ISSN: 3079-6342

Research on 3D Object Detection for Quadruped Robot in Railway Maintenance Based on Improved PV-RCNN

Taotao Li[1], Dianyun Luo[1], Ruixiang Liu[1], Yuhang Cai[1],
Yunjia Chen[1], Shenghui Zheng[1]
[1]Wuhan Railway Vocational College of Technology, Wuhan 430000

Abstract

As the core infrastructure of national comprehensive transportation system, the safe operation of railway lines is crucial. Quadruped robots have become ideal carriers for railway autonomous inspection due to their excellent terrain adaptability. High-precision environmental perception is the core prerequisite for their autonomous operation. Aiming at the defects of traditional PV-RCNN algorithm in point cloud feature extraction for railway scenarios, this paper proposes an improved TS-PV-RCNN algorithm for 3D object detection using LiDAR point cloud. By introducing transform-equivariant strategy, dual attention mechanism, and Transformer feature extraction module, the feature extraction effect is optimized. Experiments on KITTI dataset and railway field dataset show that compared with the benchmark PV-RCNN algorithm, the average precision (AP) of railway equipment/vehicles, pedestrians, and foreign objects is improved by 1.35%, 16.83%, and 9.24% respectively under medium difficulty. The proposed algorithm provides a feasible technical scheme for autonomous inspection of railway lines.

Keywords

railway maintenance; quadruped robots; LiDAR; 3D object detection; deep learning; TS-PV-RCNN

1. Introduction

1.1 Research Background and Significance

Railway, as the backbone of China's comprehensive transportation system, has significant advantages such as large transportation volume, fast speed, low energy consumption and high safety. By the end of 2024, the operating mileage of national railways has reached 168,000 kilometers, forming a nationwide railway network. With the continuous extension of railway lines and the increase of operation speed, the safety guarantee of railway lines faces unprecedented challenges. Railway lines are exposed to complex outdoor environment for a long time, leading to hidden dangers such as track deformation, fastener loosening, ballast scattering, foreign object intrusion, etc. Traditional manual inspection has problems of low efficiency, high missed inspection rate and high operation risk. Therefore, developing autonomous inspection robots with high-precision environmental perception ability has become an urgent need.

1.2 Quadruped Robot in Railway Maintenance

Quadruped robots, imitating the movement mechanism of quadruped animals, have excellent terrain adaptability and motion flexibility. They can move stably in unstructured terrain such as railway ballast, bridge steps, and tunnel cracks, effectively overcoming the limitations of wheeled and tracked robots. With the maturity of quadruped robot technology, its application

ISSN: 3079-6342

in industrial inspection has gradually emerged. Applying quadruped robots to railway maintenance can greatly improve inspection efficiency and reduce operation risks.

1.3 LiDAR-based 3D Object Detection

Three-dimensional target detection based on LiDAR point cloud is a mainstream technology for mobile robot environment perception. According to point cloud processing methods, algorithms can be divided into view-based, voxel-based, point-based, and point-voxel fusion methods. PV-RCNN is a representative point-voxel fusion algorithm, which combines the computational efficiency of voxel method and the detection accuracy of point cloud method. However, when applied to railway maintenance scenarios, PV-RCNN suffers from problems such as pseudo-2D image rotation transformation, feature loss, and insufficient receptive field, leading to low detection accuracy for small targets like track fasteners and distant foreign objects.

1.4 Main Contributions

To address these issues, this paper proposes an improved TS-PV-RCNN algorithm. The main contributions are:

A transform-equivariant strategy integrating attention mechanism is designed to solve the rotation transformation problem of BEV features and compensate for feature loss.

A point cloud feature extraction module based on Transformer is proposed to expand the receptive field and fully extract global and local features.

Extensive experiments on KITTI and railway field datasets verify the effectiveness of the proposed algorithm, achieving significant improvements over baseline PV-RCNN.

The rest of this paper is organized as follows: Section 2 introduces the joint calibration of LiDAR and camera. Section 3 details the TS-PV-RCNN algorithm. Section 4 presents experimental results and analysis. Section 5 concludes the paper.

