An Autonomous Multimodal System for Intelligent Railway Inspection

Boshi Chen[1†], Jiawei Guo[1†], Qian Zhang[2], and Yi Wang[1*]

[1]*Department of Mechanical Engineering, University of South Carolina, Columbia, SC, 29208*

jiaweig@email.sc.edu; boshi@email.sc.edu

yiwang@cec.sc.edu (Corresponding Author)

[2]*Department of Systems Engineering, College of Charleston, Charleston, SC, 29401*

zhangq@cofc.edu

† Authors have the same contribution

ABSTRACT

We propose an autonomous aerial inspection system to address growing safety concerns of railway infrastructure degradation. Unlike conventional labor- and sensor-intensive methods, our quadrotor integrates a depth camera, monocular inspection camera, Global Positioning System (GPS) module, and onboard computing unit. Combining visual-inertial fusion with GPS, it achieves robust localization even in GPS-denied environments. A lightweight deep learning model built on You Only Look Once v12 (YOLOv12) enables real-time detection of key components such as spikes and clips. To enhance autonomy, we introduce Railway Autonomous Navigation Guided by Embedded Recognition (RANGER), a novel algorithm that reconstructs 3D world coordinates from 2D detections using only onboard sensing, without requiring prior global maps. By fusing detection with localization data, RANGER enables precise track following and stable altitude control in complex or GPS-denied conditions. This reduces hardware demand while ensuring accurate navigation. Our system reduces operational costs, enhances scalability, and enables accurate, real-time inspections in complex, unstructured environments.

1. INTRODUCTION

The railway system is a vital component of national and regional transportation networks, providing a dependable and efficient means of moving freight and passengers. In the United States, the railway network stretches over 225,000 kilometers, making it the most extensive in the world (Association of American Railroads, 2024). However, as the infrastructure continues to age, safety concerns are growing, particularly the integrity of small yet essential components, including spikes, fasteners, and clips. According to the Federal Railroad Administration (2024), failures in these components contributed to more than 400 railway accidents in 2024 alone. Traditional inspection methods, including manual visual examinations and vehicle-mounted systems, are often time-consuming, labor-intensive, and subjective, along with operational constraints. These methods are ineffective for inspecting remote or hard-to-access locations, such as tunnels, bridges, and mountain regions, due to significant logistical difficulties and costs.

To address these challenges, computer vision and deep learning technologies have been increasingly adopted for automated railway infrastructure inspection. For example, Zheng et al. (2021) proposed a multi-stage detection framework based on YOLOv5 (Jocher, 2020), Mask R-CNN (He et al., 2017), and ResNet (He et al., 2016), achieving high performance in identifying and classifying defects on rails and fasteners. Wang et al. (2021) introduced AttnConv-Net, an attention-enhanced convolutional model that utilizes cascaded attention blocks and positional encoding to improve the detection of multiple rail components, such as bolts and clips, without requiring extensive pre- or post-processing. Semantic segmentation models, such as U-Net (Ronneberger et al., 2015) and DeepLabV3+ (Chen et al., 2018), have enabled pixel-level fault localization, which is crucial for detailed structural assessment (Weng et al., 2023). Hybrid models such as SSD-Faster Net (J. Wang & Yu, 2022) have combined fast object localization with refined segmentation to enhance inspection effectiveness in complex environments. These techniques have significantly improved the performance of automated inspection systems, enabling more

Boshi Chen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

comprehensive and real-time detection capabilities. Nonetheless, despite these advances, such systems remain computationally intensive and primarily rely on fixed ground-based sensors or inspection vehicles, limiting their coverage and flexibility for real-time deployment in inaccessible areas.

Recently, Unmanned Aerial Vehicles (UAVs) have emerged as a promising alternative for railway inspection due to their mobility, flexibility, and non-contact sensing capabilities. UAVs can rapidly cover large and complex terrains, making them ideal for inspecting long-distance railway lines and hard-to-access infrastructure. For instance, Qiu et al. (2024) developed a UAV-based track geometry measurement system using LiDAR and IMU sensors combined with SLAM algorithms. This system can measure track gauge, curvature, and alignment with sub-inch accuracy without interrupting regular train operations. Similarly, Xu et al. (2023) proposed a vision-based autonomous UAV inspection framework for tunnel environments. Their system leverages RGB-D cameras and dynamic mapping modules to navigate unknown and obstacle-dense construction sites, enabling 3D reconstruction and autonomous flight planning without prior mapping. Moreover, UAVs have been integrated with structural health monitoring (SHM) systems to evaluate the seismic safety and surface integrity of railway infrastructure. Liu (2023) applied deep learning models such as AlexNet (Krizhevsky et al., 2012), VGG(Simonyan & Zisserman, 2014), and ResNet (He et al., 2016) for crack segmentation in concrete components, improving the accuracy and utility of UAV-based SHM systems. Ngeljaratan et al. (2024) utilized MSER-based (Donoser & Bischof, 2006) feature extraction techniques to analyze seismic-induced deformations in linear railway structures using aerial imagery. However, despite these advantages, UAV platforms face several operational and technical limitations. Onboard computational capabilities are often limited by size, weight, and power (SWaP) constraints, making it challenging to run traditional deep learning models in real-time. UAVs are also limited by flight time, payload capacity, and their dependency on satellite signals, which may be unavailable in tunnels, dense urban settings, or forested regions. Furthermore, many UAV-based inspection methods rely heavily on pre-mapped environments, reducing their adaptability in dynamic or unknown contexts.

