ISSN: 3079-644X

Research on Multi-sensor Fusion 3D Object Detection via PMTV-RCNN for Quadruped Robot in Railway Maintenance

Taotao Li[1], Jingqi Lin[1], Yuhang Cai[1], Zihan Liu[1], Tianqi Yao[1]

Wuhan Railway Vocational College of Technology, Wuhan 430000

Corresponding Email: China468799798@qq.com

Abstract

Accurate 3D object detection is critical for quadruped robots in railway maintenance. Single LiDAR suffers from sparse point clouds and lack of texture information, while cameras are sensitive to lighting conditions. This paper proposes a multi-modal fusion algorithm PMTV-RCNN that deeply integrates LiDAR point clouds and camera images. The algorithm consists of three key modules: Voxel Transformer for efficient voxel feature extraction, adaptive key point selection to highlight discriminative features, and multi-feature aggregation network for weighted fusion of color and geometric information. Experiments on KITTI and railway field datasets demonstrate that PMTV-RCNN achieves significant improvements over baseline PV-RCNN, with AP gains of 2.5%, 18.06%, and 12.11% for railway equipment/vehicles, pedestrians, and foreign objects under medium difficulty. Field experiments on a quadruped robot verify its robustness in complex railway scenarios.

Keywords

railway maintenance; quadruped robots; multi-sensor fusion; LiDAR; camera; 3D object detection; PMTV-RCNN

1. Introduction

1.1 Challenges of Railway Maintenance and Need for Multi-sensor Fusion

Railway maintenance requires accurate identification of various targets along the line, including equipment components (tracks, fasteners, sleepers), foreign objects (branches, stones, plastic sheets), and persons illegally entering the line. Single sensor has limitations: LiDAR provides accurate 3D geometric information but lacks texture and color, and point clouds become sparse for distant small targets. Camera provides rich texture but is sensitive to lighting and cannot provide accurate depth. Multi-sensor fusion can integrate the advantages of both, improving detection accuracy and robustness.

1.2 Related Work on Multi-sensor Fusion 3D Detection

Recent years have seen many fusion algorithms, such as PointAugmenting, EPNet, DeepFusion, BEVFusion, and LoGoNet. These methods fuse LiDAR and camera at different levels: data-level, feature-level, or decision-level. Feature-level fusion is mainstream. However, existing algorithms still have problems: insufficient fusion, low accuracy for long-distance small targets, and poor adaptability to railway scenarios.

1.3 Our Approach and Contributions

This paper proposes PMTV-RCNN, a two-stage multi-modal fusion network. The main contributions are:

Voxel Transformer: Efficiently extracts voxel features using hash table to query non-empty voxels, enhancing feature representation.

ISSN: 3079-644X

Adaptive Key Point Selection: Learns key points that best differentiate similar objects, suppressing background interference.

Multi-feature Aggregation Network: Fuses camera color information with point cloud features via weighted fusion and further extracts features using point cloud Transformer.

Extensive experiments on KITTI and railway datasets demonstrate superior performance. Field experiments on a quadruped robot validate the system's robustness.

The rest of this paper is organized as follows: Section 2 introduces the robot platform and sensor calibration. Section 3 details the PMTV-RCNN algorithm. Section 4 presents experimental results. Section 5 describes system implementation and field experiments. Section 6 concludes the paper.

2. System Overview and Sensor Calibration

2.1 Quadruped Robot Platform

The hardware platform is based on "Jueying X20" quadruped robot from Hangzhou Yunshen Technology. It has industrial-grade design, IP67 protection, maximum climbing angle 30°, obstacle height 20cm, and speed 1.5m/s. The perception host is Jetson Xavier NX (21 TOPS), motion host is RK3588 (6T NPU). Sensors include LeiShen 16-line LiDAR and Intel D435i camera.

2.2 Sensor Configuration and Installation

LiDAR is installed at the top center, camera 15cm in front, with optical axis parallel to LiDAR central axis. Both are fixed with aluminum alloy bracket to reduce vibration.

