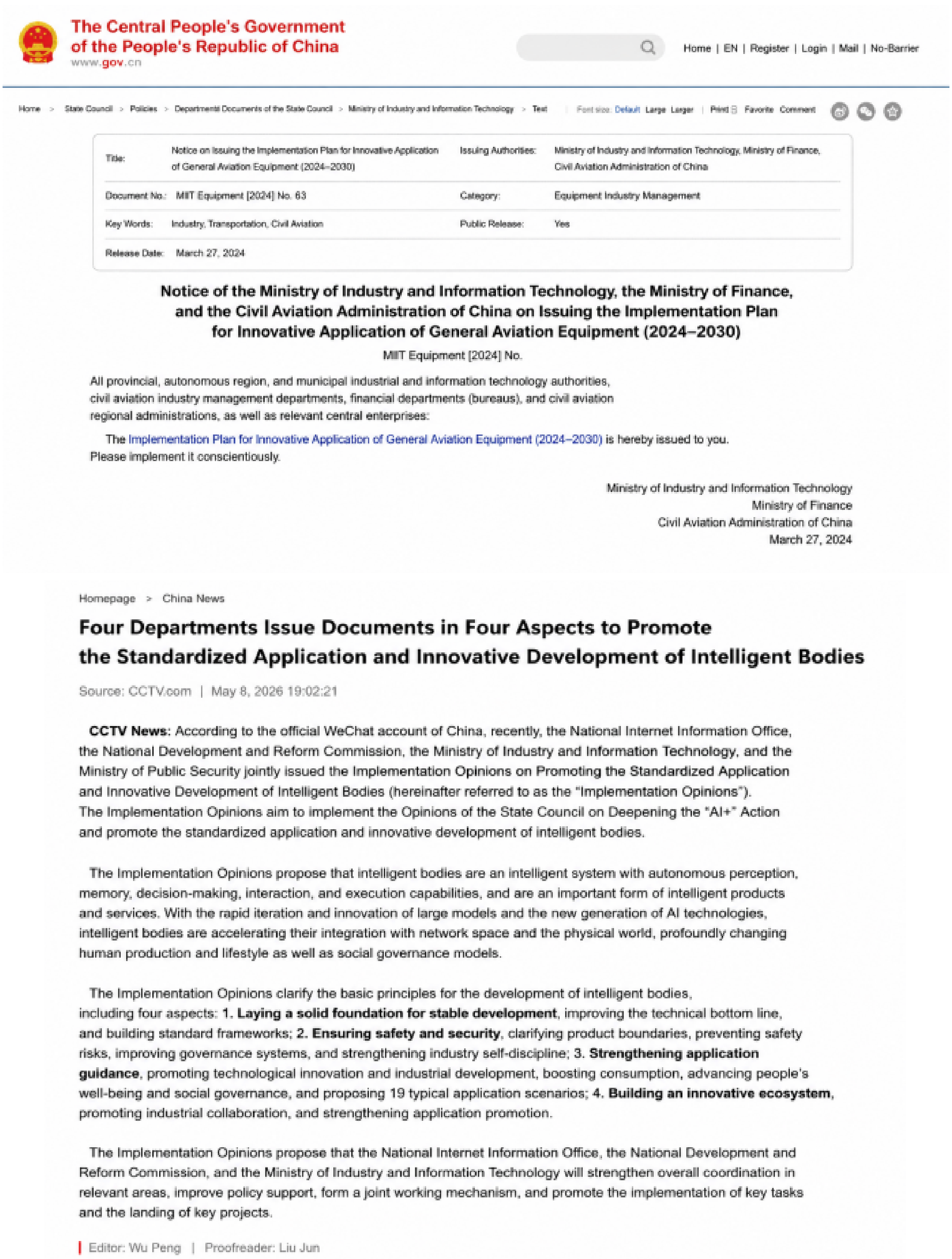


Figure 1. Forecast trend of low-altitude economic market scale (left) and distribution diagram of community security pain points in Tangshan City (right)

2.2. Policy Orientation: Intelligent Equipment and Labor Substitution by Robots

The vigorous expansion of the low-altitude economy has not only unlocked technological opportunities, but also received strong policy support from central and local authorities. As illustrated in Figure 2, relevant national ministries and commissions have issued official approvals to vigorously promote the large-scale deployment of the "UAV + Robot" integrated system in scenarios such as security defense and environmental sanitation. At present, "new quality productive forces" and the "AI Plus" initiative have become core national strategies. Governments at all levels are committed to advancing the in-depth integration of high-end equipment manufacturing and social governance. Both the *14th Five-Year Plan for National Emergency Response System* and the *Implementation Plan for Innovative Application of General Aviation Equipment* issued by the Ministry of Industry and Information Technology (MIIT) emphasize the urgent need to address industrial pain points including "surveillance blind zones" and "high risks of manual patrols". Such favorable policy context creates an optimal research and development environment for intelligent inspection and operation robots[3][4].



The Central People's Government of the People's Republic of China
www.gov.cn

Home | EN | Register | Login | Mail | No-Barrier

Home > State Council > Policies > Departmental Documents of the State Council > Ministry of Industry and Information Technology > Text | Font size: Default Large Larger | Print Favorite Comment

Title:	Notice on Issuing the Implementation Plan for Innovative Application of General Aviation Equipment (2024–2030)	Issuing Authorities:	Ministry of Industry and Information Technology, Ministry of Finance, Civil Aviation Administration of China
Document No.:	MIIT Equipment [2024] No. 63	Category:	Equipment Industry Management
Key Words:	Industry, Transportation, Civil Aviation	Public Release:	Yes
Release Date:	March 27, 2024		

Notice of the Ministry of Industry and Information Technology, the Ministry of Finance, and the Civil Aviation Administration of China on Issuing the Implementation Plan for Innovative Application of General Aviation Equipment (2024–2030)

MIIT Equipment [2024] No.

All provincial, autonomous region, and municipal industrial and information technology authorities, civil aviation industry management departments, financial departments (bureaus), and civil aviation regional administrations, as well as relevant central enterprises:

The Implementation Plan for Innovative Application of General Aviation Equipment (2024–2030) is hereby issued to you. Please implement it conscientiously.

Ministry of Industry and Information Technology
Ministry of Finance
Civil Aviation Administration of China
March 27, 2024

Homepage > China News

Four Departments Issue Documents in Four Aspects to Promote the Standardized Application and Innovative Development of Intelligent Bodies

Source: CCTV.com | May 8, 2026 19:02:21

CCTV News: According to the official WeChat account of China, recently, the National Internet Information Office, the National Development and Reform Commission, the Ministry of Industry and Information Technology, and the Ministry of Public Security jointly issued the Implementation Opinions on Promoting the Standardized Application and Innovative Development of Intelligent Bodies (hereinafter referred to as the "Implementation Opinions"). The Implementation Opinions aim to implement the Opinions of the State Council on Deepening the "AI+" Action and promote the standardized application and innovative development of intelligent bodies.

The Implementation Opinions propose that intelligent bodies are an intelligent system with autonomous perception, memory, decision-making, interaction, and execution capabilities, and are an important form of intelligent products and services. With the rapid iteration and innovation of large models and the new generation of AI technologies, intelligent bodies are accelerating their integration with network space and the physical world, profoundly changing human production and lifestyle as well as social governance models.

The Implementation Opinions clarify the basic principles for the development of intelligent bodies, including four aspects: 1. **Laying a solid foundation for stable development**, improving the technical bottom line, and building standard frameworks; 2. **Ensuring safety and security**, clarifying product boundaries, preventing safety risks, improving governance systems, and strengthening industry self-discipline; 3. **Strengthening application guidance**, promoting technological innovation and industrial development, boosting consumption, advancing people's well-being and social governance, and proposing 19 typical application scenarios; 4. **Building an innovative ecosystem**, promoting industrial collaboration, and strengthening application promotion.

The Implementation Opinions propose that the National Internet Information Office, the National Development and Reform Commission, and the Ministry of Industry and Information Technology will strengthen overall coordination in relevant areas, improve policy support, form a joint working mechanism, and promote the implementation of key tasks and the landing of key projects.

Editor: Wu Peng | Proofreader: Liu Jun

Figure 2. Screenshots of official documents issued by national ministries for promoting the innovative application of intelligent equipment

To clarify the policy transmission path of the proposed system, this study establishes a framework of the policy support system, as illustrated in Figure 3. The framework adopts a three-level progressive structure. At the top level, national strategies are driven by the concepts of "new quality productive forces" and low-altitude economy; the middle level relies on relevant plans issued by the Ministry of Industry and Information Technology (MIIT) and the Ministry of Housing and Urban-Rural Development (MOHURD); the bottom level corresponds to detailed construction regulations of Hebei Province and Tangshan City. The three tiers of policies are closely interlocked, jointly orienting the technical research priorities of the "Zhikan-Yunxun" air-ground collaborative system and providing institutional guarantees for its practical deployment and application[5][6].

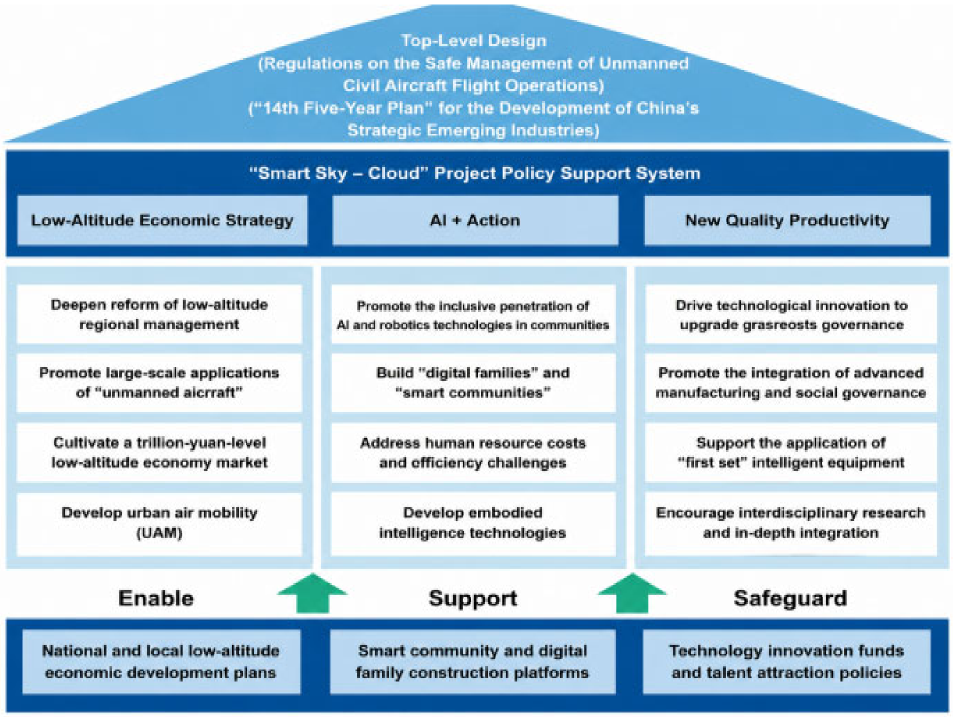


Figure 3. Diagram of national and local policy support system for intelligent robot applications

2.3. Industry Pain Points: Three Practical Dilemmas of Traditional Community Governance

Through field surveys and reviews of actual governance conditions in typical residential communities of Tangshan (Xiangfuli and Peirenli), it is found that the "fixed surveillance + manual patrol" model fails to satisfy complex community management demands and encounters three major bottlenecks, namely insufficient full-domain perception, poor operability in complicated environments, and broken closed-loop emergency disposal. Such governance difficulties constitute the core motivation for developing the "Zhikan-Yunxun" air-ground collaborative system in this project[10].

Figure 4. Panoramic diagram of core dilemmas in traditional community governance

Pain Point 1: Difficulties in three-dimensional space supervision with extensive visual blind zones and safety dead zones Aging residential blocks and high-density residential quarters feature highly three-dimensional and complicated environments.

Dense vegetation, narrow building spacing and intricate public facilities render conventional two-dimensional supervision frameworks ineffective. In practical scenarios, illegally parked electric vehicles inside building corridors are concealed within surveillance dead zones; high-rise littering lacks traceable evidence, and accumulated litter as well as damaged facilities in green belts remain unsupervised. Such systems suffer a perceptual limitation: they can only capture planar information while failing to cover three-dimensional spaces, and merely monitor main roads with no coverage of hidden corners, resulting in numerous latent safety hazards. Therefore, it is urgent to deploy UAVs with overhead bird's-eye views to construct an integrated air-ground full-domain perception network to address the above limitations[7].

Figure 5. Typical scenarios of three-dimensional supervision blind zones in aging residential communities

Pain Point 2: Difficulties in operations under complex scenarios confronted with dual challenges of urban canyons and terrain barriers The internal environment of residential communities belongs to unstructured terrain.

As illustrated in Figure 6, harsh road conditions including muddy potholed surfaces after rain and icy snow-covered pavements in winter frequently trap conventional wheeled inspection robots. Such equipment suffers wheel slipping, chassis grounding and even rollover, falling into a passive state where the robot cannot enter, move forward or evacuate from target areas. Meanwhile, dense high-rise buildings generate urban canyon effects, which trigger multipath occlusion and drift of satellite signals and completely invalidate GNSS positioning[8]. Restricted by the two above obstacles, ground equipment fails to accurately reach operation sites by surmounting terrain barriers and loses navigation and positioning capacity, breaking the full-coverage "last hundred meters" of community governance.