2. Joint Calibration of LiDAR and Camera

2.1 Introduction

Multi-sensor joint calibration is the premise for fusing LiDAR and camera. It includes time synchronization and spatial registration. Only with accurate calibration can the 3D geometric information of LiDAR be effectively integrated with the texture information of camera. In railway maintenance, the quadruped robot needs to collect sensor data in real time, and the accuracy of calibration directly affects target detection performance.

2.2 Overall Scheme Design

The overall scheme of the 3D target detection system for railway maintenance quadruped robot is divided into three stages: sensor joint calibration, algorithm research and training, robot platform deployment and on-site verification. Sensor joint calibration is the foundation.

2.3 Sensor Working Principle

2.3.1 LiDAR Working Principle

LiDAR (Light Detection and Ranging) is an active ranging sensor that measures distance by emitting laser pulses and calculating time-of-flight. The distance formula is:

$$d = \frac{c \times t}{2}$$

where (d) is distance, (c) is speed of light, (t) is flight time.

Multi-line LiDAR can obtain 3D point cloud data. This paper selects LeiShen 16-line mechanical rotary LiDAR, with parameters shown in Table 1.

Table 1: Technical parameters of LeiShen 16-line LiDAR

ISSN: 3079-6342

Parameter	Value
Number of channels	16
Horizontal scan range	360°
Vertical pitch angle	-15°~+15°
Ranging accuracy	±1cm
Max detection distance	150m (@70% reflectivity)
Point rate	320,000 pts/s
Protection rating	IP67

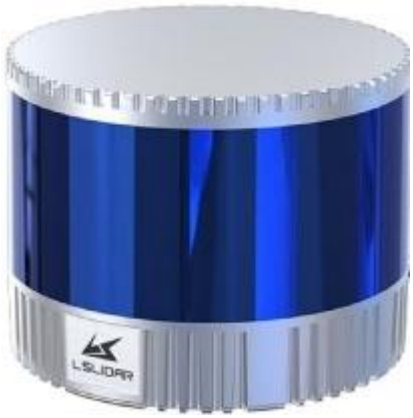


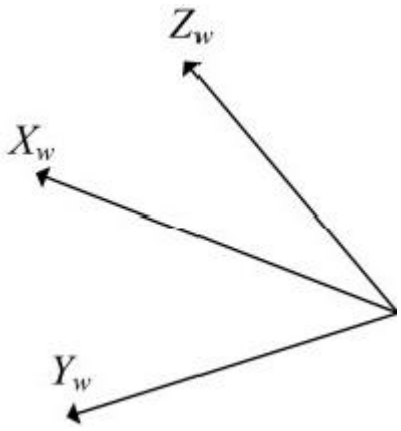
Figure 1: LeiShen LiDAR physical image

2.3.2 How the Camera Works

Camera is a passive vision sensor that converts optical signals into digital images. This paper selects Intel RealSense D435i depth camera, with parameters shown in Table 2.

Table 2: Intel RealSense D435i technical parameters

Parameter	Value
Depth range	0.2m~10m
Color image resolution	1920×1080
Color frame rate	30fps
Depth image resolution	1280×720
Depth frame rate	90fps
Field of view	~86°×57°
IMU	Accelerometer, gyroscope



ISSN: 3079-6342

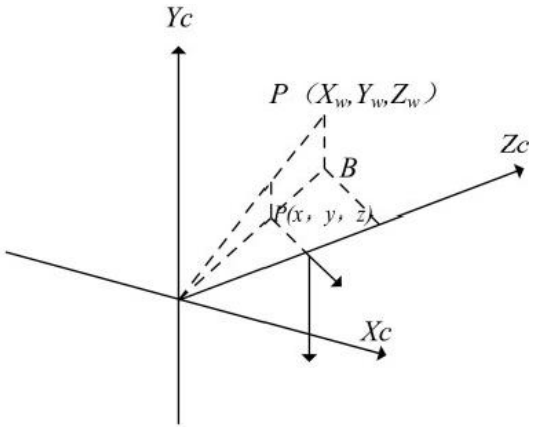


Figure 2: Camera coordinate system transformation diagram

2.4 Sensor Hardware Selection and Installation

2.4.1 Hardware Selection Rationale

The selection follows principles of detection accuracy, environmental adaptability, detection range, real-time requirement, system integration, cost and volume. LeiShen 16-line LiDAR and Intel D435i camera meet all requirements.