2.3 Calibration Results

The joint calibration of LiDAR and camera was performed using the method described in our previous work (the first paper). The obtained intrinsic matrix, extrinsic parameters (rotation matrix R and translation matrix T) achieve reprojection errors less than 0.3 pixels, ensuring accurate spatial alignment for fusion.

3. PMTV-RCNN: Multi-modal Fusion 3D Object Detection

3.1 Problem Analysis

Existing fusion algorithms often use simple feature concatenation or weighted summation, failing to fully exploit the complementarity of LiDAR and camera. For long-distance small targets, LiDAR point clouds are sparse and camera images are blurry, making fusion difficult. Moreover, similar objects (e.g., ballast vs. stones, pedestrians vs. poles) need to be distinguished using both geometric and color information.

3.2 PMTV-RCNN Network Structure

PMTV-RCNN is a two-stage network. Stage 1: LiDAR point cloud is voxelized and processed by Voxel Transformer to extract voxel features. Adaptive key point selection highlights discriminative key points. Camera image is processed by ResNet50 to extract image features. Proposal boxes are generated from BEV features. Stage 2: Multi-feature aggregation network fuses key point features with image features, then refined by voxel set abstraction and detection head

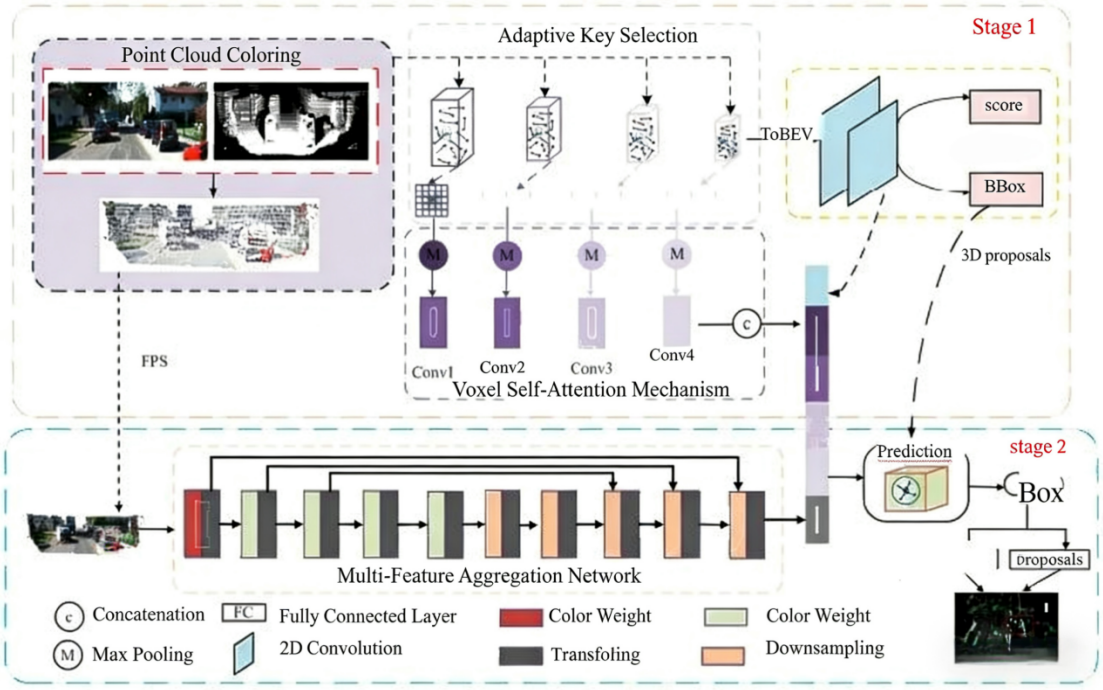

Figure 1: PMTV-RCNN overall structure diagram

3.3 Core Functional Modules

3.3.1 Voxel Transformer Feature Extraction

Voxel Transformer uses a hash table to store non-empty voxels, enabling fast query. Voxel attention mechanism weights non-empty voxel features. Empty voxel features are supplemented by sparse convolution on neighboring non-empty voxels. Multi-scale features are fused via FPN.