Figure 6. Traversal dilemmas of traditional ground inspection equipment on complex road surfaces

Pain Point 3: Broken closed-loop emergency response leading to passive conditions of "surveillance without intervention" and delayed disposal Current community governance frameworks cannot handle various emergencies (including combustion caused by illegal EV indoor charging, incipient fires and emergency medical demands of elderly residents) due to the bottleneck of "visible hazards but unavailable on-site intervention".

After surveillance systems detect abnormal events, manual teams need to transport devices to the site, yet narrow roads and disorderly parked vehicles in old residential blocks often delay the arrival of ambulances and staff, missing the golden response window for hazard disposal. The non-contact supervision mode only supports image acquisition without physical on-site intervention, which allows minor hidden risks to escalate into severe safety incidents and delays emergency medical rescue[3]. To satisfy diversified emergency demands, intelligent terminals equipped with auxiliary safety intervention and lightweight material delivery functions need to be developed, forming a rapid closed-loop chain covering hazard identification, distress signal reception, robotic on-site disposal and emergency material delivery, so as to fully protect residents' personal and property safety.

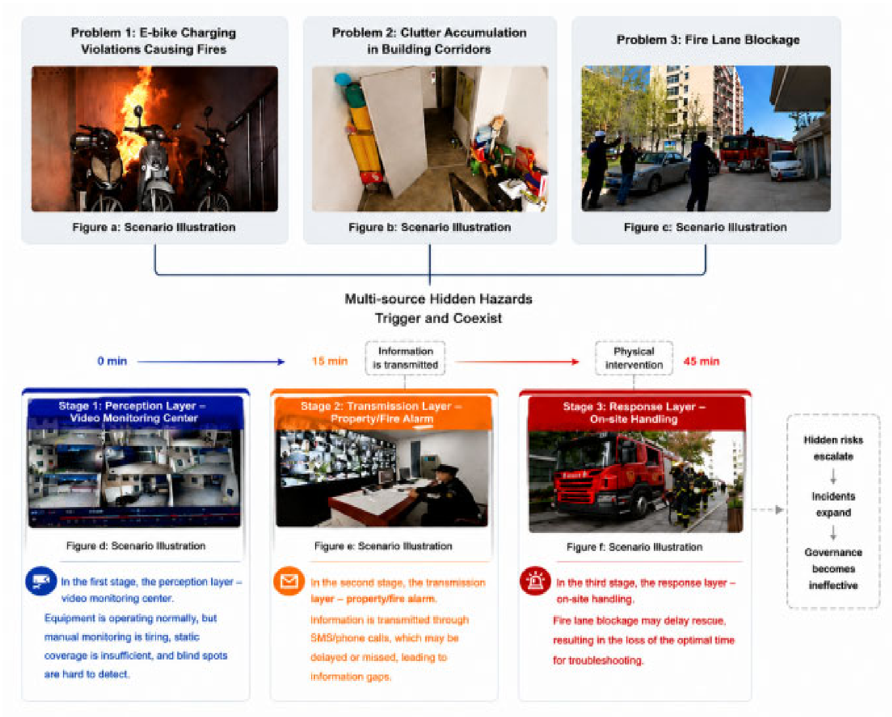


Figure 7. Flowchart illustrating "surveillance without control" and delayed emergency response in traditional community governance

2.4. Theoretical Basis for Technical Feasibility

The technical scheme of this project integrates mature technical frameworks and innovative algorithms, laying a solid foundation for engineering implementation.

(1) Mature and Reliable Air-Ground Cooperative Perception Technology This project adopts a collaborative architecture consisting of UAV wide-area perception and ground robot precise operation. Equipped with high-definition visual sensors, the UAV carries out global inspection. The improved LAF-YOLOv10 model is utilized to identify targets including litter accumulation, fire hazards, illegally parked electric vehicles and abnormal manhole covers, while ground robots conduct close-range inspection and execution in complex areas. The YOLO series algorithms exhibit favorable real-time performance and stability in smart city and unmanned system applications. The research team has completed model training and deployment verification, which confirms the reliability of the technical foundation[9].

(2) Engineering Foundation of Multi-source Fusion Localization and Autonomous Navigation Multi-source fusion localization and autonomous navigation technologies possess sound engineering practicability. To tackle issues such as building occlusion and unstable GNSS signals in residential communities, a fusion localization scheme integrating UWB, IMU, visual odometer and LiDAR is adopted. A unified framework is constructed based on ROS. SLAM mapping, autonomous positioning and dynamic path planning are deployed to realize autonomous navigation and obstacle avoidance in complex scenarios. Such technologies have been widely applied in warehousing logistics, inspection and autonomous driving with high maturity, and the team has verified the ROS platform and corresponding navigation functions[11][12].

(3) MOCHA Opportunistic Communication Ensures Stable System Operation The MOCHA opportunistic communication framework guarantees steady system operation. To address communication interruption caused by building occlusion, the MOCHA framework is introduced. It realizes information transmission over intermittent networks via the "Store-Carry-Forward" mechanism, breaking the reliance of traditional real-time communication on continuous network links and improving the communication robustness of the cooperative system under complex environments. Combined with dynamic UAV relays, task data and status information can be transmitted stably[13][14].

(4) Large Model Intelligent Decision-Making System with Promising Application Prospects Large model intelligent decision-making systems hold bright application prospects. This project introduces large language models into community governance robots. By fusing visual perception data, environmental status and historical task records, the system can accomplish target analysis, task scheduling and disposal decision-making. The large model can automatically generate execution strategies, enhance the system's environmental comprehension and autonomous decision-making capability, and deliver innovative solutions for smart communities[15].

(5) Mature Implementation Conditions of Hardware Platform The hardware platform is supported by fully mature industrialized components. The system integrates off-the-shelf industrial hardware including UAVs, wheel-legged composite robots, NUC edge computing hosts, STM32 controllers, LiDAR, UWB positioning modules and high-definition visual cameras, all of which feature high stability, reliability and convenient secondary development. Unified integration based on ROS reduces development risks and improves engineering efficiency[12].

In summary, the key technologies involved in this project, including object detection, air-ground collaboration, multi-source fusion localization, opportunistic communication and large model-based decision-making, are supported by sound theories and practical cases. The proposed technical route is feasible with strong engineering deployment capacity.

3. System Requirements Analysis

Old residential communities suffer from three major governance challenges: three-dimensional monitoring blind spots, difficult navigation in unstructured environments, and the inability to establish a closed-loop hazard disposal process. These challenges impose clear engineering constraints on the perception, localization, physical operation, air-ground communication, and human-machine interaction capabilities of an air-ground collaborative system. This chapter hierarchically analyzes the functional and performance requirements of each subsystem, providing a foundation for the overall system design presented in Chapter 4.

3.1. Air-Ground Collaborative Perception Requirements

Traditional inspection methods have inherent limitations. A comparison of the capability constraints of different solutions is presented as follows:

Table 1. Comparison of Limitations of Existing Community Inspection Solutions

Comparison Dimension	Traditional Manual Inspection + Fixed Monitoring	Single Ground Robot	Single UAV	Requirements of This System
Navigation in Complex Terrain	Limited human coverage range; fixed monitoring devices cannot move	Unable to overcome stairs and uneven terrain	No ground-based physical operation capability	Equipped with all-terrain mobility and obstacle-crossing capability
Three-Dimensional Spatial Coverage	Only provides planar monitoring; blind spots exist in vegetation areas and building corridors	Limited observation perspective from ground level	Only provides aerial top-down views; unable to perform close-range inspection	Achieve comprehensive air-ground dual-layer monitoring
Hazard Disposal Capability	Only records hazards without on-site intervention capability	Only performs image acquisition without execution mechanisms	No physical operation terminal available	Support on-site disposal of waste, fire hazards, and other incidents
Localization in Occluded Environments	No precise localization capability	Severe positioning drift inside buildings	Only provides rough aerial coordinates	Achieve centimeter-level spatial localization

Community hazards vary significantly in size, including small targets such as cigarette butts and paper scraps, as well as large targets such as electric vehicles and bulky waste. The perception module must adapt to changing illumination conditions, including daytime/nighttime environments and tree-shadowed areas. The maximum model parameter size is limited to 2.5M, and the edge computing platform must achieve an inference frame rate of no less than 20 FPS. When the ground-level field of view is obstructed, the aerial inspection device should automatically take off to compensate for blind spots. The single operation duration should not exceed 3 minutes. After completing target recognition, the aerial device should automatically return to the ground storage area for docking and storage.

3.2. Navigation and Localization Requirements in Unstructured Environments

Residential community environments contain complex terrains such as stairs, muddy roads, and snow-covered surfaces. The minimum obstacle-crossing height of the system should reach 150 mm. High-rise buildings create urban canyon environments where GNSS signals may fail. Under GNSS-denied conditions, the localization error should be controlled within 5 cm. The operating altitude range of the aerial device is 5–15 m, with its operational area restricted within a 20 m radius around the ground vehicle. The aerial unit should be capable of real-time following of the ground platform. After leaving aerial guidance, the ground device should rely on environmental sensors to perform autonomous mapping and local path planning.

3.3. Physical Operation and Emergency Intervention Requirements

The horizontal working radius of the robotic execution mechanism should be no less than 600 mm, enabling autonomous recognition and grasping of scattered objects on the ground. After returning to the docking station, the system should automatically complete UAV propeller and body cleaning and maintenance operations. For emergency scenarios, the system should support coordinated control capabilities, including: Blocking electric vehicles from entering elevators; Deploying a spraying module to suppress initial fire sources remotely and contactlessly; Avoiding secondary safety risks during emergency response.

3.4. Opportunistic Communication and Task Scheduling Requirements

Building obstructions may easily cause wireless communication interruptions. Therefore, the system should support data caching and forwarding under intermittent network conditions. The air-ground units should exchange status information once per second. When communication is lost for more than 5 seconds or the battery level becomes insufficient, the aerial device should automatically return to the ground platform. Path planning should dynamically adjust routes according to real-time communication quality. The system should also support: Natural language-based inspection command interpretation; Time-period-based scheduling of different devices; A visualization platform for synchronizing device locations and hazard images; Automatic generation of standardized inspection reports.

