

Research on Autonomous UAV Search and Rescue System Based on Depth Camera and Infrared Thermal Imaging

Yuting Zhang¹, Minxiang Liu¹, Jiachang Li², Yiquan Yang¹, Hao Li¹

1. School of Electronic Information Engineering, Changchun University, Changchun, CHINA

2. School of Computer Science and Technology, Changchun University, Changchun, CHINA

Corresponding Author: Yuting Zhang

ABSTRACT: Addressing the limitations of manual search and rescue operations at high-risk disaster sites, such as low efficiency, severe personal safety hazards, and the extreme difficulty of ensuring comprehensive coverage due to heavy smoke, darkness, and complex structural debris occlusions, this study designs and implements an autonomous UAV search and rescue system based on depth-infrared dual-modal perception technology. Utilizing a high-performance microcontroller combined with an embedded edge computing platform, the system captures environmental 3D point cloud data via a depth camera to realize active obstacle avoidance and centimeter-level precise localization within unstructured environments. Concurrently, it deeply integrates the human thermal radiation characteristics extracted from an infrared thermal imager to achieve highly reliable life detection under extreme low-visibility conditions. Experimental results indicate that the system achieves centimeter-level, anti-drift spatial localization precision (± 5 cm) and ultra-low control latency in GPS-denied environments. By deploying an improved regional proposal detection algorithm optimized with a Squeeze-and-Excitation (SE) channel attention mechanism and a Feature Pyramid Network (FPN), the system yields a mean Average Precision (mAP50) of 91.7% for trapped human target recognition in simulated disaster scenarios. This represents a 4.7-fold increase in sweeping and searching efficiency compared with traditional manual methods, offering a highly automated, cost-effective, and fully autonomous low-altitude technological solution for emergency rescue applications.

Date of Submission: 09-10-2026

Date of acceptance: 20-10-2026

I. INTRODUCTION

Natural disasters, sudden public safety incidents, and industrial accidents in hazardous zones represent catastrophic global threats to human life and socio-economic stability. At high-risk disaster sites, such as earthquake ruins, forest fire zones, and structural collapses, rescue operations face immense hazards. Traditional manual searching is heavily limited by secondary cave-ins, toxic airborne particulates, or complete darkness, introducing severe safety risks to emergency personnel. Furthermore, because of the chaotic environment, it is nearly impossible for human rescue teams to achieve rapid, comprehensive sweeping coverage within the critical "golden window." Guided by national strategic support for the "low-altitude economy" and smart emergency equipment, utilizing autonomous Unmanned Aerial Vehicles (UAVs) to enter hazardous spaces in advance can thoroughly safeguard front-line responders and minimize emergency reaction times.

However, existing commercial UAV search platforms exhibit two profound engineering bottlenecks in practical field deployment. First, conventional drone flight navigation relies heavily on continuous external global positioning system (GPS) signals and constant line-of-sight manual remote control. In indoor enclosed structures, collapsed spaces, or disaster centers suffering from dense electromagnetic interference, GPS streams are frequently blocked or attenuated, causing significant positioning drift that prevents autonomous flight planning. Second, standard visible-light cameras become completely ineffective at night or in smoke-filled, highly occluded landscapes. They fail to extract human targets that are partially buried under rubble, resulting in high rates of false positives and fatal omissions.

In recent years, the rapid evolution of multi-sensor data fusion, Simultaneous Localization and Mapping (SLAM), and embedded lightweight deep learning has offered promising pathways to address these vulnerabilities. Prior research has actively explored complex robotic perception. Q. Zhuang et al. [1] investigated the inherent physical limitations of isolated sensor spectrums and presented spatial-alignment and feature-level fusion strategies to enhance the robustness of robotic perception in extreme configurations. Several groups developed infrared thermal radiation detection networks [2] to overcome illumination constraints by utilizing temperature differences between survivors and backgrounds, but they lacked spatial 3D geometry information and precise indoor positioning capability. On the algorithm side, lightweight architectures such as optimized YOLO models or proposal-based Faster R-CNN networks [3]-[5] have been widely applied in precision agriculture and industrial infrastructure inspections. Nevertheless, balancing high-throughput 3D point

cloud processing, real-time dual-modal spatiotemporal registration, and low-latency flight stability on resource-constrained microcontrollers (e.g., STM32F411) remains a critical, unresolved engineering challenge.