$$Q_v=V\cdot W_q^v,\quad K_v=V\cdot W_k^v,\quad V_v=V\cdot W_v^v$$

$$A_v=Softmax\left(\frac{Q_v\cdot K_v^T}{\sqrt{d_{kv}}}\right)$$

$$V_{att}=A_v\cdot V_v$$

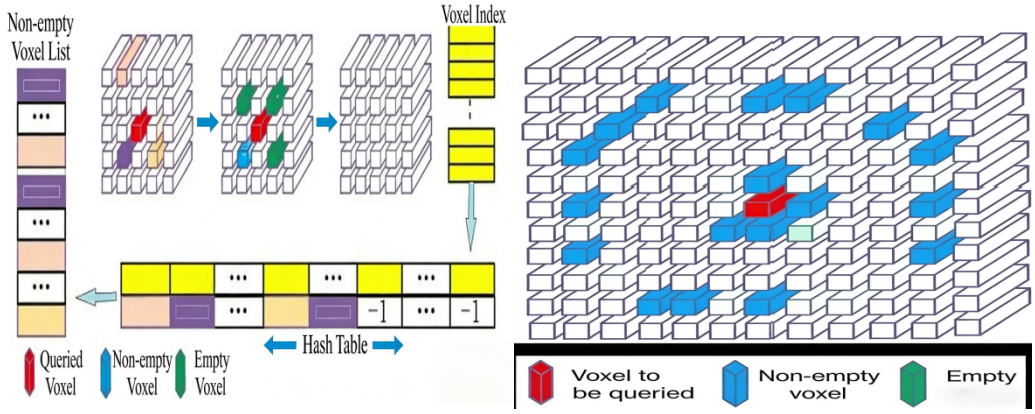

Figure 2: Voxel Transformer network structure

3.3.2 Adaptive Key Point Selection Network

Initial key points are sampled by FPS. Key point features and positions are updated via MLP to learn features that maximize differences between similar objects. Foreground-background segmentation is introduced to suppress background interference.

$$f'=ReLU(W_1 \cdot f+b_1)$$

$$p'=p+\tanh(W_2 \cdot f')$$

$$g=\sigma(W_3 \cdot f'+b_2)$$

$$f_{fg}=g \cdot f'$$

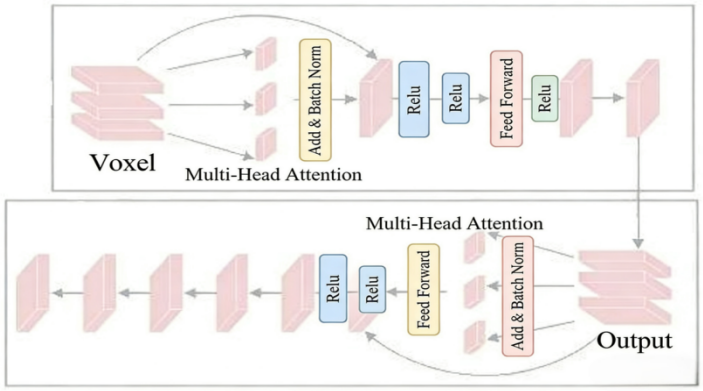

Figure 3: Adaptive key point selection network structure

3.3.3 Multi-feature Aggregation Networks

Camera image features are extracted by ResNet50. Key points are projected onto image plane to obtain corresponding pixel features, which are converted to color weights via MLP. Weighted fusion combines color weights with LiDAR key point features. Point cloud Transformer further extracts fused features with residual downsampling and upsampling.