4. System Design

4.1. Overall Architecture Design

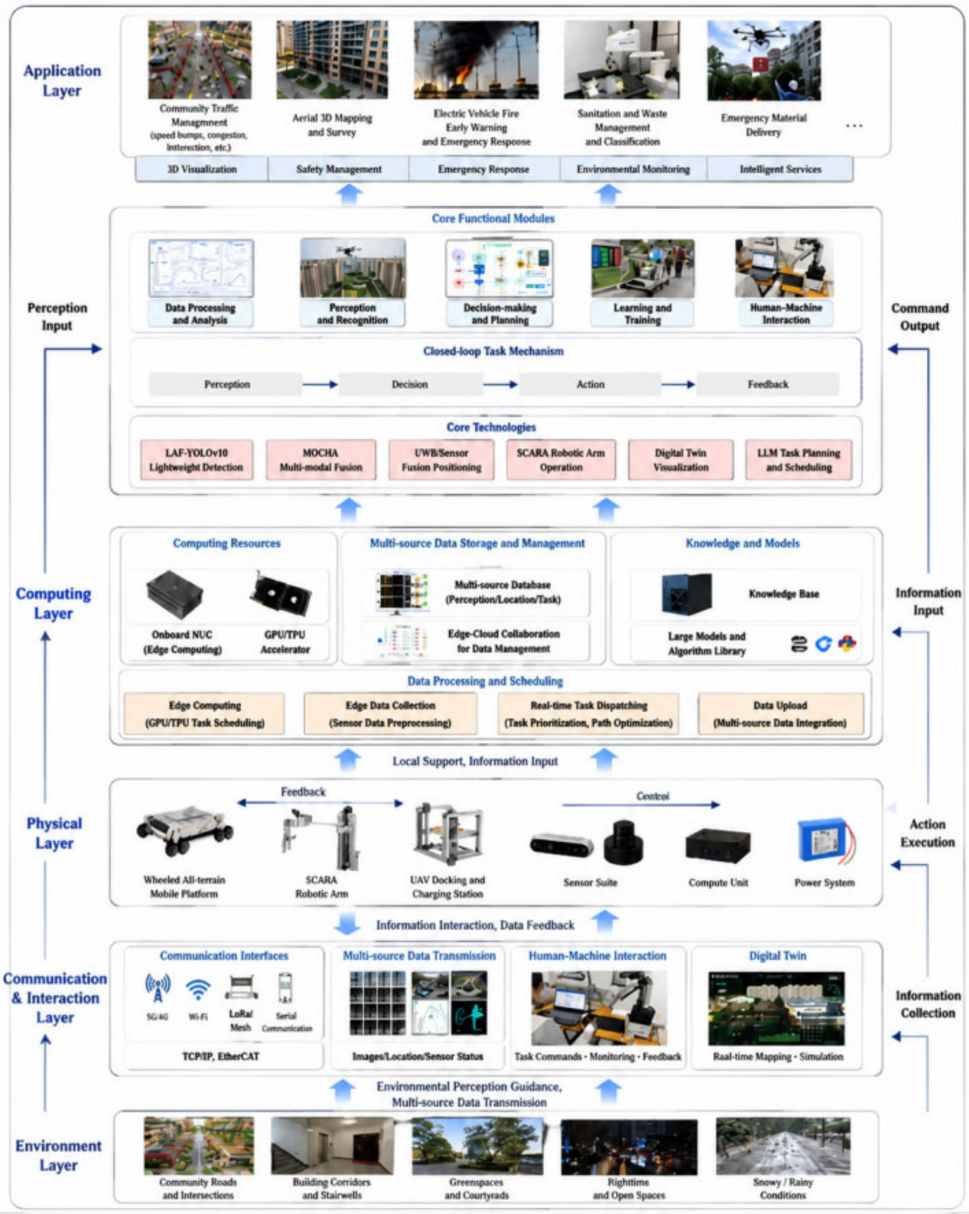


Figure 8. Diagram of the Overall System Architecture

4.2. Robot Hardware Platform and Basic System Construction

The development of the ground robot hardware platform serves as the core carrier for transforming the theoretical scheme of the "Zhikan-Yunxun" system into practical engineering implementation. To intuitively demonstrate the complete engineering composition of this platform, this paper sorts out the full research and development chain from conceptual design to physical prototype. As shown in Figure 9, the left side displays the physical engineering prototype of the platform, presenting the fully assembled and debugged ground robot system consisting of an omnidirectional mobile chassis, a SCARA manipulator, a UAV storage compartment and various sensors. Multiple rounds of functional verification have been completed in real residential community environments. The right side is the hardware exploded view, which clearly illustrates the hierarchical assembly relationship and electrical connection topology among the chassis structure, manipulator joints, battery compartment, computing unit, sensor modules and storage compartment through annotated arrows. This diagram visually reflects the platform’s design philosophy of detachable modules, standardized interfaces and convenient maintenance.

Figure 9. Physical prototype of ground robot platform (left) and hardware exploded view (right)

This design marks that the system has transcended the digital twin stage and possesses physical on-site intervention capabilities. By comparing the two diagrams, readers can clearly understand the corresponding relationship between the physical form and internal structure of the platform, laying a solid engineering foundation for subsequent system integration and application deployment.

4.2.1. High-Traversability Omnidirectional Mobile Chassis

Community road surfaces include paved roads, slippery ceramic tiles, lawns and damaged brick pavements, imposing stringent mobility requirements on the chassis. Inspired by the design concept of hybrid wheel-legged robots, this project adopts a four-wheel independent drive omnidirectional mobile chassis scheme with zero turning radius, enabling in-situ rotation in narrow corridors and unit doorways. The chassis is equipped with high-torque brushless motors and high-friction tires; the vehicle frame is made of aluminum alloy with anti-rust surface treatment, and key compartments reach the IP67 ingress protection rating.

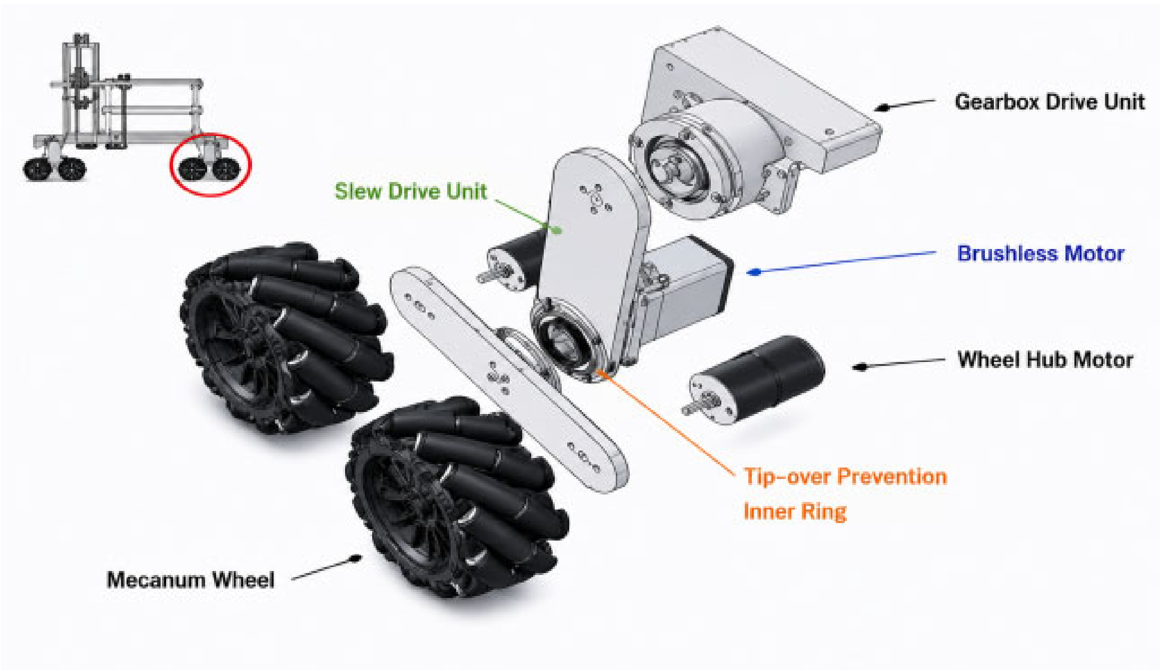


Figure 10. Schematic diagram of the WaLTER hybrid wheel-legged robot

4.2.2. Modular SCARA Manipulator and Dual-Function Operation System

The platform integrates a four-axis manipulator with SCARA (Selective Compliance Assembly Robot Arm) configurationResearchGa..., which features high rigidity along the horizontal axis and superior compliance along the vertical axis. An electromagnetic quick-change end effector is mounted at the tip, supporting rapid switching between two modes: environmental maintenance and UAV inspection & maintenance.

(I) Manipulator Configuration, Kinematics and Autonomous Grasping

The platform integrates a four-axis SCARA manipulator. This configuration exhibits high stiffness in the horizontal direction and compliance in the vertical direction, making it particularly suitable for two typical operations: horizontal grasping and vertical plugging/unplugging. Compared with six-degree-of-freedom articulated manipulators, fewer moving joints deliver higher positioning accuracy and lower control complexity.

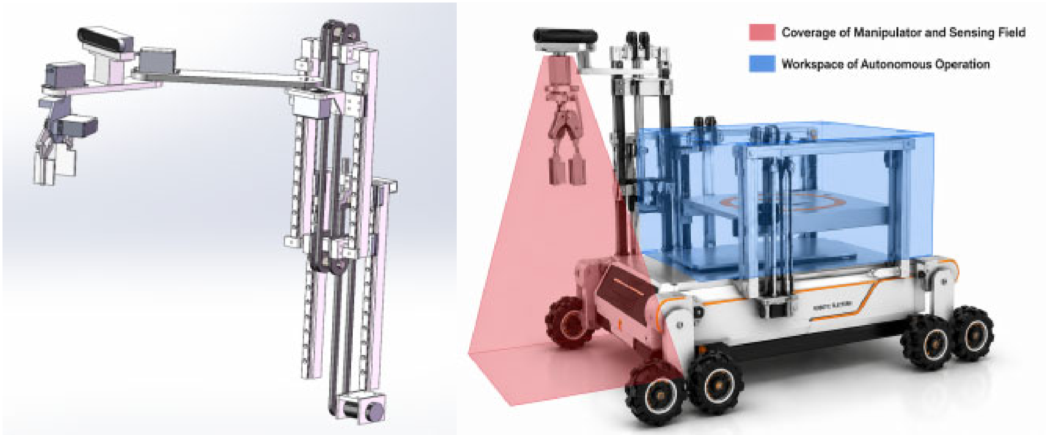


Figure 11. Mechanical structure drawing of SCARA manipulator and schematic diagram of its workspace coverage

Motion control of the manipulator is realized based on inverse kinematics solutionMDPI. Given the target pose of the end effector in three-dimensional space, the system solves the closed-form angular solutions of the four joints via the geometric analytical method, ensuring the manipulator reaches the designated position via the shortest time and optimal path. The optimized workspace naturally covers two core working areas: first, the ground operation area at the side and front of the vehicle body (horizontal coverage radius of approximately 600 mm); second, the UAV storage compartment area at the rear of the vehicle roof. In actual grasping tasks, a depth camera first acquires the 3D point cloud of the target object. After the system estimates the object pose, it invokes the inverse kinematics solver to generate joint trajectories and drive the end gripper to complete autonomous pickup and delivery into the storage bin.

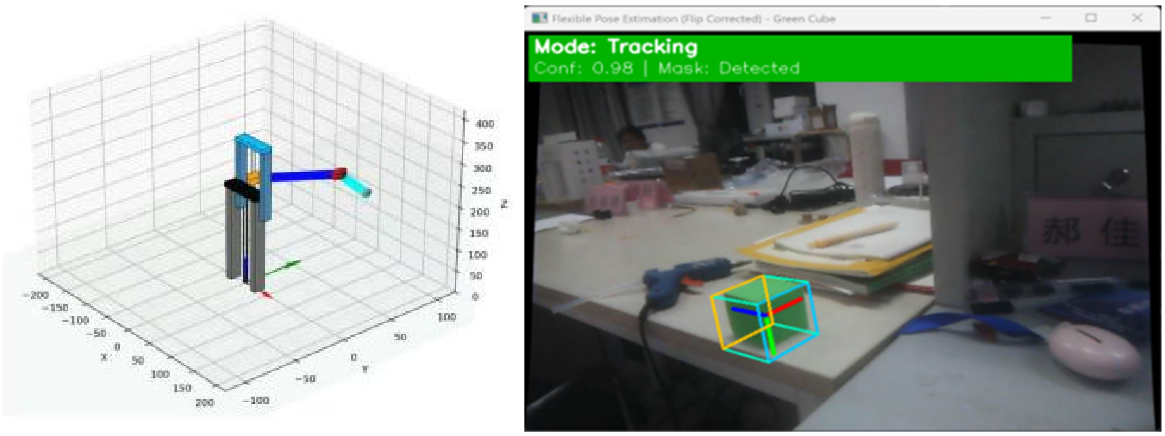


Figure 12. Schematic diagram of joint trajectory planning and object pose estimation diagram

(II) UAV Autonomous Inspection and Foreign Object Removal Tasks

When UAVs shuttle at low altitude between buildings, propellers are prone to winding airborne floating matter such as green belt branches and plastic bags; the fuselage may also accumulate dust and foreign debris, impairing flight safety and perception precision. Leveraging the workspace advantage covering the storage compartment, the SCARA manipulator undertakes autonomous inspection and maintenance tasks after UAV return, including the following operations: ① Propeller foreign object removal: The manipulator end effector is switched to a dedicated soft gripper or miniature vacuum cleaning tool. The vision system first identifies winding debris (such as grass blades, threads and thin cords) on the propeller surface and shaft gaps. The control strategy adopts a two-step mode of "recognition & positioning first, contact operation afterward": the manipulator approaches from the side of the UAV, and the gripper lightly contacts the propeller under low-torque mode to judge contact state via joint torque feedback. For linear winding materials, composite motions of "gripping – rotating – pulling out" are adopted to complete debris removal, avoiding excessive stress applied to propeller blades throughout the process. ② Fuselage surface cleaning and sensor light-transmitting surface maintenance: The manipulator end effector is replaced with a flexible brush or miniature air blowing device to blow or gently wipe UAV arms, landing gears and gimbal camera protective covers. For the light-transmitting surfaces of depth cameras and downward-facing visual sensors, a special non-woven fabric cleaning head is used for fixed-point wiping to guarantee unobstructed sensor field of view and secure perception accuracy for subsequent flight missions.

Figure 13. SCARA manipulator performing UAV inspection and maintenance tasks

4.2.3. UAV Cooperative Storage and Supply Mechanism

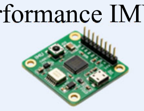
A funnel-shaped UAV storage compartment is integrated at the rear of the vehicle roof. AprilTag visual fiducial markers and wireless charging coils are embedded at the bottom of the compartment. When the UAV lands, the funnel structure guides it into position and locks it automatically; the SCARA manipulator then carries out the above inspection and maintenance procedures immediately afterward.

Figure 14. Schematic diagram of storage compartment structure and UAV

4.2.4. Core Hardware Selection and System Circuit Design

To meet the demands of autonomous inspection and precise operation in complex community environments, the hardware design of this project prioritizes system reliability, compatibility and integrability. The mobile platform achieves high mobility via omnidirectional drive, flexibly coping with narrow corridors, complex turning corners and unstructured pavements. The robot can perceive its own attitude and motion state in real time to maintain stability during navigation and motion control. The tight coordination between the SCARA manipulator and chassis drive system enables the platform to efficiently execute diversified tasks including garbage grasping, UAV maintenance and emergency assistance, realizing the organic integration of mobility, perception capacity and operation capacity, and constructing a high-performance robotic platform oriented to practical applications.