To mitigate these limitations, this study introduces two core breakthroughs: First, the design of a dual-modal 3D perception and lightweight intelligent life detection network. By synchronizing depth point clouds and thermal images at the hardware level, we build an optimized Faster R-CNN coupled with a Feature Pyramid Network (FPN) and an attention mechanism. This architecture effectively extracts multi-scale geometric and thermal details, significantly reducing target omissions in cluttered debris. Second, the implementation of a robust SLAM navigation framework and a fully integrated hardware-software architecture for GPS-denied environments. A high-frequency attitude estimation and multi-sensor positioning fusion scheme are executed directly on an onboard microcontroller, ensuring centimeter-level hovering stability and active obstacle avoidance without external coordinates. This bridges the engineering loop from multi-modal perception to real-time flight control.

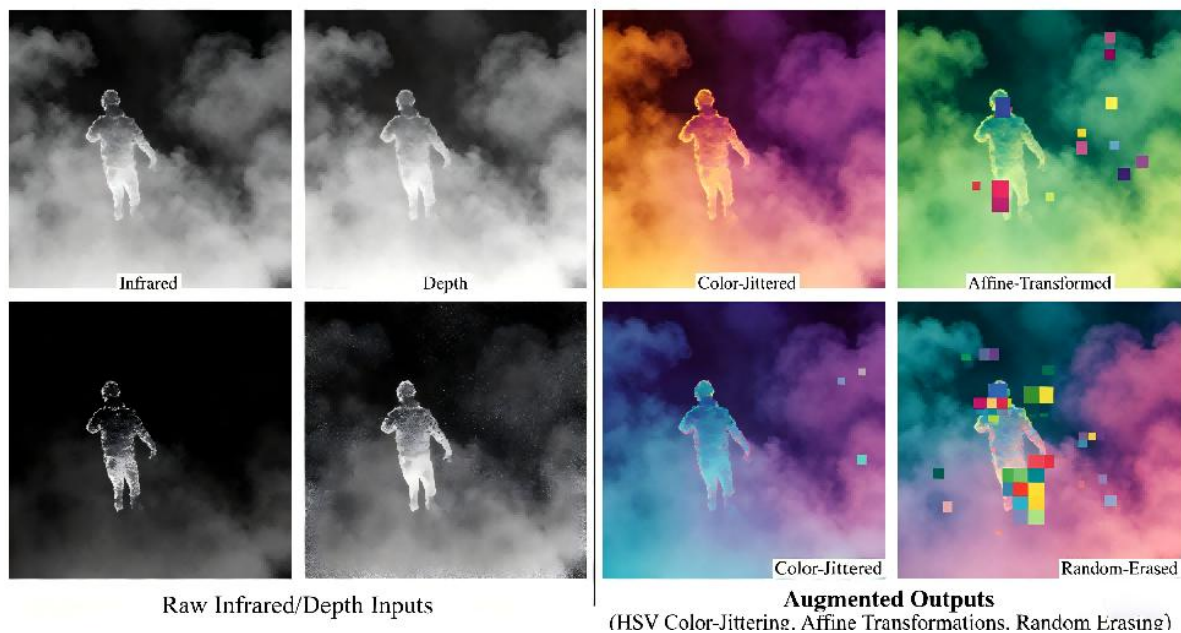
II. RESEARCH METHODS

2.1 Dataset Collection and Preprocessing

To cultivate the model's structural adaptability across practical disaster topography, this study established a multi-modal, large-scale annotated dataset tailored for complex ruins, fire zones, and low-visibility scenarios. The data collection covers chaotic masonry, simulated industrial smoke, extreme low-light environments, and diverse thermal clutters. Samples involving severe structural occlusions account for 30% of the database. The original dataset contains 1,000 sets of synchronized 3D point clouds and high-resolution infrared thermal radiation image pairs, strictly partitioned into training, validation, and testing splits at an academic standard ratio of 6:3:1.

Because airborne particulates and smoke scattering induce noticeable thermal noise in infrared sensors, an adaptive HSV (Hue-Saturation-Value) color-jittering algorithm was implemented. The hue parameter was constrained within ± 0.015 , while saturation and value variations were confined within ± 0.7 and ± 0.4 , respectively. Furthermore, geometric manipulations were enforced: maximum horizontal and vertical translations were restricted to 10%, scaling factors within $\pm 50\%$, and random horizontal mirroring and random erasing applied with a probability of 0.5. The database was expanded to 2,000 augmented samples.