$$w=\sigma(MLP(F_{img}^{k}))$$

$$F_{fusion1}=f_{fg} \cdot (1+w)$$

$$F_{fusion2}=F_{fusion1}+Conv(F_{up})$$

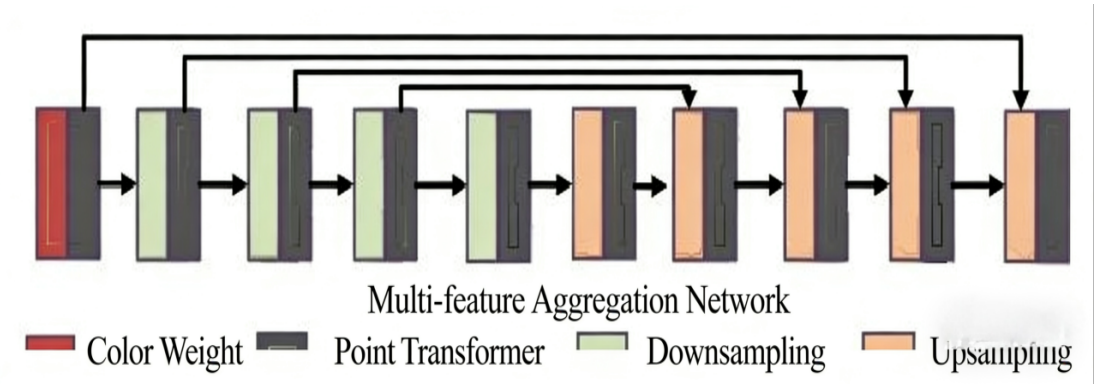

Figure 4: Multi-feature aggregation network structure

ISSN: 3079-644X

3.4 Experimental Results and Analysis

3.4.1 Experimental Setup

Same as first paper, with additional OpenCV4.5.5. Image preprocessing: Resize to 640×360, normalize to [0,1]. Training epochs: 100. Comparison algorithms: PV-RCNN, BEVFusion, DeepFusion, EPNet++.

3.4.2 Model Performance Evaluation

Table 1: Comparison results on KITTI validation set (AP40, %)

Algorithm	Vehicle (Easy/Mod./Hard)	Pedestrian (Easy/Mod./Hard)	Cyclist (Easy/Mod./Hard)	mAP (Mod.)
PV-RCNN	90.25/81.43/76.82	52.17/43.29/40.29	78.60/63.71/57.65	62.81
BEVFusion	91.05/82.15/77.56	56.89/48.36/44.98	80.83/67.66/61.36	66.06
DeepFusion	91.32/82.37/78.12	58.45/50.12/46.53	81.25/68.43/62.15	66.97
EPNet++	91.37/81.96/76.71	52.79/44.38/41.29	76.15/59.71/53.67	62.02
PMTV-RCNN	91.75/83.43/80.09	65.49/58.53/54.46	85.60/66.72/62.56	69.56

Table 2: Comparison results on railway field dataset (AP40, %)

Algorithm	Railway equipment/vehicles (Easy/Mod./Hard)	Pedestrian (Easy/Mod./Hard)	Foreign body (Easy/Mod./Hard)	mAP (Mod.)
PV-RCNN	90.12/82.17/75.67	51.23/43.29/39.87	77.89/63.71/56.89	63.06
BEVFusion	91.56/83.24/76.89	57.65/49.21/45.32	81.56/68.43/60.12	66.96
DeepFusion	92.13/83.56/77.45	59.87/51.34/47.65	82.34/69.78/61.56	67.56
EPNet++	89.67/81.25/75.32	51.89/43.21/40.12	74.36/59.87/52.45	61.44
PMTV-RCNN	93.25/84.67/81.34	69.87/61.35/55.78	91.23/75.82/67.45	73.95

3.4.3 Ablation Experiments

Table 3: Ablation experiment results on railway dataset (Mod. AP40, %)

Configuration	Railway equipment	Pedestrian	Foreign body	mAP
PV-RCNN (baseline)	82.17	43.29	63.71	63.06
+ Voxel Transformer	83.15	59.08	60.25	67.50
+ Adaptive Key Point	83.23	57.98	63.27	68.12
+ Multi-feature Aggregation	84.67	61.35	75.82	73.95

3.4.4 Detection Performance Analysis in Different Scenarios

Table 4: Detection performance in different railway scenes (Mod. AP40, %)

