ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on a Three-Dimensional Object Detection Algorithm Based on Camera and 4D Millimeter-Wave Radar Fusion

Zhikui Lu[1] *and Qimin Xu*[1,*], *Benwu Wang*[1] *and He Zhu*[1]

[1] School of Instrument Science and Engineering, Southeast University, 210096, Nanjing, China

SUMMARY: *In order to solve the problem of low precision and shortage of space information in single sensor system, a new method of 3D object detection based on 4D MMW radar is presented. The proposed method combines multi-modal feature fusion and spatiotemporal scaling, and adopts a middle level fusion strategy to model multi-element features of radar point cloud. The camera branch uses the ResNet-FPN architecture to extract multiscale semantic features, and the radar branch uses a VoxelNet-based compression structure to improve the performance of the algorithm. Experiments on the 4D-Drive and nuScenes data sets show that the proposed algorithm achieves 86.4% detection accuracy when the IoU threshold is 0.7. Compared with the single-modality baseline, there is a 12.8% increase in performance compared to a camera only mode, and 9.5% in the radar-only mode. Compared with the monocular vision approach, the absolute displacement error has been decreased to 0.184 m, and the detection accuracy remains above 85% under challenging conditions, showing strong robustness and generalization ability. The system achieves real time performance at 31 FPS on the RTX A6000 and Jetson AGX Orin platforms. This study has overcome the shortcomings of the existing detection algorithms, such as precision, latency, and environment adaptability, and provides an efficient way to realize multi-source collaboration.*

KEYWORDS: *3D object detection; Multimodal fusion; 4D millimeter-wave radar; Deep learning*

1 Introduction

In applications such as autonomous driving, intelligent surveillance and industrial detection, three-dimensional target detection plays an important role.Compared with other sensors, cameras have many advantages in target identification, such as high resolution and rich texture. However, when cameras are used in scenarios with interferences, such as low-light and haze, the localization accuracy of targets will be affected by interferences. Commillimeter-wave radar has advantages of resistance to interferences and ranging performance, which can collect three-dimensional information of targets in low-visibility scenarios, and the application of millimeter-wave four-dimensional radar can obtain target information in low-visibility scenarios, which provides basis for three-dimensional structural information. However, due to the characteristics of radar point clouds, such as sparsity and limited angular resolution, it is difficult to play its advantages when used alone. Therefore, the development direction of camera and four-dimensional millimeter-wave radar multi-modal

*18851651959@163. com
https://doi.org/10.65102/is20261195 Published: 10 June, 2026

perception is very important for improving the accuracy and robustness of spatial detection.