2.4.2 Sensor Installation Scheme

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

Figure 3: Lidar and camera installation positions

2.5 Joint Calibration of LiDAR and Camera

2.5.1 Principle of Joint Calibration

Camera internal calibration solves the intrinsic matrix (K) and distortion coefficients. The intrinsic matrix is:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

LiDAR-camera extrinsic calibration solves rotation matrix (R) and translation matrix (T):

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = R \begin{bmatrix} X_L \\ Y_L \\ Z_L \end{bmatrix} + T$$

ISSN: 3079-6342

2.5.2 Calibration Preparation

Checkerboard calibration board (12×9 grids, grid size 42mm×42mm) is used. Calibration environment is indoor with sufficient light.

2.5.3 Calibration Process

Camera internal calibration uses Zhang's method, collecting 15-20 checkerboard images with different poses. LiDAR-camera external calibration uses calibration_camera_lidar tool, recording bag files with calibration board at different distances and poses.

2.5.4 Calibration Results

Camera intrinsic matrix:

$$K = \begin{bmatrix} 915.62 & 0 & 640.32 \\ 0 & 914.87 & 359.78 \\ 0 & 0 & 1 \end{bmatrix}$$

Distortion coefficients: (k_1=-0.0521, k_2=0.0834, k_3=-0.0312, p_1=0.0012, p_2=0.0008)

Reprojection error: 0.21 pixels

Rotation matrix (R) and translation matrix (T):

$$R = \begin{bmatrix} 0.0910 & 0.1276 & 0.9876 \\ -0.0973 & -0.0508 & 0.0982 \\ 0.0627 & -0.9905 & 0.1222 \end{bmatrix}, \quad T = \begin{bmatrix} -0.0973 & 0.0246 & -0.2953 \end{bmatrix}^T$$

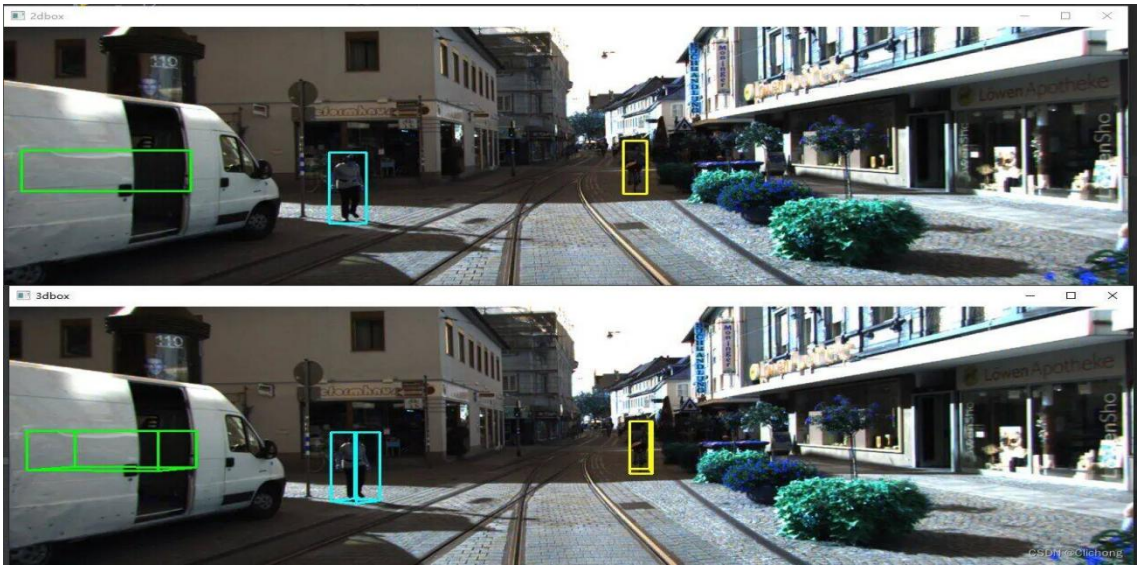
Reprojection error: 0.226 pixels

2.6 Datasets and Evaluation Indicators

2.6.1 Introduction to Datasets

KITTI dataset: 7481 training samples, 7518 test samples, with labels for cars, pedestrians, cyclists. LiDAR is Velodyne HDL-64E (64-line), camera resolution 1242×375.