Scene Type	Railway equipment	Pedestrian	Foreign body	mAP
Track (unoccluded)	89.34	65.78	80.45	78.52
Ballast (dense background)	85.67	60.12	74.32	73.37

ISSN: 3079-644X

Bridge (high, open)	87.89	62.34	77.65	75.96
Tunnel (low lighting)	83.21	58.76	72.13	71.37
Occlusion (partial)	81.56	56.43	70.89	69.62

3.4.5 Analysis of Visualization Results

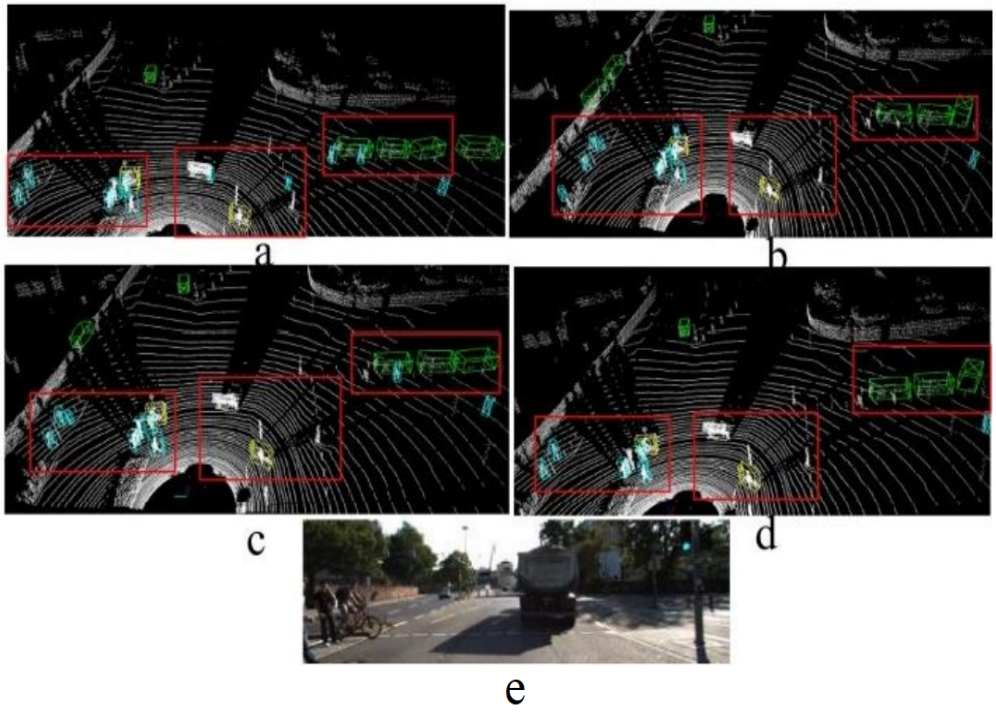

Figure 5: Detection in complex street scenes (crossing with vehicles, pedestrians, foreign objects)

Figure 6: Close-range track equipment detection (fasteners, sleepers)