Karen Liliane Riedel Hornig et al. (2025) [1] They used vibration isolation testing to explore problems related to camera stabilization in multi-view imaging and found that dynamic jitter affects the target detection process in the following link.Hu Liu et al. (2025) [2] It can eliminate the fake reflection points of radar point clouds and improve the observation effects. Chunfang Yin et al. (2025) [3] Establish a multi-scale weight fusion model to perform effective matching of point cloud data. Siyuan Wei et al. (2025) [4] designed the ST-ConvLSTM network for 3D human keypoint localization under millimeter-wave signals. Qingyuan Yang et al. (2025) [5] reviewed research progress in visual-millimeter wave fusion for pedestrian re-identification, identifying cross-modal matching and feature alignment as current bottlenecks. Zhimin Qiu et al. (2025) [6] achieved road surface recognition through multi-feature fusion, validating the fusion features' interference resistance. Ye Chen et al. (2025) [7] employ an improved extended Kalman filter for joint ranging between monocular cameras and millimeter-wave radar, enhancing depth estimation accuracy in dynamic scenes. Weigang Shi et al. (2025) [8] utilize mobile least squares to enhance the spatial fitting capability of 4D radar point clouds. Peicheng Shi et al. (2025) [9] proposed the MPVF algorithm to enhance global consistency in multimodal detection through point-level and voxel-level fusion. Rui Zheng et al. (2024) [10] developed a fusion positioning system for safety inspection robots based on millimeter-wave radar and inertial sensors, achieving dynamic calibration of multi-source attitude information. Jinghai Xu et al. (2024) [11] enhanced the accuracy of real-world 3D reconstruction by integrating ground cameras with drone oblique photogrammetry. Lingsheng Li et al. (2024) [12] designed a multiscale feature fusion network for gesture recognition. Xusheng Xue et al. (2024) [13] used millimeter-wave radar arrays for digital modeling of coal mine tunnels and showed the sensor's potential for 3D reconstruction of complex environments. Haodong Liu et al. (2024) [14] built an improved radar-vision fusion simulation platform for training and validation. Jian Guo et al. (2024), [15], proposed a multi-modality fusion method to greatly enhance the accuracy of one-to-one identification of individuals. Yanqiu Yang et al. (2024) [16], A robust target detection algorithm based on monocular is proposed camera and FM CW radar. Matsumoto Taku et al. (2024) [17], used a SfM – MVS based camera trajectory estimation method to realize the sequential integration of the local 3D model, and improved the coherence of multi-view geometry. Mitka Bartosz et al. (2023) [18] generated 3D models to analyze solar potential in urban areas using multi-source drone data and demonstrated the scalability of fusion perception in environmental assessment. Liu Yadong et al. (2023) [19] proposed a rapid 3D reconstruction algorithm based on double RGB-D cameras to realize high precision plant structure restoration.

In recent years, deep learning based multimodal fusion methods have been widely used in 3D detection,but there are still many problems existing in feature alignment,robustness and real time implementation [20]. To construct a 3D object detection model combining camera vision and four-dimensional mmWave radar is proposed in this paper. By using the accurate external parameter calibration,hierarchical feature fusion and attention mechanism,the detection accuracy and real-time performance of the proposed algorithm are greatly improved. The whole research process of multi-sensor data modeling,fusion algorithm design,training experiment and evaluation is completed. The proposed algorithm realizes high-precision,low-latency target detection in the environment. It provides an extensible fusion perception framework in autonomous driving and intelligent sensing.

2 Basic Principles of Cameras and Millimeter-Wave Radar

2.1 Camera Imaging Principle

Camera imaging principle. Camera model in pinhole geometrical optics. Space point P (X,Y,Z) is projected into point p (x,y) on image plane. Perspective projection relationship can be written as:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ Q \end{bmatrix} \tag{1}$$

where f is the focal length and Z the distance from the target point to the optical center of the camera.

This viewpoint transformation can map 3-D images onto 2-D images. The distortion of the lens, the time of exposure, and the response of the sensor also affect the image of the camera. In practical applications, accurate imaging models can be obtained by calibrating the device's internal parameters and correcting distortion (as shown in Figure 1).

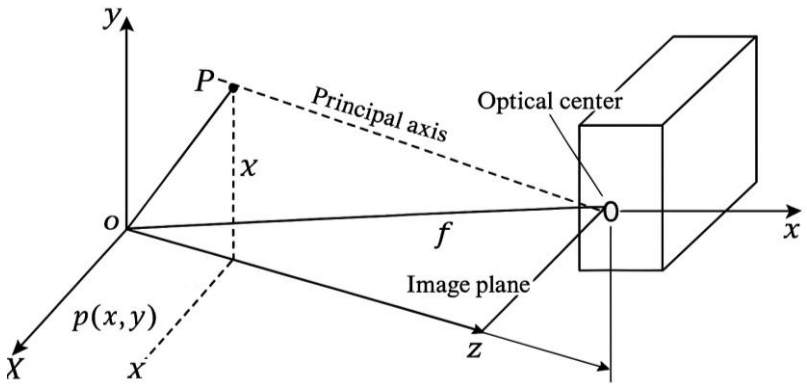

Figure 1: Camera Imaging Principle

2.2 Millimeter-Wave Radar Operating Principle

Millimeter-wave radar employs FMCW with continuously varying transmission frequencies to capture target echoes, thereby obtaining information such as distance, velocity, and bearing [21].