Railway field dataset: Collected using "Jueying X20" quadruped robot with LeiShen 16-line LiDAR and Intel D435i camera, containing 500 samples (350 train, 150 validation) with labels for railway equipment/vehicles, pedestrians, foreign objects.



ISSN: 3079-6342

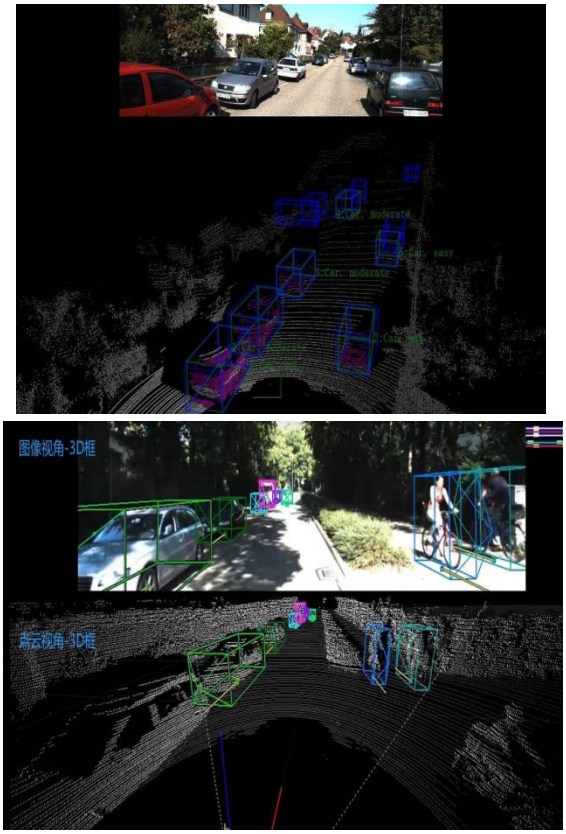


Figure 5: KITTI LiDAR visualization

2.6.2 Evaluation Indicators

IoU: Intersection over Union for 2D, BEV, and 3D.

Precision (P) and Recall (R).

Average Precision (AP): area under PR curve, using 40 recall points.

Average Orientation Similarity (AOS): measures orientation consistency.

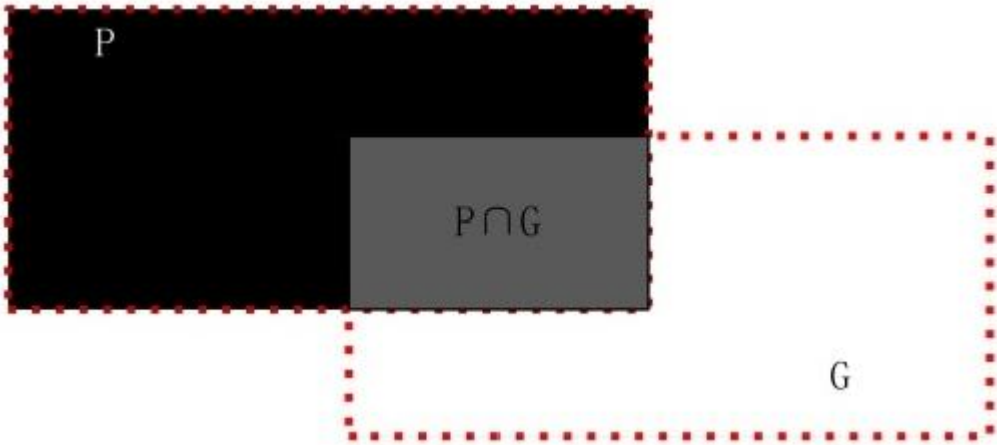
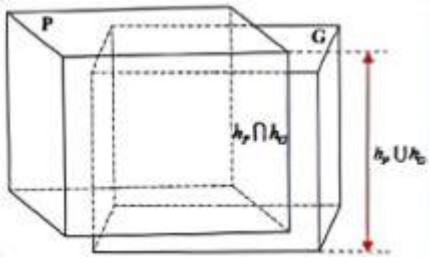