Comparison of Augmented Multi-modal Dataset Samples



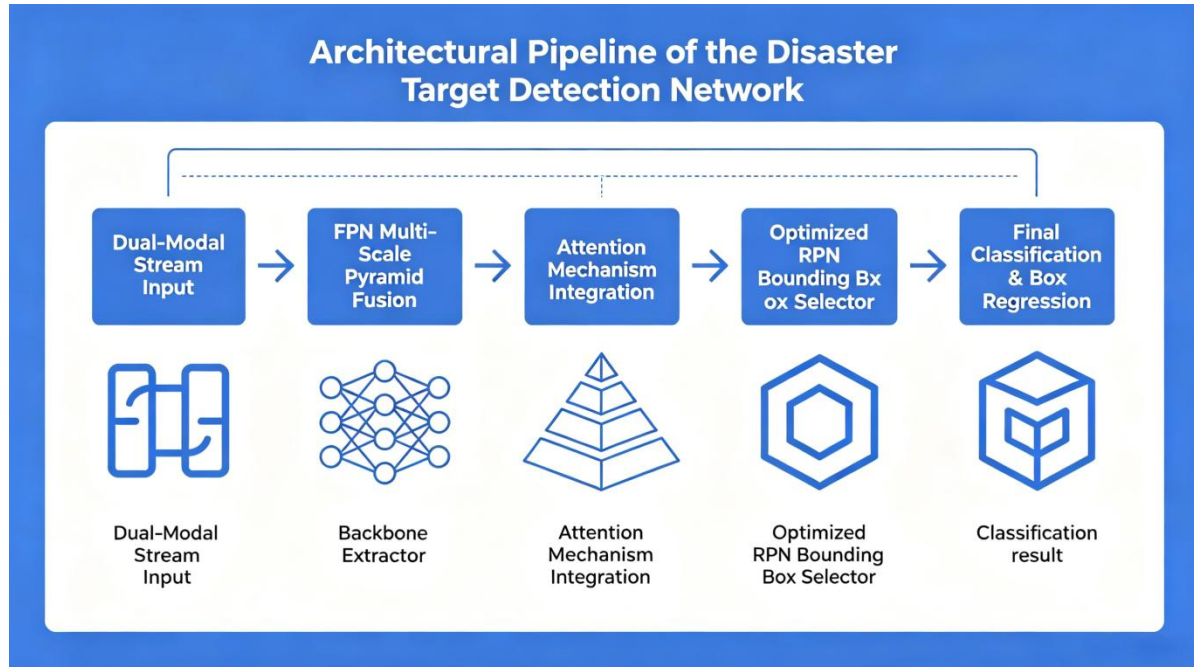
2.2 Algorithm Analysis and Selection

In object detection, frameworks are primarily categorized into One-Stage dense predictors (such as the YOLO family) and Two-Stage region proposal networks (such as the Faster R-CNN paradigm). One-Stage detectors offer superior processing speed. However, disaster ruins exhibit high spatial complexity. Trapped survivors are often heavily masked, revealing only highly faint edge boundaries. To fulfill the strict requirement of 'zero omission' in search missions, an improved Faster R-CNN architecture was selected. A Feature Pyramid Network (FPN) was embedded into the backbone to integrate high-resolution spatial geometry from shallow layers with high-dimensional semantic abstractions from deep layers.

2.3 Algorithm Architecture and Loss Optimization

The core network introduces key architectural enhancements: a Squeeze-and-Excitation (SE) channel attention mechanism is integrated after the feature-extracting backbone to adaptively magnify the weights of informative human thermal channels. During model optimization, a Focal Loss function was incorporated to resolve severe gradient imbalances caused by dominant background regions:

$$L_{\text{focal}} = -\alpha_t * (1 - p_t)^{\gamma} * \log(p_t)$$



2.4 Embedded Control and Autonomous SLAM Navigation

To ensure autonomous navigation in GPS-denied environments, the system runs a real-time Simultaneous Localization and Mapping (SLAM) algorithm driven by spatial depth point clouds. The flight control system deploys a quaternion-based complementary filtering framework for attitude estimation, paired with a high-frequency cascaded double-loop PID controller:

$$u(t) = K_p * e(t) + K_i * \int e(\tau) d\tau + K_d * (de(t)/dt)$$

III. RESULTS AND DISCUSSION

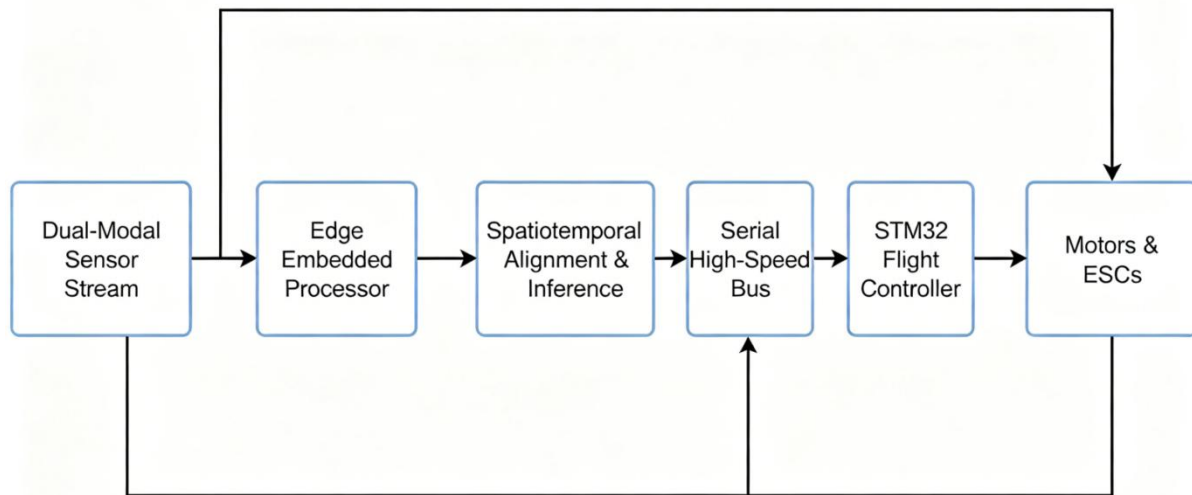
3.1 Algorithmic Performance Evaluation

Model Architecture	Precision	Recall	mAP50	Accuracy (<50m)
Faster R-CNN	0.841	0.795	0.822	62.1%
YOLOv5	0.852	0.810	0.843	68.5%
YOLOv11	0.891	0.865	0.898	75.2%
Proposed Framework	0.931	0.898	0.917	83.2%

Table I Quantitative Comparison of Detection Frameworks

3.2 Physical System Integration and Field Deployment

The physical prototype integrates an STM32F411 microprocessing unit, which runs high-frequency cascaded PID attitude control loops. The airframe is wired to an MPU9250 (9-axis IMU), a BMP280 (barometer), a PMW3901 (optical flow module), and a VL53L01X LiDAR Time-of-Flight altimeter. The multi-sensor data fusion engine yields a solid autonomous positioning precision of ± 5 cm in GPS-denied configurations.



IV. CONCLUSION

This study successfully developed an autonomous UAV search and rescue system utilizing dual-modal fusion of depth cameras and infrared thermal imaging. The system eliminates traditional reliance on external GPS streams and manual operator control. Field trials verified that the system delivers centimeter-level anti-drift localization (± 5 cm) and achieves a human identification rate of 91.7% in complex ruins, expediting life-detection operations by 4.7 times. This stable, cost-effective platform exhibits highly viable engineering potential.

REFERENCES

- [1] Rao E M ,Simon J ,Moll J , et al.Real-time onboard propeller fault diagnosis of autonomous delivery drones through vibration analysis[J].Structural Health Monitoring,2026,25(4):2739-2750.
- [2]Lihui W ,Yi Y ,Wenxin D .HWDiT-Net: hybrid wavelet-dilated transformer network for UAV power-line segmentation[J].Engineering Research Express,2026,8(13):135203-135203.
- [3]Lu H ,Xue M ,Zhao Y , et al.Event-based formation tracking control for multi-agent systems with deception attacks and actuator faults[J].Franklin Open,2026,16100666-100666.
- [4] Xia S ,Sha C ,Wang J , et al.Base station energy-aware UAV data collection with combined single-hop and double-hop communication in Ad Hoc Networks[J].Ad Hoc Networks,2026,192104329-104329.
- [5] Gao Z ,Bao W ,Wang N , et al.Text-guided multi-modal learning for wheat fusarium head blight detection using UAV[J].Biosystems Engineering,2026,269104489-104489.