Table 2. Core Hardware Selection List

Hardware Name	Application	Features
Mecanum Wheel 	Used to construct an omnidirectional mobile chassis, achieving zero-radius turning and adapting to complex unstructured terrains.	• Omnidirectional mobility, low noise, and smooth operation • Compatible with various payloads and ground environments • Widely used in robots, AGVs, and other mobile platforms
MG996 Servo Motor 	Applied to the “leg” structure of the wheel-leg hybrid chassis, driving the wheel-leg mechanism and adjusting the chassis posture to achieve active obstacle crossing.	• High torque output (9–15 kg·cm) • Digital control with high precision • Stable operation and strong compatibility
DC Brushless Geared Motor 	Provides high-torque and low-speed stable driving force for the omnidirectional chassis, ensuring strong power output and long endurance during operation on complex terrains.	• High efficiency and low noise • Large torque output and stable operation • Multiple reduction ratios available for flexible control
High-Performance IMU Module 	Enables real-time perception of chassis pose changes and assists in dynamically adjusting leg angles and motor speeds to maintain balance.	• High accuracy and low noise • Compact size and low power consumption • Supports I2C/UART interfaces
Two-Phase Hybrid Stepper Motor 	Used to drive robotic arm joints, ensuring that the manipulator can execute predefined trajectories with high precision and repeatability.	• 1.8° step angle with accurate positioning • Low vibration and low noise • Supports customized windings and parameters

The system takes STM32F407 as the main control core. Unified interfaces realize coordinated interconnection of drive motors, manipulator actuators, attitude sensors and auxiliary control units. An independent power supply architecture is adopted to guarantee stable operation of both the power system and control system, providing reliable hardware support for robot motion control, attitude perception and task execution.

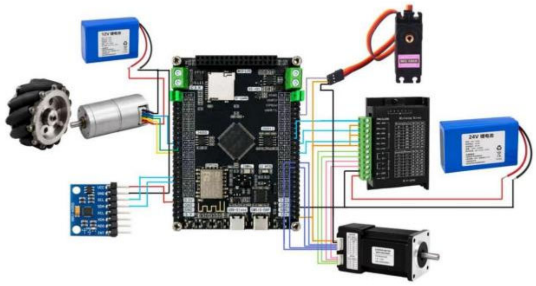


Figure 15. Wiring diagram of core hardware

4.3. Air-Ground Heterogeneous Visual Perception Network Based on LAF-YOLOv10

In the "Zhikan-Yunxun" system, the core of air-ground collaboration lies in constructing an efficient, robust and lightweight visual perception network to address unique challenges under UAV viewpoints, including extreme target scale variation, complex and variable backgrounds, and severely limited computing power on edge devices. To this end, this paper proposes the lightweight and high-efficiency LAF-YOLOv10 object detection algorithm. Through in-depth architectural optimization, the algorithm achieves an optimal trade-off between detection accuracy and inference speed, providing solid technical support for the "last kilometer" of community governance.

4.3.1. Algorithm Architecture Design

The overall architecture of LAF-YOLOv10 follows the one-stage detection paradigm, consisting of three components: backbone network, neck network and detection head. Three core innovations are proposed to realize extreme lightweight performance and high efficiency.

(1) Backbone Network: Lightweight Design Based on Partial Convolution Although the traditional CSPNet module reduces computational cost by splitting feature maps, massive standard convolution operations inside still constitute a computational bottleneck. Therefore, a partial convolution-based C2f (PC-C2f) module is designed as the basic building block of the backbone network. Its core idea is to restrict the number of channels processed by spatial convolution: the input feature map is equally split into two parts. One part is directly transmitted via identity mapping, while the other is fed into a streamlined feature extraction branch. Finally, the two branches are concatenated along the channel dimension. This design drastically cuts redundant feature map generation during computation and reduces memory access overhead.

Specifically, given input tensor $(X \in R^{C\times H\times W})$, PC-C2f partitions its channels into two groups: $(X_p \in R^{C/r\times H=\times W})$ and $(X_u \in R^{(C-C/r)\times H\times W})$, where r denotes the reduction ratio (set to 4). Only (X_p) undergoes standard 3×3 convolution:

$[X_p = \mathrm{Conv3} \times 3(X_p)]$ Subsequently, the processed (X_p) is concatenated with the unprocessed(X_u) , followed by a 1×1 convolution for channel mixing:

$[Y = \mathrm{Conv1} \times 1(\mathrm{Concat}(X_p, X_u))]$ In the formula, C represents the number of input channels, while H and W denote the height and width of the feature map, respectively. Actual tests conducted on the Intel NUC (i7 CPU) edge computing platform demonstrate that compared with the standard C2f module, PC-C2f improves inference speed by nearly 40%, with accuracy loss strictly controlled within 0.5%.

(2) Feature Fusion: Attention-Guided Multi-Scale Enhancement Community scenes captured by UAVs are highly complex. Sudden illumination changes, motion blur and background clutter severely degrade detection performance. Accordingly, an Attention-Guided Feature Pyramid Network (AG-FPN) is constructed as the neck module. The structure embeds the Squeeze-and-Excitation (SE) channel attention mechanism into the top-down path of conventional FPN. For input multi-scale feature maps (F_i), the SE module first compresses spatial information of each channel into a scalar via Global Average Pooling (GAP) to generate channel descriptors. Two fully connected layers are then used to explicitly model interdependencies between channels, ultimately outputting an importance weight (α_i) for each channel:

$[\alpha_i = \sigma(W_2^i \cdot \delta(W_1^i \cdot \mathrm{GAP}(F_i)))]$ Here, $(\mathrm{GAP}(\cdot))$ stands for global average pooling;$(W_1^i \in R^{C/s\times C})$ and $(W_2^i \in R)$ are fully connected projection matrices with squeeze ratio (s=16) ; δ refers to the ReLU activation function, and (σ) denotes the Sigmoid activation function. The weighted feature map is formulated as: $[F_i' = \alpha_i \otimes F_i]$ This enables the network to dynamically enhance feature channels corresponding to targets and suppress noise channels from the background.

Furthermore, the conventional nearest-neighbor upsampler is replaced with the DySample content-aware upsampler. DySample dynamically generates sampling point offsets based on input features and adaptively adjusts the shape of interpolation kernels, rather than adopting fixed upsampling grids. This refinement produces feature maps with sharper edges after fusion, and significantly boosts localization accuracy for tiny objects.

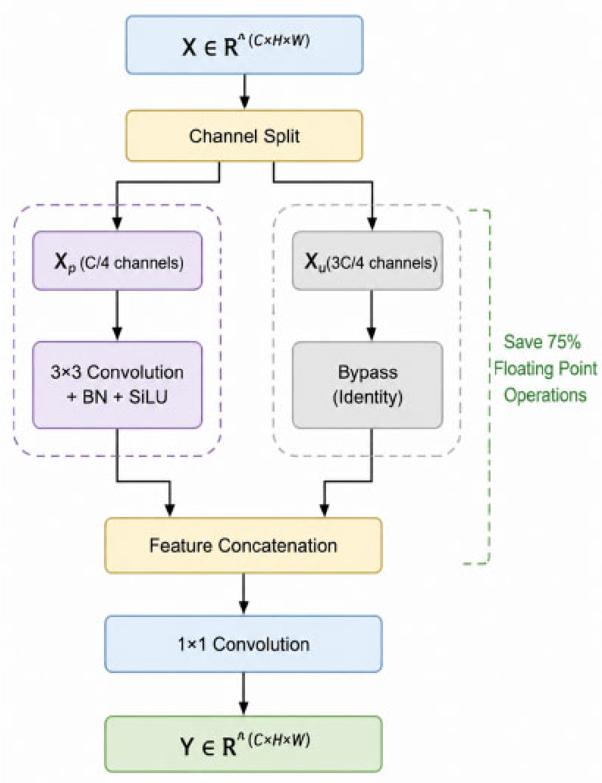


Figure 16. Detailed Diagram of the PC-C2f Module

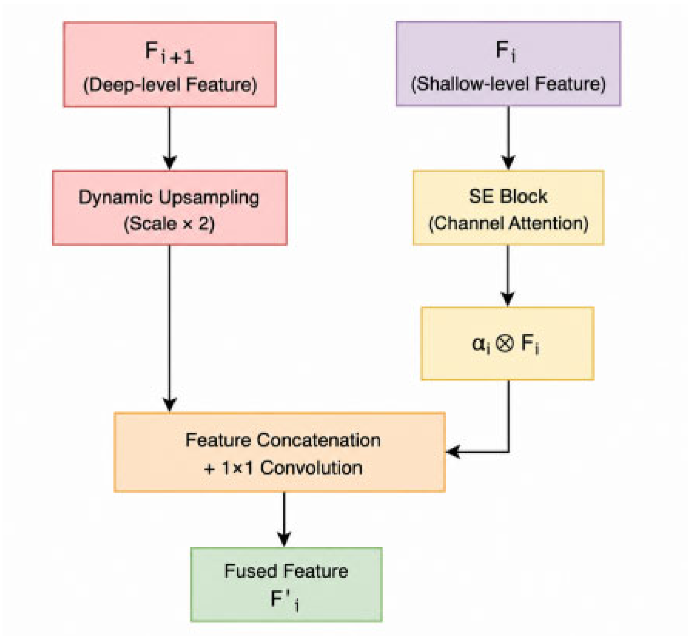


Figure 17. Fusion Procedure Diagram of AG-FPN

(3) Detection Head Reconstruction and Loss Function Improvement A reparameterized detection head is designed to handle the extreme scale distribution of targets in UAV imagery. Experiments reveal that tiny objects in communities such as cigarette butts and bottle caps suffer severe signal attenuation on high-level feature maps after multiple downsampling operations in the backbone network. Therefore, the P5 detection head specially designed for large targets under 32× downsampling is removed to save computational overhead. Meanwhile, an additional P2 detection head is introduced, whose input feature map has a resolution of 160×160, preserving abundant spatial details for detecting tiny targets ranging from 4×4 to 16×16 pixels. This "remove deep head, add shallow head" strategy allocates model parameters and computational resources more efficiently to hard samples.

For loss function optimization, Wise-IoU v3 is adopted to refine bounding box regression. When targets are densely arranged in community scenes, manual annotations often produce low-quality samples with redundant background or inaccurate bounding boxes. Such outliers generate extremely harmful gradients during training. Equipped with a dynamic focusing mechanism, Wise-IoU v3 adaptively assigns appropriate gradient gains to each predicted box according to the matching degree between anchor boxes and ground-truth boxes. Its core loss formula is defined as:

$$[L_{WIoUv3} = r \cdot R_{WIoU} \cdot L_{IoU}] \quad (1)$$

(L_{IoU}) represents standard IoU loss, and (R_{WIoU}) is a consistency penalty term based on distance and aspect ratio. The critical component is the dynamic focusing factor r, calculated as:

$$[r = \left(\frac{\beta}{\delta}\right)^{\gamma - \beta / \delta}] \quad (2)$$

In the formula,(β) denotes an outlier metric measuring the abnormality of anchor boxes; (δ) and (γ) are hyperparameters controlling the mapping relationship (default values: 3.0 and 1.9). When anchor boxes have high matching quality, r approaches 1 for normal model learning. For extremely low-quality anchor boxes, r is drastically compressed to effectively suppress harmful gradients introduced by low-quality annotations, greatly improving localization robustness of the model in dense scenes.

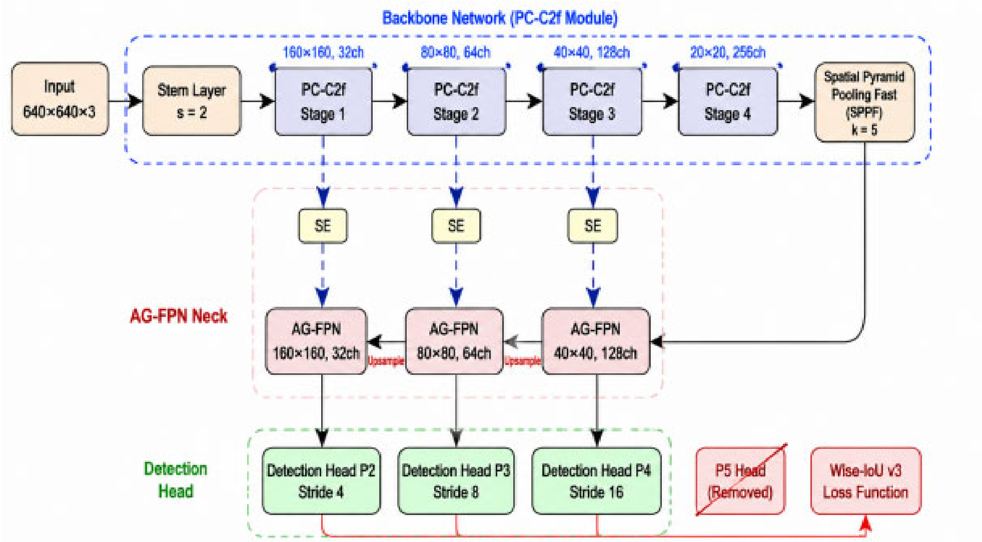


Figure 18. Overall Architecture Diagram of LAF-YOLOv10 Network

4.3.2. Model Lightweighting and Performance Verification

A comprehensive evaluation framework is established to rigorously validate the practical deployment performance of LAF-YOLOv10 on UAV edge devices. Public benchmarks VisDrone-DET2019 and UAVDT are adopted as training datasets, covering typical UAV-view scenes under diverse weather, illumination and flight altitudes. Sufficient training is conducted with Mosaic data augmentation and cosine annealing learning rate decay strategies. During validation, apart from mean Average Precision (mAP), the number of model Parameters and GFLOPs are taken as core metrics to quantify lightweight performance, focusing on the three-way trade-off among detection accuracy, parameter count and inference speed.