Figure 6: IoU calculation diagram



ISSN: 3079-6342

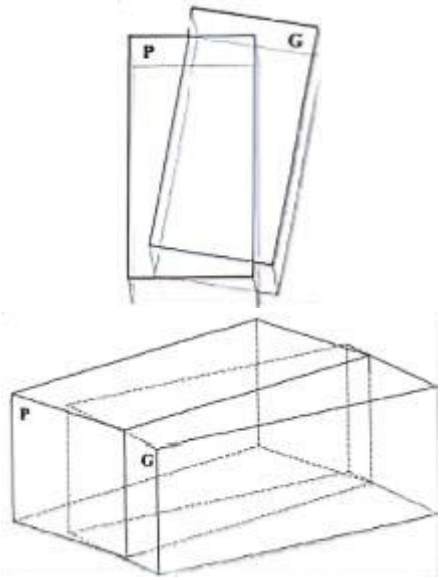


Figure 7: 3D bounding box schematic

2.7 Summary

This chapter completed the joint calibration of LiDAR and camera, achieving high-precision spatial registration with reprojection errors less than 0.3 pixels, laying foundation for subsequent algorithm research.

3. LiDAR Point Cloud Target Detection Algorithm Based on Improved PV-RCNN

3.1 Problem Analysis

When applying PV-RCNN to railway maintenance scenarios, three main problems exist:

Pseudo-2D image rotation transformation and feature loss: Compressing voxel features to BEV causes feature distortion and loss.

Insufficient point cloud feature extraction: Traditional methods have limited ability to capture details of irregular or sparse targets.

Insufficient receptive field: Long-distance small targets are hard to detect due to limited receptive field.

3.2 TS-PV-RCNN Algorithm Framework

TS-PV-RCNN is a two-stage algorithm. The first stage performs voxelization, sparse convolution, transform-equivariant BEV feature generation with dual attention, and proposal generation. The second stage samples key points, extracts features using Transformer module, aggregates voxel features, performs RoI pooling, and outputs final detection.