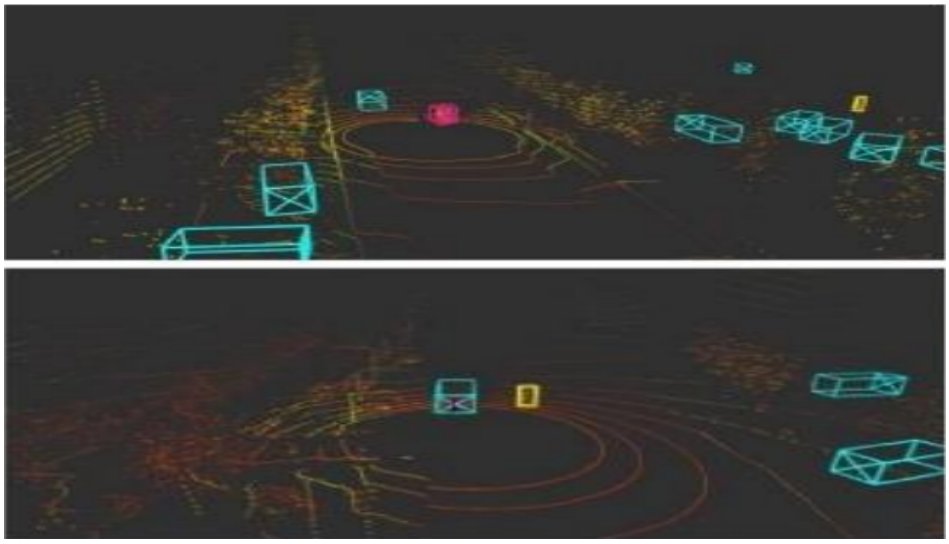

Figure 7: Detection under different lighting conditions (tunnel entrance/exit)

3.5 Summary

This chapter proposed PMTV-RCNN algorithm for multi-modal fusion. Voxel Transformer efficiently extracts voxel features. Adaptive key point selection highlights discriminative features. Multi-feature aggregation network deeply fuses color and geometric information. Experiments show significant improvements over existing fusion algorithms, especially for pedestrians and foreign objects in complex railway scenes.

4. System Implementation and Field Experiments

4.1 Design of Target Detection Hardware System

4.1.1 Hardware Components

Quadruped robot chassis: "Jueying X20", IP67, max speed 1.5m/s.

Perception host: Jetson Xavier NX (6-core ARM CPU, 384-core Volta GPU, 21 TOPS).

Motion host: Rockchip RK3588 (octa-core CPU, 6T NPU).

Sensors: LeiShen 16-line LiDAR, Intel D435i camera, four HC-SR04 ultrasonic radars, IMU.

Table 5: Jetson Xavier NX parameters

Parameter	Value
CPU	6-core NVIDIA Carmel ARMv8.2
GPU	384-core NVIDIA Volta with 48 Tensor cores
DL accelerator	2x NVIDIA DL Accelerator
Memory	8GB 128-bit LPDDR4X
Computing power	21 TOPS (INT8)

Table 6. RK3588 parameters

Parameter	Value
CPU	Octa-core (4 A76 + 4 A55)
GPU	ARM Mali-G610 MP4
NPU	6 TOPS
Memory	4GB LPDDR4X

ISSN: 3079-644X

Figure 8: Perception controller (Jetson Xavier NX) and motion controller (RK3588)

4.1.2 Hardware Connection and Installation

Perception host connects to LiDAR via Ethernet, to camera via USB 3.0, to motion host via Gigabit Ethernet.

Ultrasonic radars connect to motion host via GPIO.

All sensors fixed with aluminum alloy bracket, vibration damping pads.

4.2 Target Detection Software System Design

4.2.1 ROS Communication Mechanism

ROS Melodic with topic-based communication. Core topics: /lidar/points_raw, /camera/color/image_raw, /camera/depth/image_raw, /ultrasonic/obstacle, /detection/result, /robot/pose.

4.2.2 Data Acquisition Module

Sensor drivers start, data preprocessing (point cloud denoising and downsampling, image distortion correction and resize), multi-sensor synchronization (hardware trigger + software timestamp alignment), data storage via rosbag.

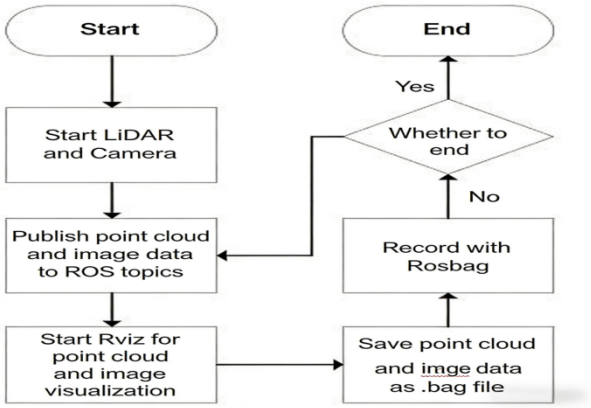

Figure 9: Data recording process flowchart

4.2.3 Algorithm Deployment Module

Trained PMTV-RCNN model is optimized via TensorRT (INT8 quantization) and deployed to perception host. Inference flow: subscribe data → feature extraction → multi-modal fusion → detection head → publish results. Multi-threading and hardware acceleration ensure real-time performance.