First, offline comparative experiments are carried out between LAF-YOLOv10 and mainstream detectors including YOLOv8 and YOLOv10s to verify its advantages in joint optimization of accuracy and efficiency. The accuracy-efficiency trade-off curve clearly demonstrates that the proposed model achieves Pareto optimality: it retains high detection accuracy while drastically reducing computational overhead.

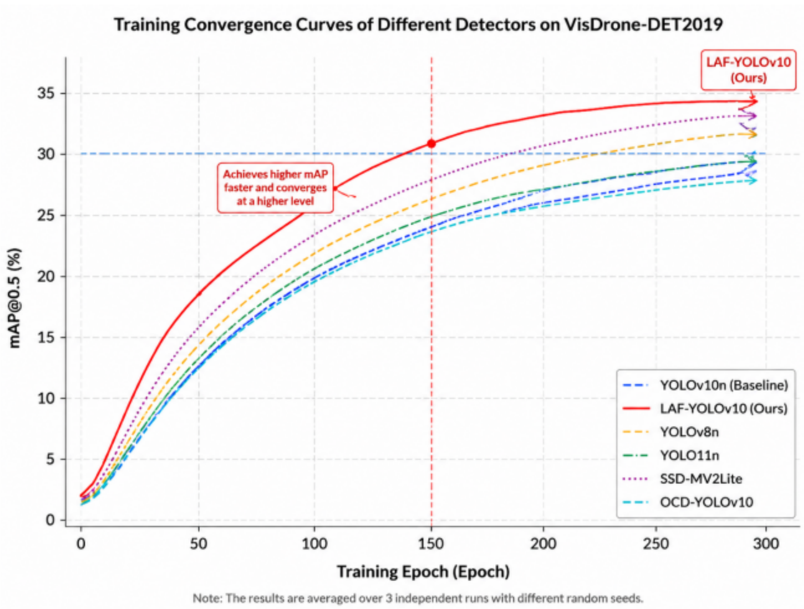


Figure 19. Performance Evaluation Comparison Chart of LAF-YOLOv10

Finally, the INT8-quantized model is deployed on the Intel NUC (i7 CPU) edge computing platform to test real inference frame rates. Deployment test results show that LAF-YOLOv10 achieves an inference speed up to 24.3 FPS under FP16 precision. Compared with the baseline YOLOv10s, the proposed algorithm raises mAP@0.5 by 1.2 percentage points while cutting computational cost by 30%. This result verifies that LAF-YOLOv10 fully satisfies the strict real-time and resource consumption requirements of UAV platforms.

To intuitively demonstrate the robustness of the model in real community environments. As illustrated in Figure 20, LAF-YOLOv10 outputs precise and stable bounding boxes under various challenging conditions, including overexposure, low illumination, mutual target occlusion, and tiny object detection (cigarette butts, bottle caps, etc.), laying a solid foundation for the fully autonomous operation capability of the "Zhikan-Yunxun" system.

Figure 20. Visualization of Detection Results

4.4. Opportunistic Communication and Dynamic Guidance System Based on MOCHA

In the "Zhikan-Yunxun" system, air-ground heterogeneous robots (UAVs and ground vehicles) perform fully autonomous operations in densely built residential communities with severe signal occlusion, confronting communication challenges including non-line-of-sight (NLOS) transmission, high node mobility and intermittent link interruption. Conventional centralized request-response communication architectures rely on stable central nodes, which become extremely fragile in such complex environments. To guarantee reliable transmission of mission data between nodes and support high-level collaborative decision-making, this paper designs and implements a decentralized opportunistic communication and dynamic guidance subsystem built upon the MOCHA framework, whose full name is Multi-robot Opportunistic Communication for Heterogeneous Collaboration.

4.4.1. MOCHA Opportunistic Communication Mechanism

The architectural principle of the MOCHA framework is derived from the decentralized Gossip protocol paradigm, which treats every unmanned robotic node as an equal peer without coordination from a master node. Its core component is a lightweight Header exchange mechanism optimized for real-time dynamic data. Each node periodically or event-triggered broadcasts a "data catalog", namely Header, to adjacent neighboring nodes. The Header only stores the robot ID (RID), topic ID (TID), timestamp and version number of each data record, rather than the complete massive data payload. Its structural layout is shown in Figure 21.

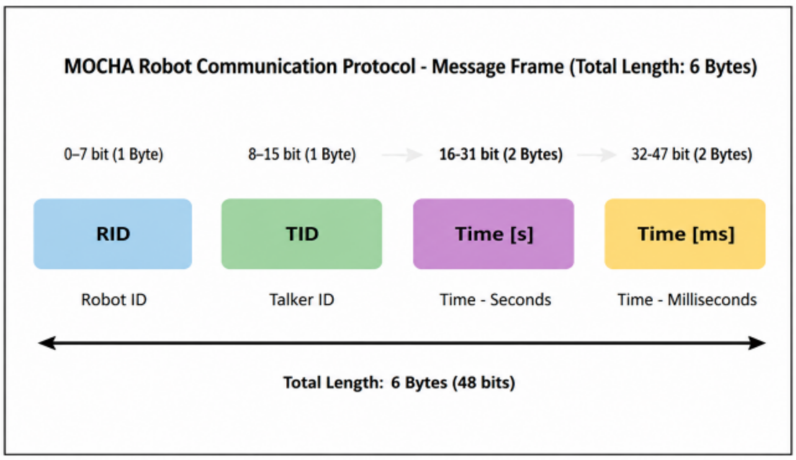


Figure 21. Structural Diagram of MOCHA Message Header

After Node A receives the Header broadcasted by Node B, it instantly identifies all updated data segments owned by Node B through local record comparison. Afterwards, Node A can accurately pull the required data blocks from Node B according to mission priority and instantaneous link quality. This on-demand synchronization mechanism fundamentally eliminates invalid full-data synchronization and effectively mitigates network congestion.

To cope with extreme scenarios featuring fully intermittent connectivity, MOCHA integrates a native Store-Carry-Forward (SCF) routing paradigm. If the target node is temporarily unreachable, the source node temporarily stores messages together with destination addresses in a local persistent queue. When mobile relay nodes (data mules), such as UAVs returning from patrol missions, enter the communication coverage range, queued messages are automatically extracted and forwarded to realize asynchronous data delivery across temporal and spatial intervals. The complete synchronization workflow is illustrated in Figure 22.

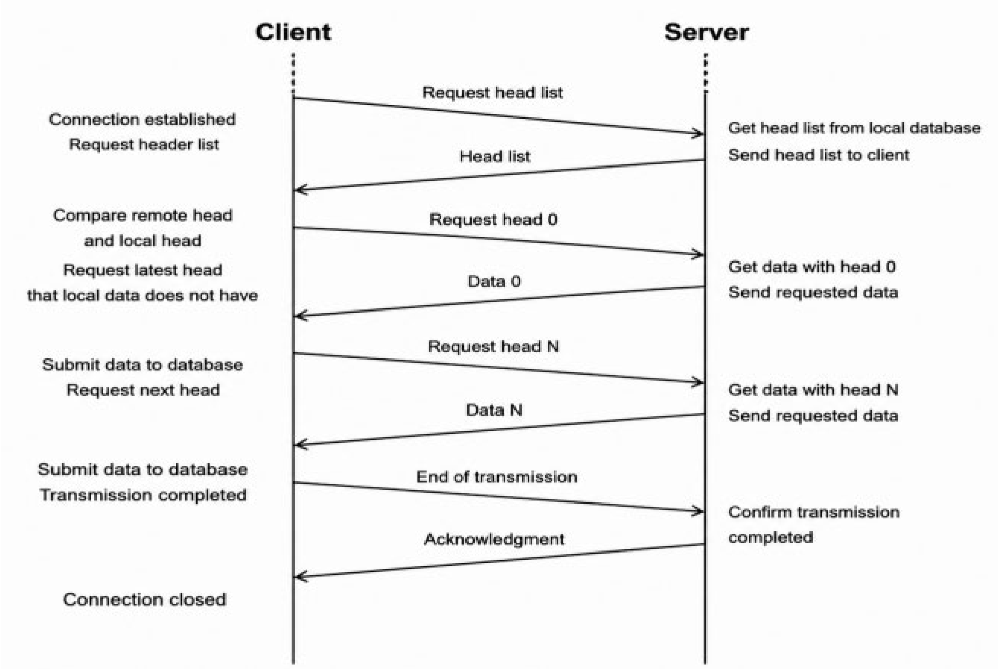


Figure 22. Synchronization Flowchart of MOCHA

For hardware integration, the C++ core library of MOCHA is compiled into ROS 2 nodes, which run natively on the Intel NUC (i7 CPU) edge computing devices carried by UAVs and ground vehicles. By subscribing to and publishing dedicated ROS topics, the framework seamlessly synchronizes critical information among all robots, including visual detection results, pose states and mission status, providing a transparent and robust cross-device data bus for upper-layer collaborative algorithms.

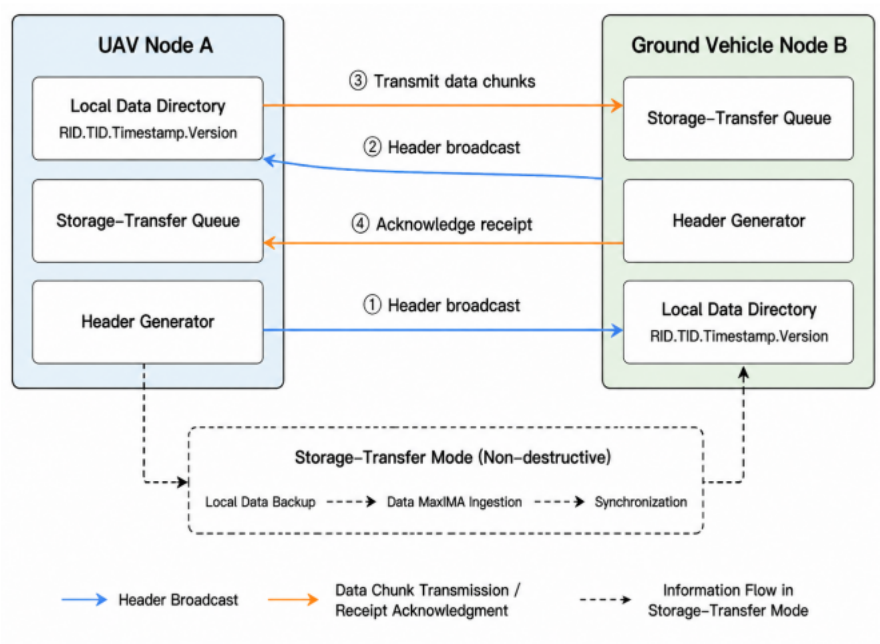


Figure 23. Schematic Diagram of MOCHA Opportunistic Communication Framework

A safety heartbeat mechanism is additionally embedded into the MOCHA communication protocol. UAVs and ground vehicles exchange status information at a fixed frequency of 1 Hz. If the UAV fails to receive heartbeat feedback from the ground vehicle for 5 consecutive seconds, or its residual battery capacity drops below 20%, an emergency return-to-storage procedure is automatically activated. Without waiting for external control commands, the UAV flies back to the visual guidance zone on top of the ground vehicle and lands precisely in the storage compartment guided by downward-facing AprilTag fiducial markers, ensuring absolute operational safety under fault conditions.

4.4.2. Communication-Aware Dynamic Path Planning

Pure geometric path planning algorithms can only avoid physical obstacles but fail to adapt to time-varying electromagnetic channel environments. This paper innovatively proposes an alternating exploration-communication path planning strategy, which incorporates real-time network quality metrics provided by the MOCHA framework, including Round-Trip Time (RTT) and Signal-to-Noise Ratio (SNR), as core components of the navigation cost function.

During mission execution, the ground vehicle continuously monitors the Round-Trip Time (RTT) and Signal-to-Noise Ratio (SNR) of the wireless link connected to the UAV beacon. Communication quality is modeled as a spatial scalar field function(C(x, y)) , whose value depends on the spatial position of the robotic node in the environment. When the ground vehicle enters communication shadow areas formed by buildings or dense vegetation, the value of (C(x, y)) drops sharply.

Therefore, the comprehensive cost function J of the global path planner integrates two cost terms: the conventional geometric path distance cost (D_{path}) and the communication quality penalty cost (C_{com}): $[J = w_1 \cdot D_{path} + w_2 \cdot (1 - C(x,y))]$ where (w_1) and (w_2) denote weighting coefficients that balance the importance of path length and communication quality. The variable $(C(x,y))$ is normalized within the interval[0, 1] , where a value of 1 represents optimal communication link quality and 0 indicates complete signal disconnection.