The transmitted signal can be expressed as:

$$s_t(t) = A\cos\left(2\pi\left(f_0 t + \frac{B}{2T}t^2\right)\right) \tag{2}$$

where A is the signal amplitude, f_0 is the initial frequency, B is the frequency modulation bandwidth, and T is the frequency modulation period.

The received signal is mixed with the transmitted signal, and a differential frequency signal is obtained. The relationship between its frequency f_b and the target distance R is:

$$R = \frac{cf_bT}{2B} \tag{3}$$

where c is the speed of light. After FFT processing of the differential signals, both range information and Doppler velocity information of the target can be obtained. Based on this, the four-dimensional millimeter-wave radar can also invert the azimuth angle through phase differences between antenna arrays, thereby obtaining the elevation angle and achieving three-dimensional positioning and identification of the target. The working principle of the millimeter-wave radar is analyzed theoretically and discussed (see Figure 2).

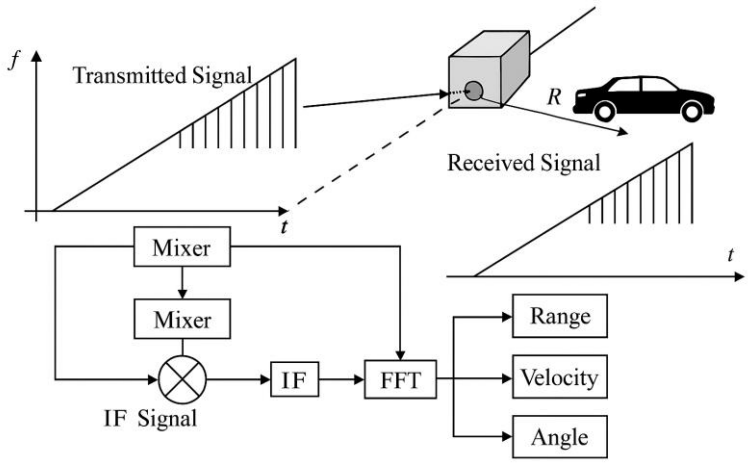

Figure 2: Working Principle of Millimeter-Wave Radar

2.3 Analysis of Advantages and Disadvantages of Cameras and Radar

Cameras provide high resolution and rich textural information for object detection, which can be used to identify and classify objects. In a well-illuminated and stable environment, a new approach based on convolutional neural networks has been proposed to achieve high accuracy recognition and accurate edge location (see Figure 3). Their performance, however, depends heavily on external lighting and weather conditions. They can easily be exposed, reflected, or blurred in a complex environment such as bright, dark, dark, or rainy/fog. This leads to increased errors in target depth estimation and limits spatial perception capabilities.