To mitigate the limitations of deep learning on resource-constrained UAV platforms, researchers have begun to develop lightweight neural networks tailored for aerial systems. Nguyen et al. (2019) proposed MAVNet, a compact segmentation model inspired by ERFNet (Romera et al., 2017), designed for real-time execution on micro aerial vehicles (MAVs). This model significantly reduces the number of parameters while maintaining acceptable segmentation performance. Lee et al. (2023) developed WATT-EffNet, which uses width-wise incremental feature modules and attention mechanisms to achieve high classification accuracy with minimal computational overhead. Guo et al. (2023) introduced AWL-NanoDet, a lightweight object detection model with less than 2 MB in size and 1.52 GFLOPs of computation, capable of real-time defect detection on embedded systems through dynamic loss weighting and teacher-student knowledge distillation. While these lightweight models represent a significant advancement in enabling deep learning on UAVs, most operate independently of the UAV's flight control and planning modules. This disconnect leads to suboptimal performance, with underutilized onboard resources. Additionally, their inspection paths are not dynamically adjusted according to real-time visual input. In essence, these systems detect without influencing navigation or flight behavior, resulting in inefficiencies and missed opportunities for intelligent planning and adaptive coverage.

To bridge this gap, we propose a fully integrated aerial inspection system that combines lightweight deep learning-based perception with real-time flight control and planning. The system is built on a quadrotor platform equipped with a high-resolution inspection camera, a depth camera, and an onboard computing unit. Our framework adopts a tightly coupled sensor fusion approach. The system primarily relies on visual-inertial odometry (VIO). When GPS signals are available, GPS measurements are fused with VIO estimates to improve positioning accuracy. In GPS-denied environments such as tunnels or dense urban areas, the system falls back to VIO alone to ensure robust localization and mapping. At the core of our detection module is a compact, real-time object detection network based on You Only Look Once v12 (YOLOv12) (Tian et al., 2025), optimized for identifying essential railway components, such as rails, fasteners, spikes, and clips, from aerial imagery with minimal latency and computational cost. We introduce the Railway Autonomous Navigation Guided by Embedded Recognition (RANGER) module to convert detected 2D features into global 3D coordinates without relying on LiDAR or stereo vision. This allows the UAV to dynamically adapt its inspection trajectory, improving spatial awareness and ensuring consistent inspection coverage. Additionally, the system features a computationally efficient motion planner capable of generating real-time, collision-free paths without the need for pre-mapped environments. This enables the UAV to autonomously adjust its flight path based on detection confidence, component density, or obstacle presence, improving resource utilization and inspection behavior. Unlike conventional inspection platforms that require expensive hardware or extensive infrastructure, our system is lightweight, cost-effective, and easily deployable across large-scale railway networks. It enhances inspection coverage, reduces labor demands, and enables accurate, real-time defect identification even in challenging or previously inaccessible environments.

The remainder of this paper is organized as follows: Section 2 examines current approaches in UAV autonomous

navigation and discusses limitations of existing methods. Section 3 details the pipeline of our proposed method. Section 4 describes the experimental setup and results with interpretation. Section 5 concludes the paper and outlines directions for future work.

2. PRELIMINARY

This section first reviews common navigation methods used in UAV-based inspection, then analyzes their limitations and motivates the need for a tightly integrated, perception-aware solution.

2.1. Existing Navigation Methods

GPS-based navigation remains the most widely adopted method for UAV positioning in aerial inspection tasks. Global Navigation Satellite Systems (GNSS), such as GPS, GLONASS, Galileo, and BeiDou, estimate absolute positions by triangulating signals from multiple satellites. Under open-sky conditions, these systems typically achieve meter-level accuracy. To meet the precision requirements of infrastructure inspection tasks, Real-Time Kinematic (RTK) correction is often applied (Frodge et al., 1994). RTK improves accuracy to the centimeter level by comparing satellite signal phases between a stationary base station and the UAV receiver. This is particularly crucial for applications such as railway monitoring, where sub-meter spatial resolution is required. As a result, GPS-RTK systems form the foundation of modern UAV navigation pipelines in structured, open environments such as railways, highways, and powerlines.

Building upon GNSS positioning, waypoint-based navigation defines UAV missions as ordered sequences of spatial coordinates, typically specified in either global or local frames. The UAV autonomously follows these waypoints using a position-holding flight controller, pausing at each location to perform tasks such as image capture or sensor measurement. These waypoint trajectories are often pre-defined using ground control software, where operators configure flight paths and behaviors in advance. This method is commonly employed in structured inspection scenarios, including tunnel mapping, bridge deck scanning, and railway corridor surveillance. Repetitive waypoint paths enable consistent spatial coverage and support temporal comparison across inspection intervals. For instance, densely spaced waypoints can guide UAVs to capture overlapping high-resolution images of rail fasteners or tunnel walls, facilitating pixel-level change detection. Commercial UAV platforms further support advanced waypoint features, such as adaptive hovering, gimbal control, and synchronized multi-sensor activation, making this strategy particularly effective for infrastructure monitoring. Though mission plans are pre-defined, many systems allow in-flight adjustments to accommodate dynamic environments or updated inspection goals.

Visual Simultaneous Localization and Mapping (SLAM) provides an alternative navigation strategy, enabling UAVs to estimate their position and map their surroundings without relying on GPS. This makes SLAM especially suitable for GPS-denied environments such as tunnels, underpasses, or dense urban areas. SLAM systems integrate visual input from monocular, stereo, or depth cameras with inertial measurements to perform simultaneous pose estimation and map construction. Modern visual SLAM frameworks, such as ORB-SLAM3(Campos et al., 2021), VINS-Mono(Qin et al., 2018), and DSO(Engel et al., 2017), employ multi-stage pipelines involving visual feature extraction, matching, pose graph optimization, loop closure, and bundle adjustment. These methods support real-time onboard execution by leveraging keyframe-based optimization and tightly coupled VIO, ensuring robust performance in the presence of motion blur, lighting variations, and occlusions. SLAM-based navigation has been deployed in various UAV inspection scenarios, including bridge span modeling, rail tunnel surveying, and confined space exploration. By enabling drift-corrected localization without external positioning infrastructure, SLAM allows UAVs to operate autonomously in previously inaccessible environments and adapt their paths based on the surrounding geometry.