ISSN: 3079-6342

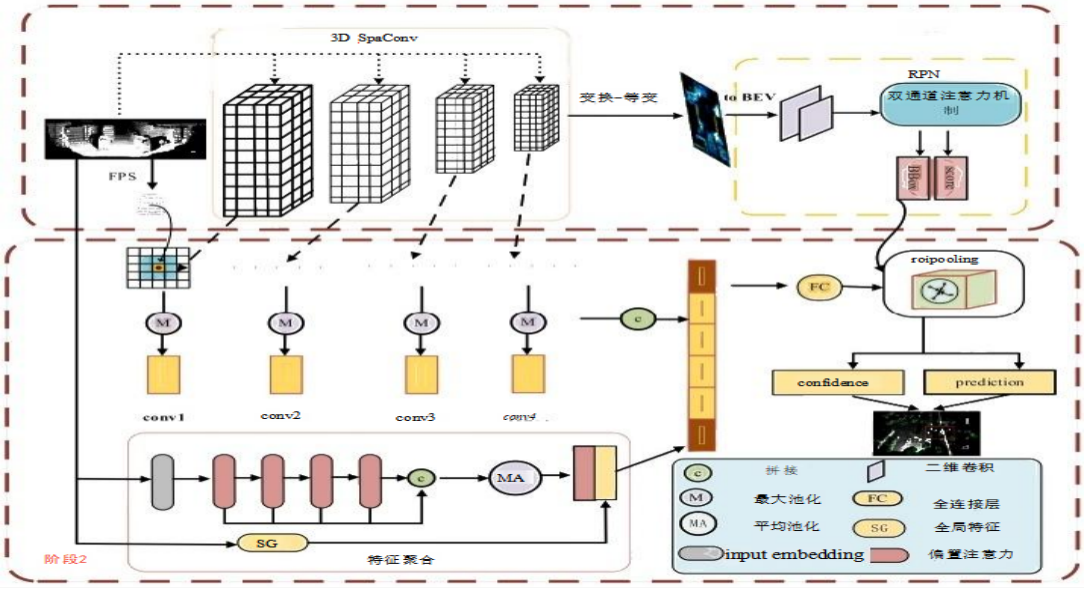


Figure 8: TS-PV-RCNN flowchart

3.3 Core Improvement Module

3.3.1 Transform-Equivariant Strategy with Attention Mechanism

The transform-equivariant strategy aligns BEV features under different rotations using bilinear interpolation and max pooling. The dual attention mechanism (PSA) includes channel self-attention and spatial self-attention to enhance important features.

Channel self-attention:

$$A_{ch}(A^) = \sigma(W_2 \cdot \text{Softmax}(W_1 \cdot A^) \cdot W_q \cdot A^*)$$

Spatial self-attention:

$$A_{sp}(A^) = \sigma(\text{GlobalPool}(W_q \cdot A^) \cdot W_k \cdot A^*)$$

Final attention feature:

$$A_{psa} = A^* \oplus A_{ch}(A^) \oplus A_{sp}(A^)$$

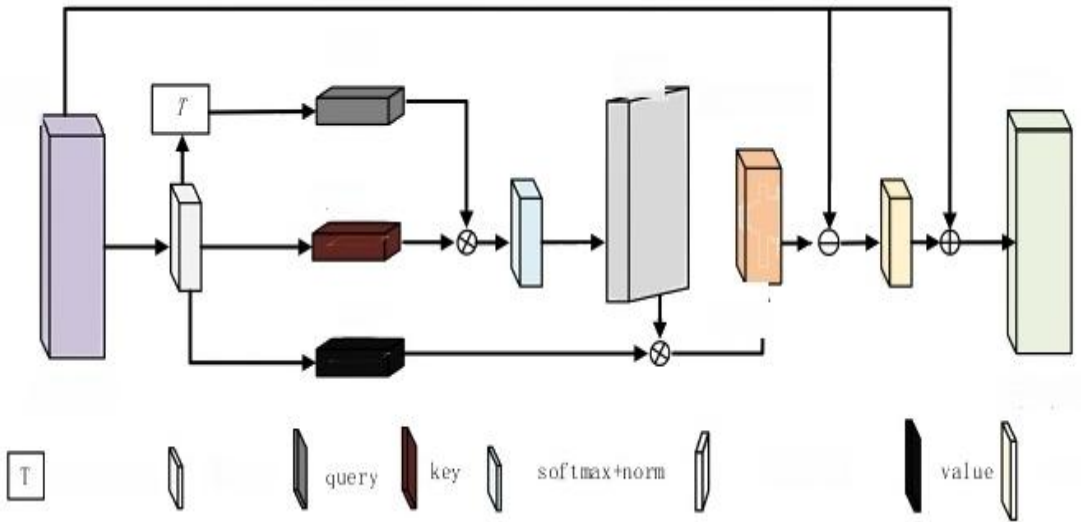


Figure 9: Transform-equivariant strategy with attention

3.3.2 Point Cloud Feature Extraction Module Based on Transformer

The module consists of point cloud embedding (two-layer LBR), biased attention mechanism, and local neighborhood feature extraction (two sampling grouping layers). Biased attention incorporates local neighborhood information to enhance local correlation.

$$Q = F_{embed} \cdot W_q, \quad K = F_{embed} \cdot W_k, \quad V = F_{embed} \cdot W_v$$

ISSN: 3079-6342

$$A = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot \text{Norm}(L)$$
$$F_{sa} = A \cdot V$$