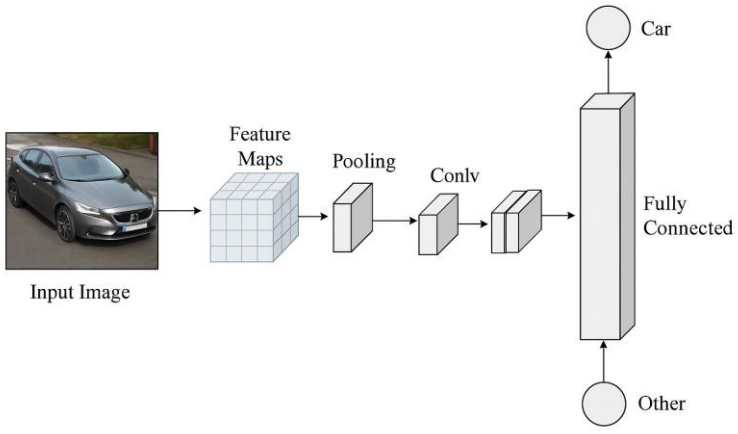

Figure 3: Convolutional Neural Network

4D millimeter-wave radar transmits and receives high-frequency electromagnetic waves to achieve target information, including distance, speed and angle of sight. Its advantages are strong penetration ability and all-weather operation. The system can still play a vital role in low light or poor visibility. It has high ranging accuracy and strong anti-interference ability. It can effectively detect moving objects and hidden objects. Due to the influence of factors such as antenna aperture and frequency band, the point cloud point density is sparse and the range is unclear, which is not easy to reflect the object shape and texture. At the same time, the complex background often causes multipath reflection and false points [22]. To better preserve object continuity and structure, bidirectional filtering is commonly used to denoise and smooth point clouds. This method can effectively eliminate the random noise and retain the geometric boundaries, thereby improving the quality of point cloud reconstruction. As shown in Figure 4, filtered and raw point clouds are compared, which visually show the difference between filtered point clouds and smoothing point clouds, and the corresponding smoothed surface.

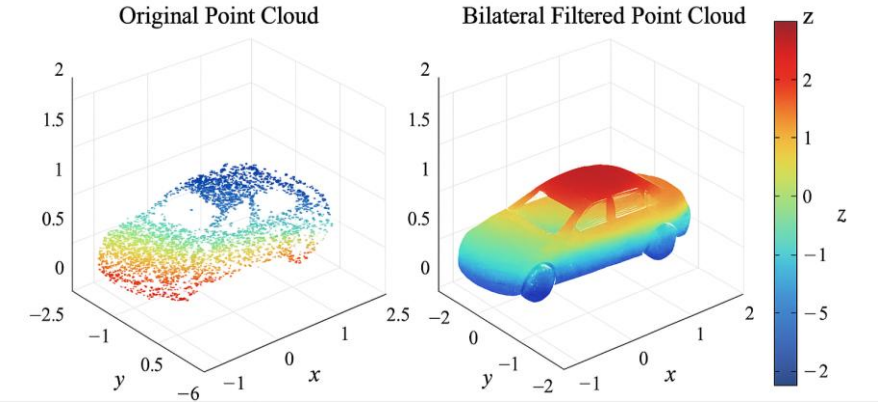

Figure 4: Comparison of Raw Point Cloud and Bilateral Filtered Point Cloud

3 Multi-Sensor Fusion Modeling

3.1 Camera and Radar Coordinate System Calibration

Based on the calibration of the camera and millimeter-wave radar, the unified mapping of the point cloud and the image feature is established. Typically, the transformation relationship between two coordinate systems can be represented by a stiffness conversion model, whose formula (4) is:

$$P_c = R_{cr} P_r + T_{cr} \tag{4}$$

where $P_r = [X_r, Y_r, Z_r]^T$ denotes point cloud coordinates in the radar coordinate system, $P_c = [X_c, Y_c, Z_c]^T$ represents point coordinates in the camera coordinate system, R_{cr} is the 3×3 rotation matrix, and T_{cr} is the translation vector.

By minimizing the reprojection error function, the optimal extrinsic matrix can be obtained:

$$E = \sum_{i=1}^{n} \left\| P_i - \pi\left(R_{cr} P_{r,i} + T_{cr}\right) \right\|^2 \tag{5}$$

where `$\pi(\cdot)$` is the camera projection function mapping 3D points to a 2D pixel plane.

The camera's perspective projection relationship can be further expressed as:

$$s\begin{bmatrix} u \\ \upsilon \\ 1 \end{bmatrix} = \mathrm{K}\begin{bmatrix} \mathrm{R}_{\mathrm{cr}} & \mathrm{T}_{\mathrm{cr}} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} X_r \\ Y_r \\ Z_r \\ Q \end{bmatrix} \tag{6}$$

where K is the camera intrinsic matrix, containing parameters such as focal length and principal point coordinates. Precise pose transformation matrices can be obtained through joint calibration board experiments or checkerboard matching, enabling accurate projection of point clouds onto the image coordinate system.