2.2. Limitations of Existing Navigation Methods

Despite significant advancements, existing UAV navigation methods remain largely decoupled from the core inspection and detection objectives. GPS-based and waypoint-following approaches depend on pre-defined trajectories or satellite-based localization, which ensure spatial coverage but may fail to align the UAV's perception system with inspection targets such as clips, fasteners, or surface defects. These targets are regularly placed along the track, but a pre-defined flight path may not keep them within view. These methods typically treat detection as a post-processing step or as an independent module that does not influence flight behavior. Similarly, while visual SLAM provides real-time localization in GPS-denied environments, it is primarily designed for mapping and pose estimation rather than task-specific, perception-driven navigation. Most SLAM implementations focus on full-area coverage or loop closure, without the capability to prioritize regions of higher inspection relevance or dynamically modify paths in response to real-time visual cues. As a result, flight trajectories are often rigid or heuristically defined, limiting adaptability in complex or cluttered environments.

In practice, detection modules on UAV platforms tend to operate passively, collecting data without influencing flight control or motion planning. This results in suboptimal inspection performance, inefficient resource utilization, and limited opportunities for focused, high-value data collection. While a few recent studies have begun exploring perception-aware navigation, only a limited number have attempted to tightly couple real-time object detection with motion

planning, especially in unstructured or large-scale environments such as railway networks.

3. PROPOSED RAILWAY AUTONOMOUS NAVIGATION GUIDED BY EMBEDDED RECOGNITION (RANGER)

To address the limitations of conventional GPS-based or SLAM-based UAV navigation, we propose RANGER, Railway Autonomous Navigation Guided by Embedded Recognition, a real-time, vision-driven waypoint generation framework tailored for autonomous railway inspection. Unlike traditional systems that decouple perception from motion planning, RANGER tightly integrates rail detection and waypoint generation within a unified, vision-based navigation pipeline. The core idea is to extract spatially meaningful waypoints directly from visual observations, enabling the UAV to dynamically follow the physical railway path without relying on external maps or GPS infrastructure.

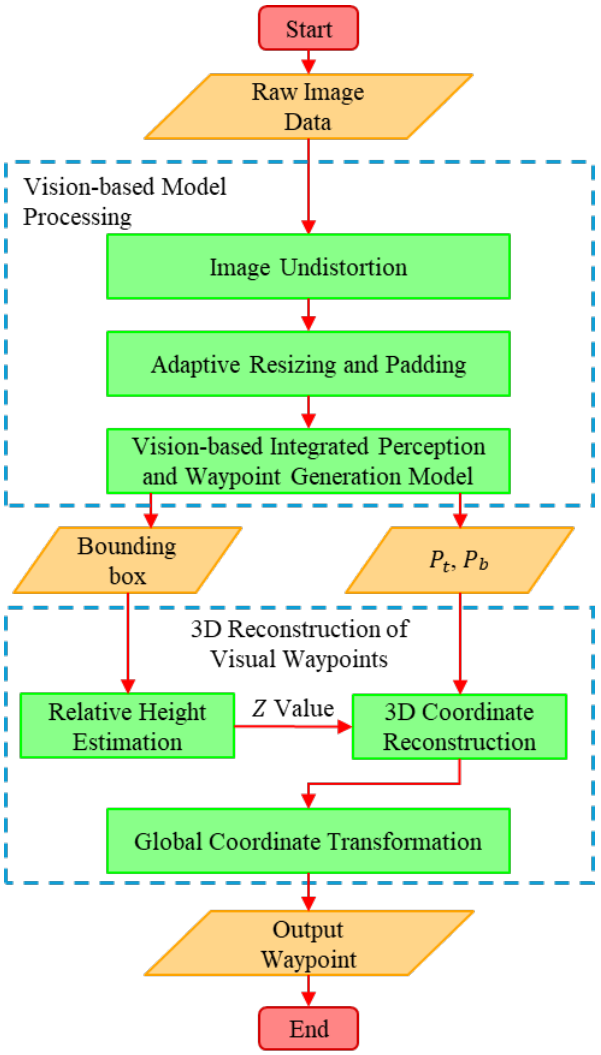

Figure 1. Pipeline of Proposed Railway Autonomous Navigation Guided by Embedded Recognition (RANGER)

As shown in Figure 1, our proposed RANGER pipeline comprises two main modules: (1) Embedded Recognition, which utilizes the Vision-based Integrated Perception and Waypoint Generation Model to obtain centerline reference points of the railway, and (2) Railway Autonomous Navigation, which corresponds to the 3D Reconstruction of Visual Waypoints.

This unified perception-to-planning pipeline ensures that each waypoint is both visually grounded and spatially accurate, allowing the UAV to continuously adapt its flight path based on actual rail geometry. By bridging object detection with onboard navigation, RANGER enables robust, map-free inspection along complex or unstructured railway corridors. throughout.

3.1. Vision-based Integrated Perception and Waypoint Generation Model

To enable autonomous UAV navigation along railway tracks, we design a vision-based integrated perception and waypoint generation model built on the lightweight YOLOv12-n (Tian et al., 2025) architecture. The detector identifies the positions of the rail in each image frame by generating bounding boxes that tightly align with the orientation of the tracks.