When the ground vehicle detects that the SNR of the link with the UAV falls below the predefined threshold, the optimization solver automatically generates a sequence of intermediate waypoints, guiding the robot to temporarily deviate from the shortest geometric path and actively navigate toward open areas with superior communication quality to establish stable wireless links. This forms an exploratory collaborative mode of "move a short distance, check link status", which fundamentally guarantees the robustness of the air-ground data link and ensures timely transmission of critical monitoring information and reliable delivery of control commands for the "last kilometer" of community governance.

Figure 24. Scenario Diagram of Communication-Aware Path Planning

4.5. Multi-Sensor Fusion-Based Cooperative Localization and Visual Servo Control

In the "Zhikan-Yunxun" system, realizing accurate following and close-range observation of ground unmanned vehicles by UAVs is a core link to realize the "last kilometer" of community governance. Confronted with complex adverse conditions including building occlusion, illumination fluctuation and dynamic disturbance, a single sensor cannot simultaneously satisfy the dual requirements of robustness and precision for fully autonomous operation. Therefore, this research constructs a hierarchical coarse-to-fine cooperative localization framework, which mainly relies on UWB modules, IMUs and AprilTag visual fiducial markers deployed on ground vehicles (observed by the downward-looking camera of UAVs). On this basis, a highly robust visual servo control loop is designed to realize centimeter-level dynamic following and precise hovering of UAVs relative to ground vehicles.

Meanwhile, to guarantee flight safety and regulatory compliance, a 3D electronic geofence is preconfigured in the onboard flight controller: the X-axis and Y-axis ranges are limited within ±20 meters around the ground vehicle, while the Z-axis (altitude) is constrained between 5 and 15 meters. Once the detected UAV position exceeds the geofence boundary, the flight controller immediately executes a deceleration and pull-back strategy to confine the UAV within the safe operation airspace, eliminating risks such as lost flight or boundary violation causing public disturbance.

4.5.1. Hierarchical Multi-Sensor Fusion Localization Based on UWB/IMU/Vision (UAV Following)

UWB modules, IMUs and downward-looking cameras possess respective strengths and drawbacks in complex community environments. UWB can provide meter-level absolute range constraints but is susceptible to multipath interference under non-line-of-sight (NLOS) conditions; IMUs deliver high-frequency motion state updates yet suffer from uncorrectable cumulative drift; the downward-looking camera equipped on the UAV can detect AprilTag markers mounted on ground vehicles to output high-precision relative poses, but its performance is vulnerable to illumination variation and rapid movement. To overcome the above limitations, the proposed localization system deeply integrates UWB ranging modules, UAV downward-looking cameras and IMUs, and adopts a hierarchical "coarse localization at long range, fine localization at short range" strategy to achieve full-coverage, seamless high-precision pose estimation.

Long-range Navigation Stage (Distance > 20 m) The UAV acquires coarse GNSS coordinates of the ground vehicle via the MOCHA communication framework. In this stage, the UAV mainly relies on IMUs and barometers to conduct Dead Reckoning (DR). Meanwhile, the UWB module carried by the UAV performs two-way ranging with the UWB tag on the ground vehicle to obtain high-precision range constraints. The DR trajectory and UWB ranging measurements are constructed into a Factor Graph, and the iSAM2 optimizer is adopted for incremental graph optimization. This process effectively suppresses the cumulative drift of DR, stably guiding the UAV into the valid visual sensing area within a 20-meter radius of the ground vehicle and laying a foundation for subsequent precise localization.

Short-range Following Stage (Distance < 20 m) After entering the valid visual area, the system seamlessly switches to a tightly coupled fusion localization mode based on the Extended Kalman Filter (EKF). A 12-dimensional system state vector X is defined, containing the position, attitude, linear velocity and angular velocity of the UAV:

$$X = [px, py, pz, \phi, \theta, \psi, vx, vy, vz, \omega x, \omega y, \omega z]^T \quad (3)$$

The prediction step of EKF is driven by acceleration and angular velocity measurements from the IMU at a frequency of 100 Hz, and a discretized kinematic model is utilized for state propagation. In the update step, when the downward-looking camera of the UAV successfully captures the AprilTag deployed on the top of the ground vehicle, the system solves the Perspective-n-Point (PnP) problem to acquire low-frequency (approximately 10 Hz) yet absolutely accurate 6-degree-of-freedom (6-DoF) relative pose observation $Zvision$. This observation vector is correlated with the state vector via the nonlinear observation function $h(\bullet): Zvision = h(X)$

The core update procedure of EKF follows the formulas below:

$$K = P H^T (H P H^T + R)^{-1} \quad (4)$$

$$X \leftarrow X + K(Zvision - h(X)) \quad (5)$$

$$P \leftarrow (I - K H) P \quad (6)$$

Where K denotes the Kalman gain,P represents the state covariance matrix, $H = \partial h / \partial X$ is the Jacobian matrix of the observation function, and R stands for the covariance matrix of visual observation noise. Through this tightly coupled fusion mechanism, high-frequency drift-prone IMU measurements are repeatedly corrected by low-frequency drift-free absolute visual observations, finally outputting stable smoothed centimeter-level localization results at 100 Hz.

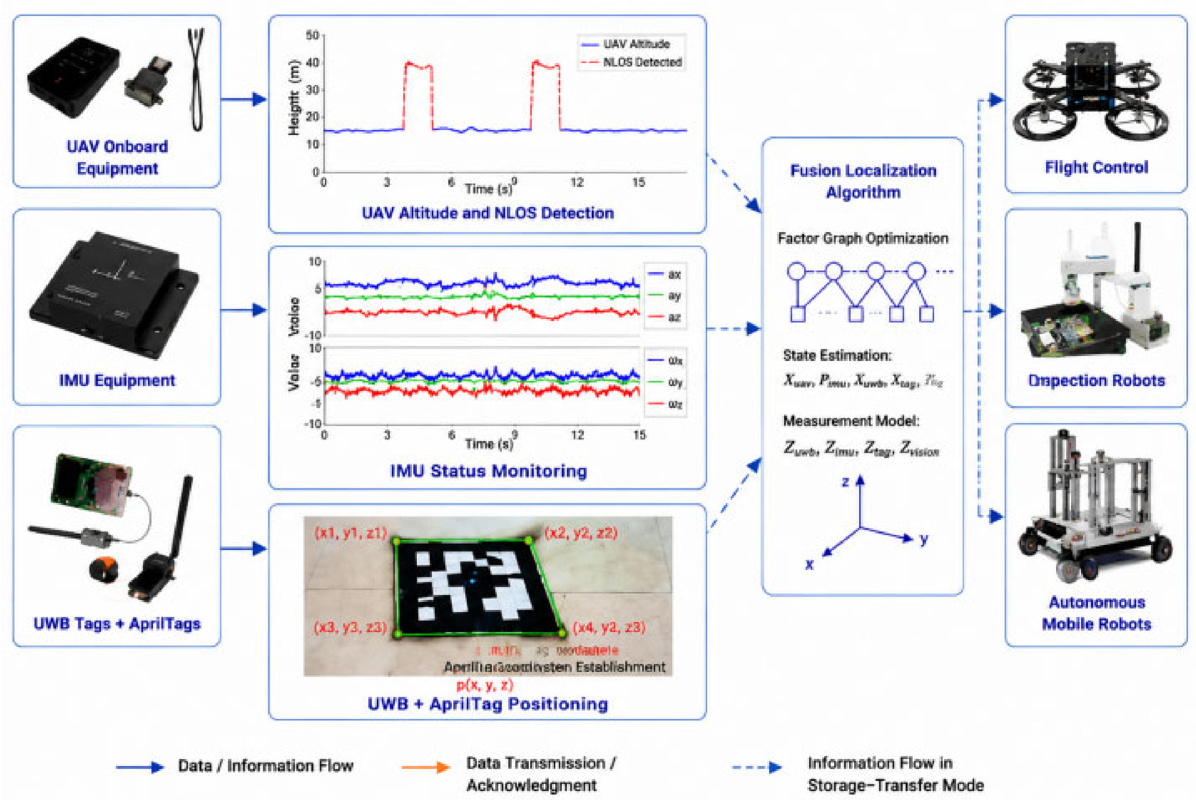


Figure 25. Hierarchical Fusion Localization Architecture Diagram Based on UWB/IMU/Downward-Looking Camera

4.5.2. Image-Based Visual Servoing for Precise UAV Following

Precise localization answers the question of "where the UAV is", while visual servoing directly addresses the control problem of "how to realize stable following". To achieve autonomous precise following and fixed-point hovering of UAVs relative to ground vehicles, an Image-Based Visual Servoing (IBVS) control strategy is adopted. This method directly defines feature errors on the 2D image plane, and drives UAV motion via closed-loop control to minimize such errors, which eliminates the dependence on 3D environmental reconstruction and improves the real-time performance and robustness of the system. The visual feedback of this module mainly relies on the UAV's downward-looking camera.

The core of the following task lies in high-precision relative pose estimation. The downward-looking camera of the UAV captures the AprilTag marker mounted on the top of the ground vehicle in real time, and the pinhole camera model is adopted to calculate the relative depth D from the UAV to the ground vehicle. Assume that the true physical side length of the AprilTag marker is W, the camera focal length is f (unit: pixel), and its projected side length on the image plane is w (unit: pixel). Based on the principle of similar triangles, the depth D is formulated as:

$$D = (f \cdot W) / w \tag{7}$$

After obtaining the depth value D, the horizontal pixel deviation ex (relative to the image center) of the marker's central point is utilized to solve the horizontal relative offset. A classical PID controller is applied for motion control to realize decoupled regulation of the UAV's forward/backward linear velocity v and yaw angular velocity ω .

The control law for forward/backward velocity is designed to adjust the longitudinal distance between the UAV and ground vehicle:

$$v = K_{p,v} \cdot (D - D_{target}) + K_{i,v} \cdot \int (D - D_{target})\, d\tau \tag{8}$$

The control law for yaw angular velocity eliminates horizontal alignment errors:

$$\omega = K_{p,\omega} \cdot e_x + K_{d,\omega} \cdot \dot{e}_x \tag{9}$$

Where D_{target} denotes the expected following standoff distance (generally set to 2–3 m), and $K_{p,v}$, $K_{i,v}$, $K_{p,\omega}$, $K_{d,\omega}$ represent PID gain parameters corresponding to each control channel. Considering attitude fluctuations of the UAV caused by wind disturbance or agile maneuvering, the real angular velocity ω_{imu} fed back by the IMU is introduced for feedforward compensation of the yaw command in each control iteration. The final yaw control command is expressed as $\omega_{cmd} = \omega + \omega_{imu}$. This design constructs a robust closed-loop control framework, which adopts the onboard downward-looking camera to locate the target and leverages IMU inertial measurements to stabilize attitude.

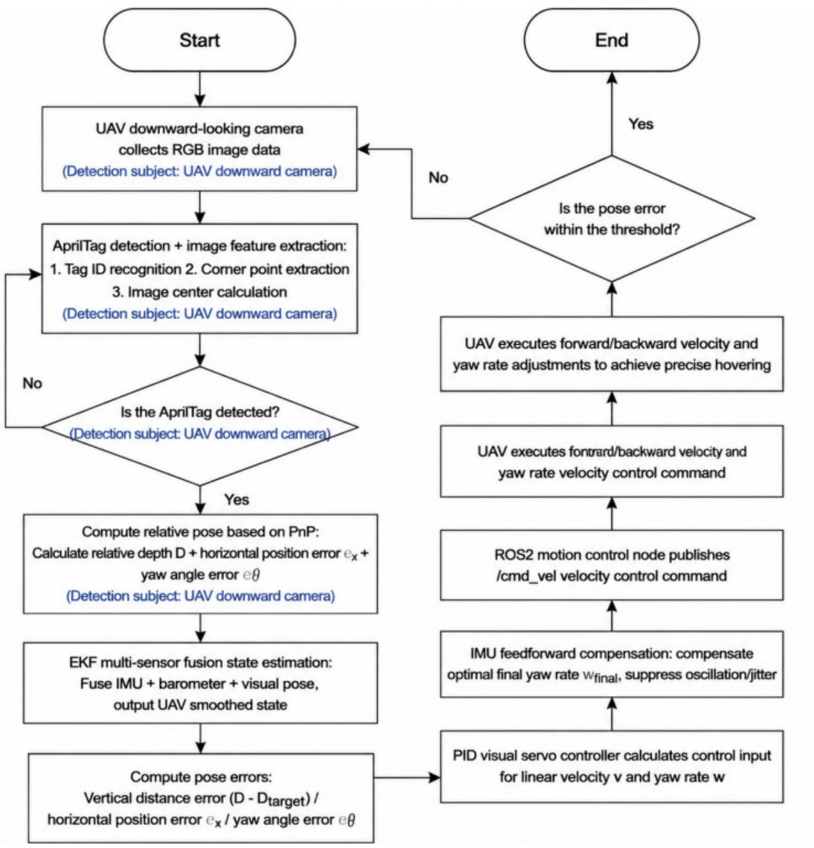


Figure 26. IBVS Control Loop Diagram Based on UAV Downward-Looking Vision and IMU Feedforward Compensation

4.5.3. Multi-Sensor Fusion SLAM Mapping for Autonomous Ground Vehicle Localization

After losing UAV guidance, ground vehicles need to independently perceive their own position and surrounding environments within GNSS-denied corridors between buildings. To solve this problem, this research develops a multi-sensor fusion SLAM system based on RTAB-MAP. Its hardware specifications are listed in Table 3.