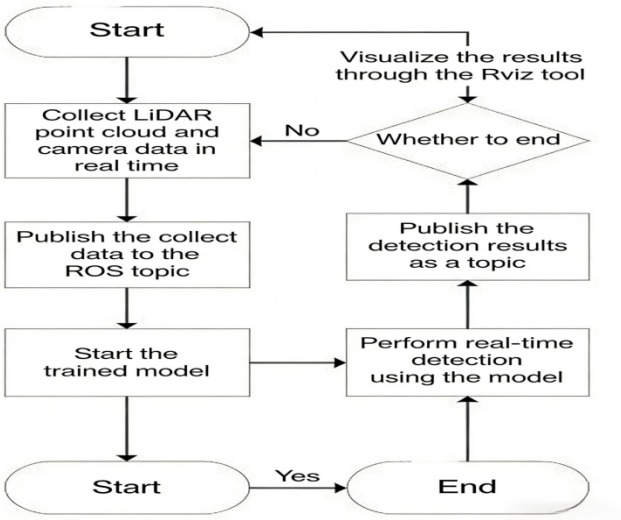

Figure 10: Real-time detection flowchart

4.2.4 Result Visualization Module

Rviz tool visualizes LiDAR point cloud, camera image, 3D detection boxes, robot pose, and sensor status.

4.3 Railway Field Experiment Verification

4.3.1 Experimental Scenario and Protocol

Experiments conducted on a railway section including tracks, ballast, bridge, tunnel. Robot speed 0.5-1.0 m/s. Targets: railway vehicles, fasteners, sleepers, pedestrians (simulated), foreign objects (branches, stones, plastic sheets). Detection results compared with manual labels.

4.3.2 Experimental Results and Analysis

Detection Accuracy:

Table 7: Detection accuracy in field experiment (AP40, %)

Target Category	Easy	Mod.	Hard
Railway equipment/vehicles	93.25	84.67	81.34
Pedestrian	69.87	61.35	55.78
Foreign body	91.23	75.82	67.45
mAP (Mod.)	-	73.95	-

Real-time Performance:

Table 8: Real-time latency (ms)

Index	Average	Max	Min
Data acquisition delay	15.2	22.3	10.5
Algorithm inference delay	89.7	112.4	75.3
Overall delay	104.9	134.7	85.8

Robustness Analysis:

Table 9: Robustness under different conditions (mAP, Mod., %)

Scene/Weather	Sunny day	Cloudy day	Evening (low light)	Average

Track	78.52	77.34	75.67	77.18
Ballast	73.37	72.15	70.89	72.14
Bridge	75.96	74.87	73.45	74.76
Tunnel	71.37	70.12	68.98	70.16

Visualization Results:

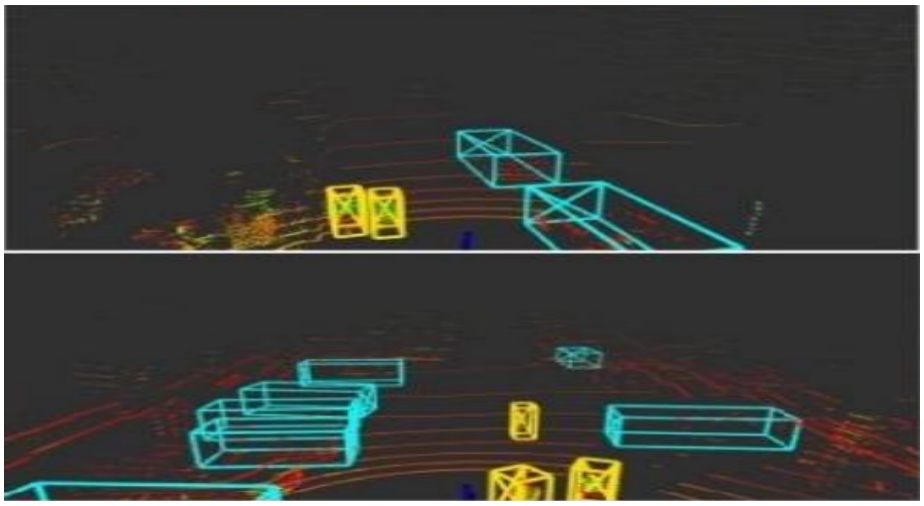

Figure 11: Field experiment detection results (multiple scenes)

4.4 Summary

This chapter presented the hardware and software implementation of the target detection system for railway maintenance quadruped robot. Field experiments verified that the system achieves high accuracy (mAP 73.95% Mod.), real-time performance (average delay 104.9ms), and robustness in complex railway scenarios.

5. Conclusion and Future Work

This paper proposed PMTV-RCNN, a multi-modal fusion 3D object detection algorithm for quadruped robots in railway maintenance. The algorithm integrates LiDAR and camera data through Voxel Transformer, adaptive key point selection, and multi-feature aggregation network. Extensive experiments on KITTI and railway datasets show significant improvements over baseline. Field experiments on a quadruped robot demonstrate the system's practical applicability.

Future work includes: (1) incorporating infrared thermal imaging for temperature anomaly detection; (2) improving performance in extreme weather (rain, snow, dust); (3) developing fault warning module and path planning for fully autonomous inspection; (4) multi-robot collaborative inspection.

References

[1] Shi S, Guo C, Jiang L, et al. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection[C]//CVPR. 2020: 10529-10538.

[2] Liu Z, Chen X, Gao S, et al. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework[C]//CVPR. 2023: 17923-17932.

[3] Li Y, Yu A W, Meng T, et al. DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection[C]//CVPR. 2022: 17182-17191.

ISSN: 3079-644X

[4] Huang Y, Huang H, Zhu X, et al. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection[C]//CVPR. 2020: 11024-11033.

[5] Wang Y, Solomon E. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection[C]//CVPR. 2020: 11003-11012.

[6] Li Y, Chen X, Shen S, et al. LoGoNet: Local-Global Fusion Network for LiDAR-Camera 3D Object Detection[C]//ICCV. 2021: 11134-11143.

[7] Jiao J, Zhang Y, Jiang H, et al. MSMD Fusion: Multi-Scale Multi-Depth Fusion for LiDAR-Camera 3D Object Detection[C]//CVPR. 2023: 17943-17952.

[8] Wang S, Liu Y, Wang T, et al. Exploring object-centric temporal modeling for efficient multi-view 3d object detection[C]//ICCV. 2023: 3621-3631.

[9] Park J, Xu C, Yang S, et al. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection[OL]. arXiv preprint arXiv:2210.02443, 2022.

[10] Lin X, Lin T, Pei Z, et al. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion[OL]. arXiv preprint arXiv:2211.10581, 2022.

[11] Wang L, Li R, Sun J, et al. Multi-view fusion-based 3D object detection for robot indoor scene perception[J]. Sensors, 2019, 19(19): 4092.

[12] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//CVPR. 2016: 770-778.

[13] Girshick R. Fast R-CNN[C]//ICCV. 2015: 1440-1448.

[14] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//CVPR. 2016: 779-788.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]//NIPS. 2015: 91-99.

[16] Chen G, Hong L. Research on environment perception system of quadruped robots based on lidar and vision[J]. Drones, 2023, 7(5): 329.

[17] LYU Y, JIA Y, ZHUANG Y, et al. Obstacle avoidance approach for quadruped robot based on multi-modal information fusion[J]. Chinese Journal of Engineering, 2024, 46(8): 1426-1433.

[18] Gao F, Tang W, Huang J, et al. Positioning of quadruped robot based on tightly coupled LiDAR vision inertial odometer[J]. Remote Sensing, 2022, 14(12): 2945