From each bounding box, we extract two centerline reference points: the top center (P_t) and bottom center (P_b), which represent the visual axis of the railway in the current frame. These two points are used to dynamically guide UAV flight. Formally, the waypoint centers are computed as:

$$P_t, P_b = \left(\frac{x_t^l + x_t^r}{2}, \frac{y_t^l + y_t^r}{2}\right), \left(\frac{x_b^l + x_b^r}{2}, \frac{y_b^l + y_b^r}{2}\right) \quad (1)$$

where (x_t^l, y_t^l) and (x_t^r, y_t^r) refer to the topmost points on the left and right sides of the bounding box, respectively. Similarly, (x_b^l, y_b^l) and (x_b^r, y_b^r) refer to the bottommost points on the left and right sides.

As shown in the Figure 2, these points are visualized in the rail. Specifically, the yellow dots at the top of the image denote the topmost points on the left and right boundaries of the bounding box, while the red dot indicates the top center point. Likewise, the yellow dots at the bottom correspond to the bottommost points on the left and right sides, and the red dot marks the bottom center point. They are used as dynamic waypoints to guide UAV motion frame by frame. This approach enables real-time flight without relying on external maps or localization infrastructure.

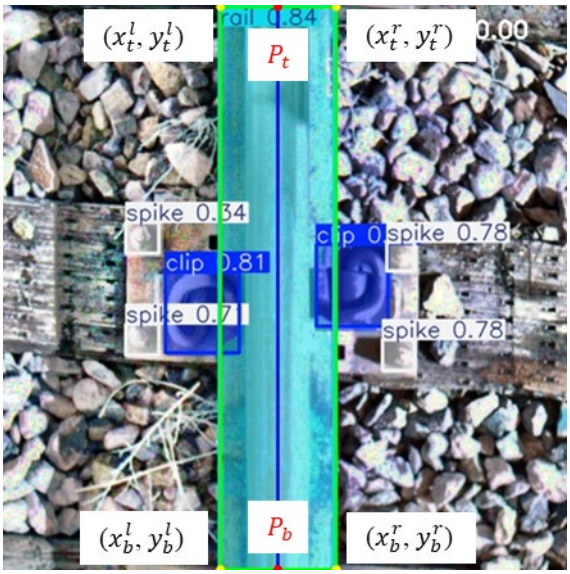

Figure 2. Bounding Box and Reference Points Extracted by Vision-Based Integrated Perception and Waypoint Generation Model

3.2. 3D Reconstruction of Visual Waypoints

This module aims to enable real-time, vision-based rail tracking and autonomous navigation for UAVs using a monocular inspection camera as the primary sensor. The downward-facing inspection camera is rigidly mounted on the underside of the UAV, aligned perpendicularly to the ground. During flight, this camera continuously captures sequential images of the railway beneath the UAV. These image frames are processed through a combination of computer vision and geometric reasoning techniques to extract two centerline reference points as mentioned in Section 3.1. To convert 2D image observations into real-world 3D coordinates, the system employs a reconstruction pipeline based on the pinhole camera model.

Initially, raw images are corrected for geometric distortion induced by the lens, eliminating both radial and tangential aberrations. Subsequently, the image is resized to fit the input resolution of the vision-based integrated perception and waypoint generation model. This resizing step records the applied scaling factor and padding offsets, allowing accurate inverse projection of the model's output to the original image frame.

One critical requirement for accurate 3D back-projection is the estimation of the UAV's relative altitude (Z-axis distance) to the railway surface. The system provides two operating modes to support different application scenarios: adaptive height estimation mode and preset height mode. In adaptive height estimation mode, the system dynamically estimates UAV altitude based on the observed pixel width of the railway in the image. Given the intrinsic camera matrix and the real-world rail width w_r, the altitude Z can be computed by applying the triangle Similarity Principle:

$$Z = f \cdot \frac{w_r}{w_p} \tag{2}$$

where f is the focal length of the camera (in pixels) and w_p is the measured rail width in pixels. This mode is especially useful in dynamically changing environments such as uneven terrain or hilly regions, where adaptive altitude estimation improves spatial localization accuracy. Alternatively, in stable environments or latency-sensitive scenarios, users can specify a constant height using preset height mode, which remains unchanged throughout the mission.

Once the altitude Z is known (either adaptively estimated or manually defined), the spatial position of the centerline reference points can be calculated. The model identifies two essential centerline reference points on the rail centerline: the upper endpoint P_t (top center) and the lower endpoint P_b (bottom center), both in pixel coordinates. These points act as visual anchors representing the projected direction and spatial extent of the rail within the image.

To reconstruct the 3D positions of P_t and P_b, the system utilizes the standard pinhole camera model. Let $p = (u, v)^\top \in R^2$ be the pixel coordinates of a keypoint in the image. Assuming a known altitude Z, and given the camera's intrinsic parameters: focal lengths f_x, f_y and principal point offsets c_x, c_y, the back-projected 3D position $P_{cam} = (X_c, Y_c, Z_c)^\top \in R^3$ in the camera coordinate frame is calculated as:

$$X_c = \frac{(u - c_x) \cdot Z}{f_x} \tag{3}$$

$$Y_c = \frac{(v - c_y) \cdot Z}{f_y} \tag{4}$$

$$Z_c = Z \tag{5}$$

Alternatively, this projection can be expressed compactly using the intrinsic matrix K as:

$$\tilde{\boldsymbol{p}} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{6}$$

where $\tilde{\boldsymbol{p}} \in R^3$ represents the normalized image coordinate in the camera frame. The corresponding 3D position of the point in the camera coordinate frame $\boldsymbol{P}_{cam} \in R^3$ is then given by:

$$\boldsymbol{P}_{cam} = Z \cdot \tilde{\boldsymbol{p}} \tag{7}$$

This result is then transformed into the UAV's body frame using a known extrinsic matrix $T_{cb} \in SE(3)$, which defines the rigid-body transformation from the camera to UAV coordinate system:

$$\boldsymbol{P}_{\boldsymbol{body}} = T_{cb} \cdot \boldsymbol{P}_{cam} \tag{8}$$