Table 3. Hardware Configuration of Ground Vehicle SLAM System

<table>
<tr><td rowspan="2">
</td><td>Application
The Intel NUC (Next Unit of Computing) is a compact high-performance mini PC integrated with CPU, memory, storage and abundant peripheral interfaces. It is widely deployed in scenarios including robot control, edge computing and industrial automation.</td></tr>
<tr><td>Features
Compact form factor with high computing performance, compatible with multiple generations of Intel Core i3 / i5 / i7 processors Low-power architecture supporting stable 24/7 continuous operation Native compatibility with Windows, Linux and real-time operating systems such as Ubuntu and ROS</td></tr>
<tr><td rowspan="2">
</td><td>Application
LiDAR calculates distances to target objects by emitting laser beams and generates high-precision environmental point cloud data. It is extensively applied to robot navigation, SLAM, topographic surveying and autonomous driving.</td></tr>
<tr><td>Features
High-precision ranging within ±2 cm, detection range spanning several meters to hundreds of meters
Supports 2D / 3D scanning with high angular resolution ranging from 0.1° to 0.9°
Lightweight, low-power design for seamless integration onto robotic platforms</td></tr>
<tr><td rowspan="2">
</td><td>Application
The Intel RealSense D435 is a stereo vision depth camera that adopts fused binocular infrared imaging and RGB camera technology. It outputs high-resolution depth maps and color images, and is applicable to robotic vision, gesture recognition, 3D reconstruction and other tasks.</td></tr>
<tr><td>Features
Maximum depth resolution: 1280 × 720 at 30 FPS Global-shutter infrared sensor optimized for capturing dynamic scenes Built-in Vision Processing Unit (VPU) to reduce computing overhead of host equipment</td></tr>
</table>

Equipped with the above LiDAR and depth camera, the ground vehicle implements tightly coupled mapping via the RTAB-MAP framework. The back-end global graph optimization module fuses three categories of sensing information: ① Visual Odometry: The depth camera extracts ORB features to conduct frame-to-frame pose estimation at 30 Hz; ② Depth Point Cloud Registration: Dense local point clouds are generated from depth images, and Iterative Closest Point (ICP) matching is adopted to provide geometric constraints for map alignment; ③ LiDAR Auxiliary Constraint: 16-line LiDAR point clouds are incorporated into the optimization as depth constraints to eliminate drift of visual odometry in long straight corridors and low-texture wall surfaces.

IMU outputs measurements at 200 Hz for point cloud distortion removal and motion compensation. The back-end maintains global map consistency via Bundle Adjustment (BA) and loop closure detection, outputting 6-DoF poses and dense point cloud maps.

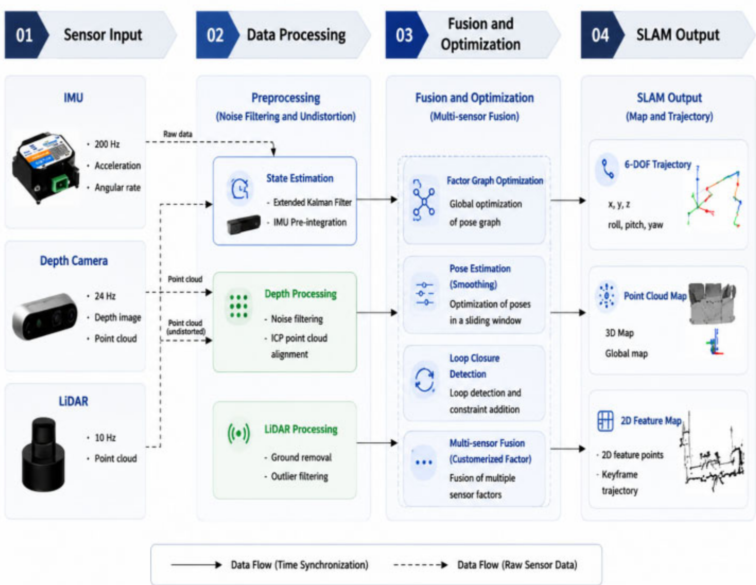
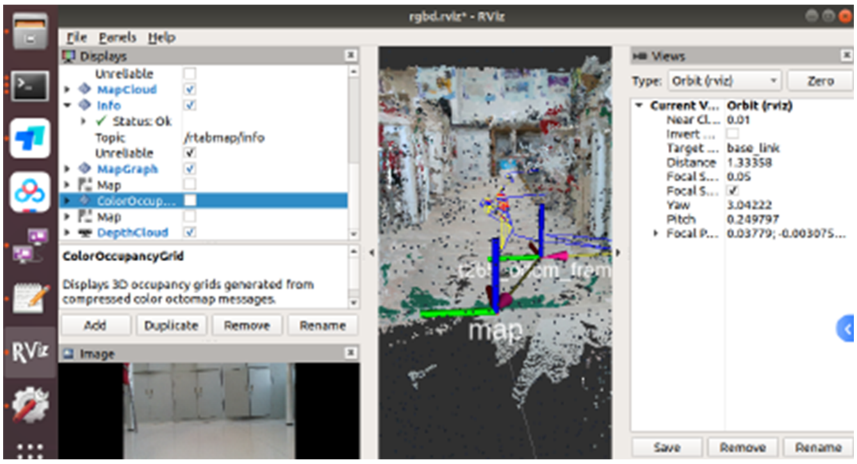


Figure 27. Multi-Sensor Fusion SLAM Architecture Diagram

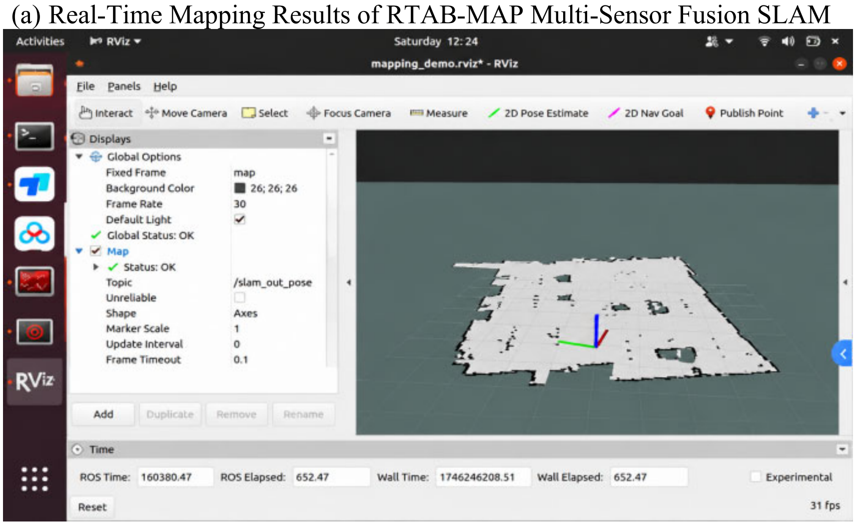
4.5.4. Local Path Planning and Navigation

Based on the 2.5D cost map constructed by SLAM (a 2D occupancy grid map projected from 3D point clouds), the Timed Elastic Band (TEB) algorithm is adopted for local path planning. This algorithm models the robot pose sequence, control commands and obstacle constraints as a hypergraph optimization problem, supports kinematic models of differential-drive and omnidirectional chassis, and can generate smooth collision-free local trajectories in narrow building corridors.

As illustrated in Figure 28, gray areas represent the SLAM-built 2D occupancy grid map, black pixels denote static obstacles (walls, green belt boundaries), and blue coordinate axes indicate the current robot pose (ROS Topic: /slam_out_pose). When temporary obstacles such as illegally parked electric vehicles are detected, the TEB algorithm can replan trajectories within millisecond scale to generate detour paths and return to the original global route.



(a) Real-Time Mapping Results of RTAB-MAP Multi-Sensor Fusion SLAM



(b) 2D Occupancy Grid Map and Robot Pose Visualization

Figure 28. Visualization of Local Path Planning and Navigation Environment

Subgraph (a) of Figure 28 demonstrates the colored dense point cloud map and 3D occupancy grid constructed in real time by the RTAB-MAP framework within building corridors. RGB coordinate axes mark the real-time robot pose, presenting the mapping effect fused by visual odometry, depth point cloud registration and LiDAR auxiliary constraints. Subgraph (b) shows the corresponding projected 2D occupancy grid map, which serves as the underlying environmental model of the TEB algorithm and directly supports local path planning and real-time navigation.

4.6. Task Planning and Digital Twin Visualization

To realize efficient scheduling of multi-robot collaborative tasks and algorithm verification in complex community environments, the "Zhikan-Yunxun" system introduces a large language model (LLM)-driven embodied AI decision-making hub, and constructs a high-fidelity general-purpose digital twin platform.

4.6.1. LLM-Driven Task Comprehension and Decomposition

The system integrates lightweight large language models (GPT-4o-mini or LLaMA 3.1 8B) as offline auxiliary tools for task planning. When administrators issue natural language instructions, the LLM automatically parses task objectives and decomposes them into a sequence of subtasks, outputting task schemes for manual confirmation or modification. This module does not participate in real-time scheduling; after the generated scheme is confirmed, it is executed by an embedded Finite State Machine (FSM).

When administrators input natural language instructions, the system automatically judges whether the current time falls within the permitted operation window during task decomposition. If the current time is outside the allowed window, all UAV-related tasks will be suspended, or only ground robots will be activated for operation, with a feedback prompt: "The UAV is in silent hours; ground vehicle patrol has been scheduled". If the target task can be completed solely by the ground robot, the task will be executed normally. This design reflects the system's human-machine friendly design philosophy that fully respects residents' daily rest schedules.

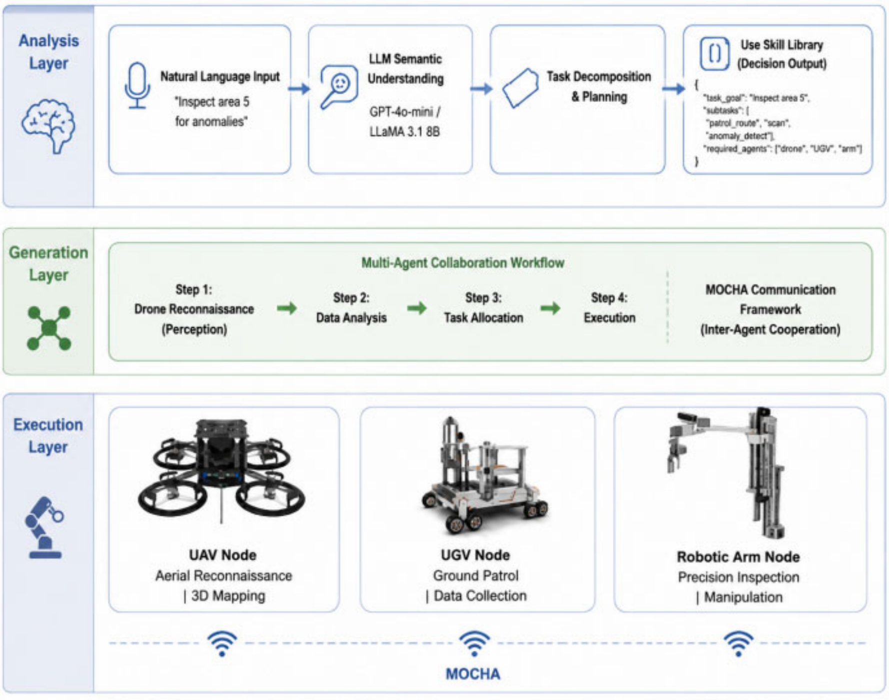


Figure 29. Flowchart of LLM Task Parsing and Instruction Generation

4.6.2. Construction of Digital Twin Simulation Platform

This system develops a digital twin visualization front-end oriented to community governance, whose core functions include real-time state mapping and operational monitoring. Based on 3D reconstructed maps of typical residential communities such as Xiangfuli Community in Tangshan City, the platform establishes geometric and semantic models of buildings, roads and vegetation. Real-time data including pose information, sensor measurements and task status of UAVs and ground robots are received via the MOCHA communication framework. Key indicators such as robot motion trajectories, detected target coordinates and communication link quality are synchronously rendered in the 3D virtual scene. The platform supports historical task playback and multi-view switching, providing administrators with full-domain situational awareness to assist manual supervision and decision-making for abnormal event intervention.