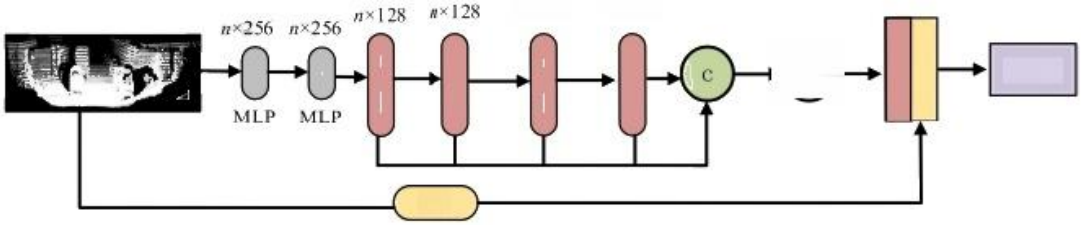


Figure 10: Transformer-based point cloud feature extraction module

3.4 Experimental Results and Analysis

3.4.1 Experimental Setup

Hardware: Intel Core i7-12700F, NVIDIA RTX3080 (10GB), 32GB RAM.

Software: Ubuntu 18.04, PyTorch 1.11.0, CUDA 11.1.

Parameters: voxel size [0.05,0.05,0.1]m, learning rate 0.01, batch size 1, 80 epochs.

Comparison algorithms: PointPillar, 3DSSD, EPNet++, PV-RCNN.

3.4.2 Model Performance Evaluation

Table 3: Comparison results on KITTI validation set (AP40, %)

Algorithm	Vehicle (Easy/Mod./Hard)	Pedestrian (Easy/Mod./Hard)	Cyclist (Easy/Mod./Hard)	mAP (Mod.)
PointPillar	79.05/74.99/68.30	52.08/43.53/41.49	75.78/59.07/52.92	59.20
3DSSD	88.36/79.57/74.55	54.64/44.27/40.23	82.48/64.10/56.90	62.63
EPNet++	91.37/81.96/76.71	52.79/44.38/41.29	76.15/59.71/53.67	62.02
PV-RCNN	90.25/81.43/76.82	52.17/43.29/40.29	78.60/63.71/57.65	62.81
TS-PV-RCNN	91.62/82.70/88.52	67.24/59.07/53.62	89.64/71.83/67.28	67.87

Table 4: 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.)
PointPillar	78.21/73.45/66.89	50.12/41.35/38.76	73.25/58.67/51.32	57.82
3DSSD	86.54/78.32/72.45	52.36/42.18/39.54	80.12/63.45/55.67	61.32
EPNet++	89.67/81.25/75.32	51.89/43.21/40.12	74.36/59.87/52.45	61.44
PV-RCNN	90.12/82.17/75.67	51.23/43.29/39.87	77.89/63.71/56.89	63.06
TS-PV-RCNN	92.34/83.52/87.89	68.45/60.12/52.98	90.12/72.95/66.34	68.86

3.4.3 Ablation Experiments

Table 5: Ablation experiment results on KITTI (Mod. AP40, %)

Configuration	Vehicle	Pedestrian	Cyclist	mAP
PV-RCNN (baseline)	81.43	43.29	63.71	62.81
+ Transform-Equivariant	82.15	52.80	67.31	67.42
+ PSA	82.69	54.87	70.52	69.36
+ Transformer	82.70	59.07	71.83	67.87