Here, $\boldsymbol{P}_{\boldsymbol{body}} \in R^3$ represents the 3D position in the UAV's body coordinate frame. With P_t and P_b now reconstructed in the UAV body frame, the system computes a normalized

direction vector that represents the railway's forward extension. Specifically, a directional unit vector d is computed from P_b to P_t as follows:

$$\boldsymbol{d} = \frac{P_t - P_b}{\|P_t - P_b\|} \tag{9}$$

This unit vector captures the spatial direction of the railway in the UAV's local frame. To identify a navigable target point ahead of the UAV, the system projects a new point along this direction by a fixed distance λ, such as 2 meters. This yields the intermediate target waypoint in the body frame:

$$\boldsymbol{W}_{body} = \boldsymbol{P}_{\boldsymbol{body}} + \lambda \cdot \boldsymbol{d} \tag{10}$$

This operation predicts a flyable waypoint ahead of the UAV that lies along the perceived rail axis, enabling smooth and progressive navigation aligned with the railway.

To integrate this waypoint into the UAV's global navigation framework, $\boldsymbol{W}_{body}$ is transformed from the UAV body frame into the global coordinate frame. The transformation uses the UAV's current pose, represented by its global position vector $t \in R^3$ and orientation quaternion, which is converted to a rotation matrix $R \in SO(3)$. The global 3D waypoint W_{global} is then computed as:

$$\boldsymbol{W}_{\text{global}} = R \cdot \boldsymbol{W}_{body} + \boldsymbol{t} \tag{11}$$

This final waypoint, now expressed in the global coordinate frame, encapsulates both the geometric alignment of the railway and the UAV's spatial context. It can be passed to the downstream flight control module for real-time trajectory tracking, enabling robust, map-free autonomous navigation along the railway solely using visual input.

4. EXPERIMENTAL SETUP AND RESULT

To evaluate the performance of our proposed system, we conducted a series of experiments, including both simulation and real-world flight tests. The simulation experiments were carried out within a Gazebo environment. The real-world experiments took place at an abandoned railway site. A customized quadrotor UAV was deployed, equipped with a RealSense D435 depth camera, a PX4-based flight controller, an Intel NUC onboard computing unit, a ZED-F9P GPS module, and a downward-facing inspection camera. A VIO algorithm was used to estimate the UAV's pose and motion state. All computations were performed onboard, enabling fully autonomous operation.

4.1. System Hardware Setup

The overall UAV hardware configuration is illustrated in Figure 3. The platform is equipped with a comprehensive suite of sensors and computing modules designed to support autonomous railway inspection and navigation tasks.

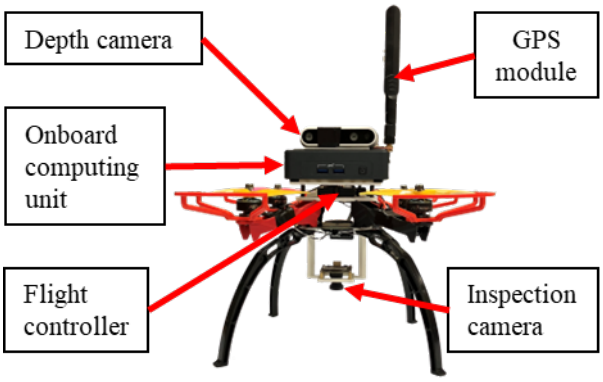

Figure 3. Hardware Setup

A RealSense D435 depth camera is mounted on the UAV, providing depth images for real-time mapping and obstacle avoidance. Additionally, the RealSense D435 camera provides a monocular image stream used as visual input for the VIO module. A PX4-based flight controller interfaces with the onboard computer, executing low-level flight control and providing inertial measurements from its integrated IMU. These measurements are fused with visual input to enable accurate pose estimation. The primary onboard computing unit is an Intel NUC, which delivers sufficient computational performance for VIO, waypoint planning, and perception tasks. Its integrated Iris GPU further supports real-time object detection using lightweight neural networks. To enhance localization accuracy, a ZED-F9P GPS module provides high-precision positioning data that is fused with VIO outputs. In addition, a downward-facing RGB inspection camera is installed on the underside of the UAV to capture high-resolution imagery of the railway tracks during flight.

4.2. System Software Architecture

The software architecture of our system, illustrated in Figure **4**, is implemented using the Robot Operating System (ROS) Noetic. For position estimation, we utilize the Global Visual-Inertial Navigation System (GVINS) (Cao et al., 2022). This module significantly mitigates the long-term drift typically associated with standalone VIO systems, enhancing both robustness and global consistency. In addition, GVINS provides the UAV's position and orientation estimates, which are used as state information inputs to the RANGER module. The proposed RANGER module generates the waypoints in 3D space once the railroad is detected. These waypoints act as targets guiding the UAV along the inspection route. To ensure safe traversal between waypoints and facilitate responsive navigation in dynamic or cluttered environments, local obstacle avoidance is handled by EGO-Planner(Zhou et al., 2020). This planner computes safe, dynamically feasible trajectories.

All perception, state estimation, planning, and control modules run fully onboard the UAV on the Intel NUC computer, enabling complete autonomy without dependence on external computation or communication.