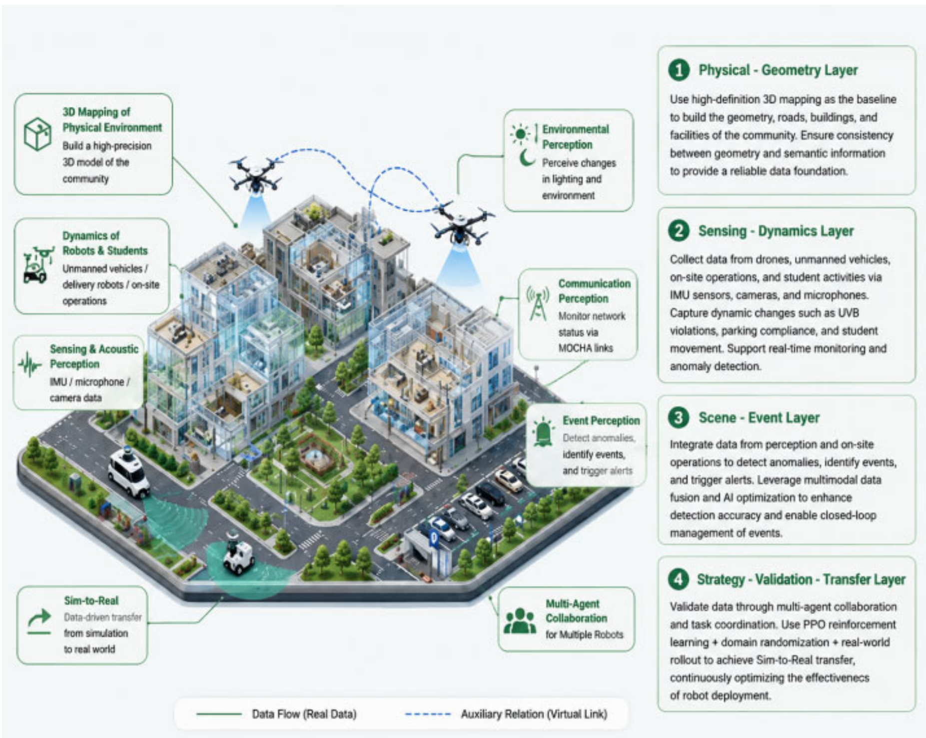


Figure 30. Architecture Diagram of Digital Twin Environment

5. Experimental Verification

5.1. Experimental Scheme Design

5.1.1. Experimental Site and Prototype Parameters

To quantitatively verify whether each performance indicator of the "Zhikan-Yunxun" air-ground collaborative system meets the engineering constraints for old residential community governance, a typical old residential complex, Xiangfuli Community in Tangshan City, is selected as the field test site. The total area of the community is 10000 square meters, covering typical unstructured working conditions including 15 cm-high steps, muddy potholed roads, snow-covered pavements, dense vegetation occlusion, and GNSS-denied urban canyons formed by high-rise buildings. Meanwhile, an indoor simulation testbed with adjustable illumination and communication occlusion is built for offline comparative testing of algorithms. The hardware configuration and algorithm architecture of the full experimental prototype are completely consistent with the system design described in Chapter 4, without any hardware trimming or algorithm simplification.

5.1.2. Control Group Setup

Two types of mainstream traditional inspection equipment are set as parallel control samples. Each group of experiments is repeated 20 times, and the arithmetic mean value is adopted to eliminate experimental errors caused by random disturbances. Control Group 1 is a conventional two-wheeled ground inspection robot without active obstacle-crossing mechanisms, and no onboard UAV or manipulator for field operation. Control Group 2 is an independent aerial photography UAV, which is not equipped with a ground carrying chassis or physical operation units. The evaluation system of this experiment covers five core dimensions: terrain traversability, object detection performance, localization accuracy under GNSS-denied environments, full-area inspection efficiency, and operation & maintenance (O&M) cost. Since the independent UAV has no supporting ground hardware, its O&M cost cannot be calculated and thus will not be included in the comparative analysis.

5.2. Mobile Chassis and Complete Machine Operation & Maintenance Performance Test

Continuous traversal tests are carried out under four obstacle conditions: steps, muddy roads, snow-covered pavements, and narrow 90° corridors. The experimental results show that conventional wheeled robots can only pass through narrow corridors with limited accessibility, achieving an overall traversal success rate of merely 8.75%. The independent aerial UAV has no ground mobility, so all ground traversal indicators are invalid. The wheel-legged composite chassis proposed in this paper combines attitude adjustment mechanisms and Mecanum wheels, achieving a 100% traversal success rate for all four road conditions. It can stably surmount obstacles up to 15 cm in height, operates without slipping on slippery surfaces, and realizes zero-radius turning in narrow passages, satisfying the unstructured terrain traversal index requirements specified in Section 3.2.

The closed-loop full-machine inspection test results indicate that the traditional manual patrol plus fixed surveillance mode requires 125 minutes for a single full-area inspection, with 3 on-duty staff assigned and an annual comprehensive O&M cost of 210,000 RMB. This mode can only record images of hidden dangers without any on-site intervention capability. The conventional wheeled robot takes 92 minutes for a single inspection with one maintenance staff, with an annual O&M cost of 140,000 RMB. However, it cannot reach most concealed hazard locations due to terrain limitations. The integrated air-ground collaborative system proposed in this paper only consumes 18 minutes for a complete full-community inspection, requiring only one staff member for remote monitoring, with an annual O&M cost of 80,000 RMB. The supporting SCARA manipulator can perform physical disposal tasks including picking up road debris, suppressing aerosol combustion at incipient fire scenes, and intercepting electric vehicles entering elevator shafts. The statistical standard for annual O&M costs in this paper uniformly covers all expenditures including staff salaries, equipment depreciation, and routine inspection & maintenance.

5.3. Experiment on Lightweight Detection Algorithm LAF-YOLOv10

A test dataset containing 1200 aerial images captured in real residential community scenes is constructed. The samples cover multi-scale hidden hazards such as cigarette butts, waste paper scraps, illegally parked electric vehicles, debris stacked in corridors, and incipient fires. Comparative edge-device inference experiments are conducted between the proposed algorithm and mainstream lightweight detectors YOLOv11n and OSD-YOLOv10, and the quantitative evaluation metrics are listed in Table 4.

Table 4. Comparison of Performance Metrics of Lightweight Detection Models

Model	Parameters / M	Inference Frame Rate / FPS	mAP@0.5	Miss Detection Rate of Tiny Objects
YOLOv11n	4.1	17.2	30.9%	12.7%
OSD-YOLOv10	3.5	19.5	32.1%	9.4%
LAF-YOLOv10 (Proposed)	2.3	24.3	33.3%	4.8%

5.4. Multi-Sensor Fusion Localization Accuracy Experiment

A 20-meter straight-line walking experiment is carried out in alleys where GNSS signals are completely blocked by high-rise buildings. The average trajectory errors of three schemes are collected separately, namely pure IMU, standalone UWB, and UWB+IMU+AprilTag fusion. The test data are listed in Table 5.

Table 5. Average Localization Error of Different Localization Schemes for 20 m Traversal

Localization Scheme	Average Localization Error (20 m Travel)
Dead Reckoning with IMU Only	1.23 m
Standalone UWB Ranging Localization	0.38 m
UWB/IMU Vision Hierarchical Fusion	0.028 m

5.5. Reliability Experiment of MOCHA Opportunistic Communication

A non-line-of-sight (NLOS) communication environment with dual occlusion of buildings and trees is constructed, where image data packets and task control commands are cyclically transmitted for a continuous duration of 5 minutes. Conventional point-to-point WiFi communication suffers severe interference from occlusion, with a packet loss rate of 31.7%. After adopting the MOCHA store-and-forward opportunistic communication framework, the data packet loss rate is reduced to 1.2%. The system adopts a 1 Hz heartbeat interaction mechanism. If the heartbeat signal of the counterpart device is not received for 5 consecutive seconds, or the remaining battery power of the UAV is lower than 20%, the emergency return logic will be automatically activated. When the communication link is disconnected, pending transmission data is cached locally, and asynchronous automatic forwarding is executed after the link recovers, which effectively solves the problem of intermittent communication disconnection in occluded environments of old residential communities.

5.6. Comprehensive Experimental Analysis

1.Mobile Operation Platform: The wheel-legged composite chassis achieves stable traversal on all types of road conditions. The modular SCARA manipulator supports various types of physical on-site intervention, overcoming the drawbacks of traditional inspection equipment, namely limited terrain adaptability and lack of on-site disposal functions.

2.Visual Perception: The lightweight LAF-YOLOv10 network realizes a balanced trade-off between detection accuracy and computing overhead. Its recognition performance for tiny hidden hazards is remarkably superior to other similar lightweight detection models, satisfying the real-time perception constraints of onboard equipment.

3.Localization and Communication: The UWB/IMU/vision fusion localization method achieves centimeter-level high-precision following in GNSS-denied areas. By adopting the store-and-forward mechanism, MOCHA opportunistic communication drastically reduces the packet loss rate under occluded scenarios, and the stability of the air-ground collaborative data link is greatly improved.

4.Engineering Deployment: The integrated air-ground autonomous inspection drastically shortens the full-area patrol duration and cuts labor and equipment operation & maintenance costs. A complete closed-loop community governance system is formed, consisting of hidden hazard perception, air-ground collaborative navigation and physical on-site disposal. All performance indicators of the system satisfy the engineering constraints proposed in Chapter 3, and the system is qualified for large-scale practical deployment in old residential communities.

6. Conclusion

This paper develops the "Zhikan-Yunxun" air-ground collaborative community governance system. Aiming at three prominent practical bottlenecks in old residential communities, namely surveillance blind zones, poor terrain traversability and discontinuous emergency response, an integrated collaborative architecture integrating aerial blind-area supplement and ground on-site operation is constructed. Field real-scene tests verify that all core performance indicators meet design requirements, and the overall scheme possesses practical engineering deployment value.

In terms of hardware, the wheel-legged composite chassis adapts to complex road conditions including steps and potholed surfaces. An integrated storage compartment is mounted on the vehicle roof to realize autonomous takeoff, landing and charging of UAVs. The matched SCARA manipulator can execute waste cleaning and incipient fire intervention, eliminating the critical drawback of traditional inspection equipment that only supports monitoring without physical on-site disposal.

From the perspective of algorithms and perception, the lightweight LAF-YOLOv10 network achieves accurate identification of tiny hidden hazards in residential areas. The MOCHA opportunistic communication framework mitigates signal occlusion interference caused by high-rise buildings. The UWB/IMU/vision fusion localization algorithm realizes centimeter-level UAV following performance under GNSS-denied environments. Equipped with lightweight large language models and a digital twin platform, the system supports natural language-based task scheduling and full-domain situational visualization monitoring.

The proposed system forms a complete closed-loop governance chain of "Perception – Communication – Navigation – Disposal", upgrading the conventional passive patrol mode to active air-ground intelligent operation and maintenance. The contributions of this work can be summarized in four dimensions: theoretically, it improves the methodology of multi-robot collaborative operation in community scenarios; technically, it breaks through technical difficulties including localization in GNSS-denied environments and edge-side lightweight object detection; socially, it reduces the manpower burden of grassroots community staff; economically, it requires no reconstruction of existing community infrastructure and significantly cuts long-term operation and maintenance costs, providing a feasible technical solution for empowering smart communities via low-altitude economy.

Future research directions include optimizing the environmental generalization capability of the object detection model, investigating multi-robot swarm collaborative strategies, expanding emergency operation functions of the manipulator, and gradually establishing a full-coverage long-term intelligent operation and maintenance system.

References

[1] Ministry of Industry and Information Technology of China. (2025). White paper on China's low-altitude economy (2025). Ministry of Industry and Information Technology of China.

[2] State Council of the People's Republic of China, & Central Military Commission of the People's Republic of China. (2023, May 31). Interim measures for the administration of unmanned aircraft flight: Order No.761 of the State Council and the Central Military Commission.

[3] State Council of the People's Republic of China. (2022). The 14th five-year plan for national emergency response system.

[4] Ministry of Industry and Information Technology of China, Ministry of Science and Technology of China, Ministry of Finance of China, et al. (2024, March 27). Implementation plan for innovative application of general aviation equipment (2024-2030): MIIT Lianchongzhuang〔2024〕No.52.

[5] People's Government of Hebei Province. (2025). Implementation rules for innovative development of low-altitude economy in Hebei Province.

[6] People's Government of Tangshan City. (2025). Implementation plan for smart community construction of Tangshan City.

[7] Zhang, L., & Wang, J. (2024). Research on security shortcomings of old residential communities and UAV inspection application. Urban Management and Science & Technology, 26(2), 45-49.

[8] Liu, H., & Chen, Y. (2024). Review on GNSS positioning failure and multi-source positioning technology under urban canyon environment. Robot Technology and Application, (3), 34-41.

[9] Wang, A., Chen, H., Liu, L., et al. (2024, October 30). YOLOv10: Real-time end-to-end object detection. https://arxiv.org/pdf/2405.14458

[10] Research Group of This Project. (2026). Field investigation report on Xiangfuli and Peirenli communities in Tangshan city [Internal research report].

[11] Gao, W. J., & Zhou, M. (2025). Research on multi-source fusion SLAM of lidar-UWB-IMU based on ROS. Computer Engineering and Applications, 61(8), 231-239.

[12] Hu, C. X. (2022). ROS robot development practice. China Machine Press.

[13] Zhang, Y. (2023). Research on store-carry-forward routing mechanism in opportunistic networks [Doctoral dissertation]. Xidian University.

[14] Li, X., & Ma, Y. (2025). Air-ground cooperative opportunistic communication relay technology in complex urban environment. Telecommunication Engineering, 65(4), 512-519.

[15] Chen, M., & Zhao, B. (2025). Research on autonomous decision-making of smart community robot empowered by large model. Computer Simulation, 42(5), 489-494.