ISSN: 3079-6342

3.4.4 Visualization Results Analysis

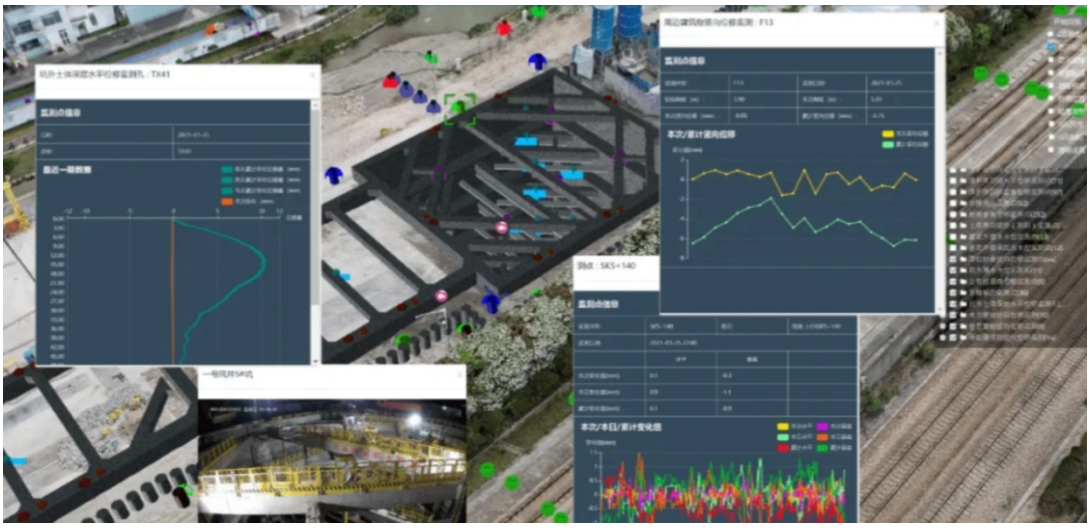


Figure 11: Detection results visualization in railway scene (point cloud and bounding boxes)

3.5 Summary

This chapter proposed TS-PV-RCNN algorithm for LiDAR point cloud detection. The transform-equivariant strategy with attention mechanism solves rotation transformation and feature loss. The Transformer-based module expands receptive field and improves feature extraction. Experiments show significant improvements over baseline PV-RCNN, especially for pedestrians and foreign objects.

4. Conclusion and Future Work

This paper addresses the problem of 3D object detection for quadruped robots in railway maintenance using LiDAR point cloud. We improved PV-RCNN by introducing transform-equivariant strategy, dual attention, and Transformer feature extraction, resulting in TS-PV-RCNN. Experimental results on KITTI and railway field datasets demonstrate the effectiveness of the proposed method.

Future work includes: (1) further improving detection accuracy for extremely small targets; (2) optimizing the algorithm for real-time performance on embedded platforms; (3) exploring the integration with camera data to leverage multi-modal information.

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] Zhou Y, Tuzel O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection[C]//CVPR. 2018: 4490-4499.

[3] Qi C R, Su H, Mo K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]//CVPR. 2017: 652-660.

[4] Qi C R, Yi L, Su H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]//NIPS. 2017: 5099-5108.

[5] Yang Z, Sun Y, Liu S, et al. 3DSSD: Point-Based 3D Single Stage Detector[C]//CVPR. 2020: 11044-11053.

[6] Liang M, Yang B, Wang S, et al. Range R-CNN: Detector for 3D LiDAR Point Clouds[C]//CVPR. 2020: 11108-11117.

ISSN: 3079-6342

[7] Fan B, Yang Z, Zhu S, et al. RangeDet: Efficient 3D Object Detection from LiDAR Range Images[C]//CVPR. 2022: 16216-16225.

[8] Ye H, Lee S, Kim J, et al. HVNet: Hybrid Voxel Network for LiDAR Based 3D Object Detection[C]//ICCV. 2021: 11094-11103.

[9] Kuang Y, Zhang Y, Chen Y, et al. Voxel-FPN: Multi-scale Voxel Feature Aggregation for 3D Object Detection[C]//AAAI. 2020, 34(07): 11987-11994.

[10] Shi S, Wang X, Li H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud[C]//CVPR. 2019: 770-779.

[11] Mao J, Jiang H, Qian C, et al. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection[C]//CVPR. 2021: 13013-13022.

[12] Noh H, Lee Y, Whang J, et al. HVPR: Hybrid Voxel-Point Representation for 3D Object Detection[C]//ICCV. 2021: 11114-11123.

[13] 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.

[14] Li B, Ouyang W, Sheng L, et al. Gs3d: An efficient 3d object detection framework for autonomous driving[C]//CVPR. 2019: 1019-1028.

[15] Wang T, Zhu X, Pang J, et al. Fcos3d: Fully convolutional one-stage monocular 3d object detection[C]//ICCV. 2021: 913-922.

[16] Luo S, Dai H, Shao L, et al. M3dssd: Monocular 3d single stage object detector[C]//CVPR. 2021: 6145-6154.