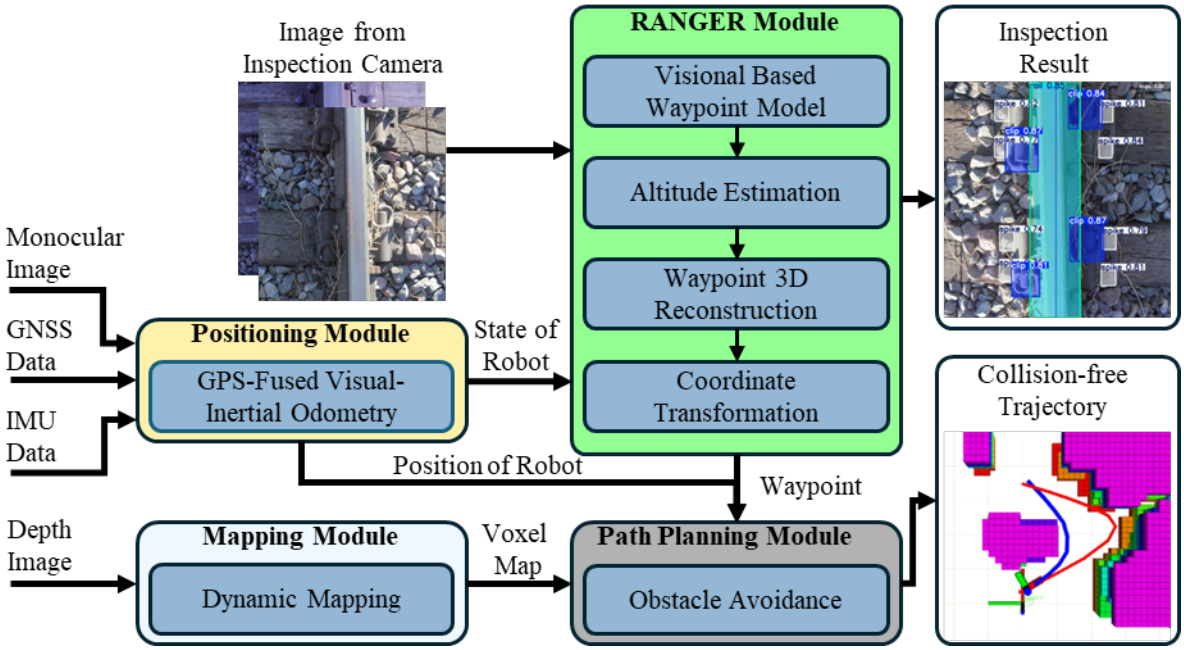

Figure 4. Software Architecture

4.3. Flight Experiment in Simulation

We conducted independent testing of the RANGER module in a simulated environment using the Gazebo platform. A realistic railway scene was created, as shown in Figure 5, to evaluate the module's adaptability to typical inspection tasks. To better approximate real-world flight conditions, wind disturbances were introduced into the simulation. The tests focused on verifying the module's ability to detect railway tracks and generate corrective waypoints when the UAV deviated from the track centerline. These simulation-based, module-level evaluations enabled us to assess the functional stability and reliability of the RANGER module under complex environmental conditions before full system integration and real-world deployment.

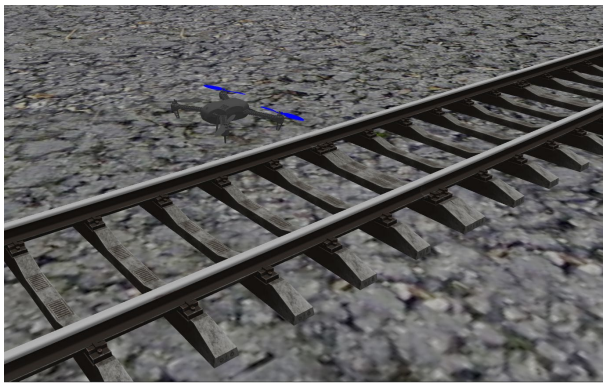

Figure 5. Gazebo Simulation Environment

4.4. Real World Flight Experiment

To validate the effectiveness and robustness of our proposed autonomous inspection system, we conducted real-world flight experiments at an abandoned railway site located in a semi-structured outdoor environment. This environment presents several challenges, including uneven terrain, varying lighting conditions, and the presence of static obstacles such as poles, vegetation, and infrastructure debris, making it a suitable testbed for evaluating the system's perception, localization, and planning capabilities.

The primary objective of the experiment was to autonomously inspect the railway while maintaining safe flight and avoiding obstacles. The UAV was launched autonomously. During the mission, the UAV used the onboard positioning module to continuously estimate its pose. The downward-facing RGB inspection camera captured high-resolution images of the railway during the flight. The RANGER module was configured to operate in preset height mode, as the surrounding terrain was relatively flat. The UAV followed the waypoints generated by the RANGER module, while EGO-Planner computed collision-free trajectories in real time based on onboard depth perception and continuously updated maps. All system modules, including VIO, navigation, obstacle avoidance, and image logging, ran fully onboard, without requiring any offboard processing or human intervention.

Figure 6. Real World Flight Experiment

4.5. Result

We evaluated the detection function of RANGER module using a custom-labeled dataset of railway images. The vision-based integrated perception and waypoint generation model in RANGER module achieved a mean Average Precision (mAP@0.5) of 0.954 across all classes, with a precision of 0.894 and a recall of 0.97, demonstrating robust detection performance.

We further validated the effectiveness of the proposed RANGER module in simulation scenarios. As shown in Figure 7, the simulated environment differs from the real world in lighting conditions, object textures, and other visual features, the model was never trained using simulated images. Figure 8 shows the detection results in the simulation. During testing, the model successfully detected railway tracks within the simulation and generated segmentation masks for use by the RANGER module. This demonstrates the model's strong generalization capability. As illustrated in Figure 9, during simulated flights, by utilizing the RANGER module to detect and send target points, the railway tracks were consistently maintained near the center of the downward-facing camera's field of view.

Figure 7. Comparison between railway tracks in the simulated environment (left) and the real-world environment (right)

Figure 8. RANGER Result in Simulation

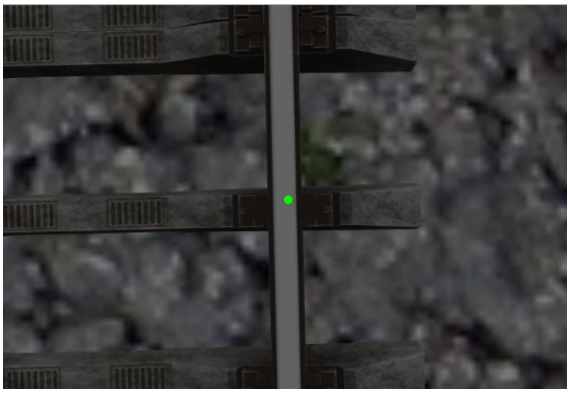

Figure 9. Visualization of the onboard inspection camera placement on the UAV. The green dot indicates the optical center of the camera.

5. CONCLUSION

This paper presents a fully integrated aerial inspection system that combines real-time object detection, onboard sensing, and perception-aware navigation for autonomous railway inspection. Leveraging a lightweight quadrotor platform equipped with a monocular camera, depth sensor, GPS, and onboard computing, the system achieves robust localization through visual-inertial-GPS fusion. The proposed vision-based detection module enables real-time identification of key railway components with minimal computational overhead. To close the loop between perception and motion, we introduce RANGER, a novel navigation algorithm that reconstructs 3D target positions from 2D detections, guiding the UAV without requiring pre-mapped environments or high-end sensors. This approach significantly enhances autonomy, efficiency, and adaptability in GPS-denied or cluttered scenarios. Field evaluations demonstrate accurate target detection, stable flight, and intelligent path adjustment in real time. Future work will focus on extending multi-target prioritization strategies, improving robustness in adverse weather conditions, and validating long-term deployment performance across diverse railway infrastructures.

REFERENCES

Association of American Railroads. (2024). *Resources*. Https://Www.Aar.Org/Resources/. https://www.aar.org/Resources/

Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M. M., & Tardós, J. D. (2021). Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, *37*(6), 1874–1890.

Cao, S., Lu, X., & Shen, S. (2022). GVINS: Tightly coupled GNSS–visual–inertial fusion for smooth and consistent state estimation. *IEEE Transactions on Robotics*, *38*(4), 2004–2021.

Donoser, M., & Bischof, H. (2006). Efficient maximally stable extremal region (MSER) tracking. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, *1*, 553–560.

Engel, J., Koltun, V., & Cremers, D. (2017). Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(3), 611–625.

Federal Railroad Administration. (2024). *Railroad Equipment Accident/Incident Source Data (Form 54)*. https://data.transportation.gov/Railroads/Railroad-Equipment-Accident-Incident-Source-Data-F/aqxq-n5hy/about_data

Frodge, S. L., DeLoach, S. R., Remondi, B., Lapucha, D., & Barker, R. A. (1994). Real-Time on-the-Fly Kinematic GPS System Results. *Navigation*, *41*(2), 175–186.

Guo, J., Zhang, S., Qian, Y., & Wang, Y. (2023). A NanoDet Model with Adaptively Weighted Loss for Real-time Railroad Inspection. *Annual Conference of the PHM Society*, *15*(1).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Jocher, G. (2020). *Ultralytics YOLOv5*. https://doi.org/10.5281/zenodo.3908559

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Lee, G. Y., Dam, T., Ferdaus, M. M., Poenar, D. P., & Duong, V. N. (2023). Watt-effnet: A lightweight and accurate model for classifying aerial disaster images. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5.

Liu, K. (2023). Learning-based defect recognitions for autonomous UAV inspections. *ArXiv Preprint ArXiv:2302.06093*.

Ngeljaratan, L., Bas, E. E., & Moustafa, M. A. (2024). Unmanned Aerial Vehicle-Based Structural Health Monitoring and Computer Vision-Aided Procedure for Seismic Safety Measures of Linear Infrastructures. *Sensors*, *24*(5), 1450.

Nguyen, T., Shivakumar, S. S., Miller, I. D., Keller, J., Lee, E. S., Zhou, A., Özaslan, T., Loianno, G., Harwood, J. H., & Wozencraft, J. (2019). Mavnet: An effective semantic segmentation micro-network for mav-based tasks. *IEEE Robotics and Automation Letters*, *4*(4), 3908–3915.

Qin, T., Li, P., & Shen, S. (2018). Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, *34*(4), 1004–1020.

Qiu, L., Zhu, M., Park, J., & Jiang, Y. (2024). Non-Interrupting Rail Track Geometry Measurement System Using UAV and LiDAR. *ArXiv Preprint ArXiv:2410.10832*.

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, *19*(1), 263–272.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Tian, Y., Ye, Q., & Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. *ArXiv Preprint ArXiv:2502.12524*.

Wang, J., & Yu, N. (2022). SSD-faster net: A hybrid network for industrial defect inspection. *ArXiv Preprint ArXiv:2207.00589*.

Wang, T., Zhang, Z., Yang, F., & Tsui, K.-L. (2021). Automatic rail component detection based on AttnConv-Net. *IEEE Sensors Journal*, *22*(3), 2379–2388.

Weng, Y., Li, Z., Chen, X., He, J., Liu, F., Huang, X., & Yang, H. (2023). A railway track extraction method based on improved DeepLabV3+. *Electronics*, *12*(16), 3500.

Xu, Z., Chen, B., Zhan, X., Xiu, Y., Suzuki, C., & Shimada, K. (2023). A vision-based autonomous UAV inspection framework for unknown tunnel construction sites with dynamic obstacles. *IEEE Robotics and Automation Letters*, *8*(8), 4983–4990.

Zheng, D., Li, L., Zheng, S., Chai, X., Zhao, S., Tong, Q., Wang, J., & Guo, L. (2021). A defect detection method for rail surface and fasteners based on deep convolutional neural network. *Computational Intelligence and Neuroscience*, *2021*(1), 2565500.

Zhou, X., Wang, Z., Ye, H., Xu, C., & Gao, F. (2020). Ego-planner: An esdf-free gradient-based local planner for quadrotors. *IEEE Robotics and Automation Letters*, *6*(2), 478–485.

BIOGRAPHIES

Boshi Chen received his B.S. degree in Biomedical Engineering from Hefei University of Technology, Hefei, China, in 2023. He is currently pursuing a Ph.D. in Mechanical Engineering at the University of South Carolina. His research interests include non-destructive evaluation, robotics, and autonomous systems.

Jiawei Guo earned his B.S. from Tianjin University of Technology and Education (2019), Tianjin, China, and his M.S. from the University of Southern California, CA, USA. He is now pursuing his Ph.D. of mechanical engineering with the University of South Carolina. His research interests are in computer vision and machine learning for engineering applications.

Qian Zhang is an Assistant Professor of Systems Engineering at the College of Charleston. She completed her postdoctoral research at Houston Methodist Hospital in Texas in 2022 and earned her Ph.D. in Industrial and Systems Engineering (Human Factors and Ergonomics) from the State University of New York at Buffalo (2017-2022). She holds an M.S. in Automotive Engineering from Jilin University (2011-2014) and a B.A. in Mechanical Engineering from Zhengzhou University (2007-2011). Her research interests include human-robot collaboration, human factors in medical device development, and wearable technologies. Dr. Zhang is a member of the Human Factors and Ergonomics Society and the American Society of Safety Professionals and serves as a reviewer for journals such as IEEE Human-Machine Systems, Virtual Reality, and the International Journal of Human-Computer Interaction. Dr. Zhang was bestowed South Carolina Manufacture Maven Award 2024.

Yi Wang earned his B.S. and M.S. from Shanghai Jiao Tong University (1998, 2000), and his Ph.D. from Carnegie Mellon University (2005). Currently, he is a Professor at the University of South Carolina. His research focuses on computational and data-enabled science and engineering, multi-fidelity surrogate modeling, machine learning, computer vision, and autonomous systems.

APPENDIX

The total cost of the UAV system is detailed in Table 1, amounting to USD 1,967.15. This remains under the target budget of USD 2,000, fulfilling the requirements for a low-cost, modular, and functional platform suitable for research and prototyping.

Table 1. Cost breakdown of the Intelligent Railway Inspection System

Item	Unit	Qty	Unit Price	Total Price (Each)
Intel NUC 11th gen Intel Core i5-1135G7 Tiger Canyon	set	1	$555.00	$555.00
G.SKILL Ripjaws 16GB DDR4 Laptop Memory	piece	1	$29.99	$29.99
Samsung - 990 PRO 1TB Internal SSD	piece	1	$119.99	$119.99
SpeedyFPV Q250 250mm Quadcopter Drone Frame Kit	kit	1	$18.86	$18.86
Universal Landing Gear	pack	1	$6.39	$6.39
EMAX Formula Series 45A ESC support	piece	4	$44.99	$179.96
F60PRO IV V2.0 Fpv Racing Drone Motor 4-6S KV2550	box	4	$26.90	$107.60
Pixhawk 4	set	1	$142.03	$142.03
Radiolink R12DSM 2.4Ghz 12 Channels Micro RC Receiver	piece	1	$26.99	$26.99
Radiolink AT9S Pro 2.4G Radio Controller Transmitter	piece	1	$120.02	$120.02
HDMI Dummy Plug	pack	1	$6.99	$6.99
Tattu 14.8V 2300mAh 4S 75C LiPo Battery Pack	pack	1	$47.99	$47.99
Fpv Drone Props Propelle 51477 Tri-Blade 5 Inch	pack	1	$19.99	$19.99
MP1584EN 3A Mini DC-DC Buck 5V Voltage Regulator	pack	1	$8.69	$8.69
Intel RealSense Depth Camera D435	set	1	$307.41	$307.41
Multicopter Propeller Guards Prop Protector	pack	1	$17.99	$17.99
2PCS Smoke Stopper for FPV Drone,Short-Circuit Protection	pack	1	$15.99	$15.99
M3x8mm Round Aluminum Standoff Column Spacer	pack	1	$7.89	$7.89
Aluminum Spacer Posts M3 x 10mm	pack	1	$8.99	$8.99
Aluminum Spacer Posts M3 x 15mm	pack	1	$8.99	$8.99
Aluminum Spacer Posts M3 x 20mm	pack	1	$9.99	$9.99
10cm Breadboard Jumper Wires Assorted Kit	pack	1	$6.98	$6.98
15cm Breadboard Jumper Wires Assorted Kit	pack	1	$7.49	$7.49
Double Sided Tape 1in x 16.5ft, Mounting Tape Heavy Duty	tape	1	$17.99	$17.99
M3-0.5 x 6mm Flat Head Socket Cap Screws Bolts	pack	1	$6.99	$6.99
14AWG Flexible Extension Cord	pack	1	$18.99	$18.99
18AWG 2 Conductors Flexible Wire	pack	1	$15.99	$15.99
20AWG OFC 12V/24V DC Extension Wire	pack	1	$17.99	$17.99
120fps USB Camera Module	piece	1	$106.99	$106.99
			Total Price (All) :	$1,967.15