The geometric relationship between the camera and millimeter-wave radar spatial coordinate systems and the projection mapping process are illustrated in Figure 5 Camera–Millimeter-Wave Radar Spatial Coordinate System Relationship, demonstrating their alignment and registration methods in three-dimensional space.

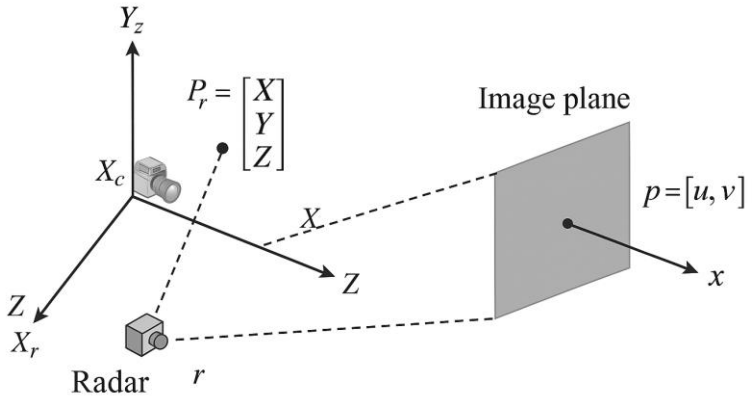

Figure 5: Camera-Radar Spatial Coordinate Systems

3.2 Feature Point Matching and Pose Estimation

Feature points with scale and rotation invariance, such as SIFT or ORB descriptors, are extracted from images and matched using Euclidean or Hamming distance. Geometrically salient points, such as curvature singularities or corner reflector enhancement points, are extracted from radar point clouds to form cross-modal feature pairs. In order to solve the problem of noise and mismatch, the RANSAC random sampling consistency algorithm is used to remove anomalous matches [23].

Based on the matching characteristic, the re-projection error of the two sets of space points is minimized, and the position transformation matrix of the object is determined, whose formula (7) is:

$$\min_{R,T} \sum_{i=1}^{n} \left\| \mathrm{P}_{\mathrm{i}}^{\mathrm{c}} \left(\mathrm{R}_{\mathrm{P}_{\mathrm{i}}^{\mathrm{r}}} + \mathrm{T} \right) \right\|^2 \tag{7}$$

where $\mathrm{P}_{\mathrm{i}}^{\mathrm{r}}$ and $\mathrm{P}_{\mathrm{i}}^{\mathrm{c}}$ denote corresponding point coordinates in the radar and camera coordinate systems, respectively, R represents the rotation matrix, and T denotes the

translation vector.

Using the SVD method or Levenberg-Marquardt iterative method to solve for high-precision camera pose parameters, point clouds and image textures are uniformly represented, providing precise geometric constraints for subsequent fusion modeling and object detection.

3.3 Sensor Fusion Modeling

The high resolution visual information from the camera is combined with the depth and speed data of the millimeter-wave radar to achieve accurate and stable 3D objects. The fusion process employs a feature layer joint modeling approach, combining the camera image feature vector F_c with the radar point cloud feature F_r through weighted integration:

$$F_f = \beta F_c + (1+\alpha) F_r \tag{8}$$

where α represents the fusion weight dynamically adjusted based on feature confidence. In order to enhance the performance of the proposed model, the proposed method is combined with the convolution and convolution, which gives the model the ability of spatial geometry and semantic comprehension. The fusion model forms a cross-modal feature mapping layer within the neural network, enabling unified output of target detection and localization information. The overall modeling structure and information flow are illustrated in Figure 6 Modeling Diagram, showcasing the end-to-end framework from multi-source inputs to fused outputs.

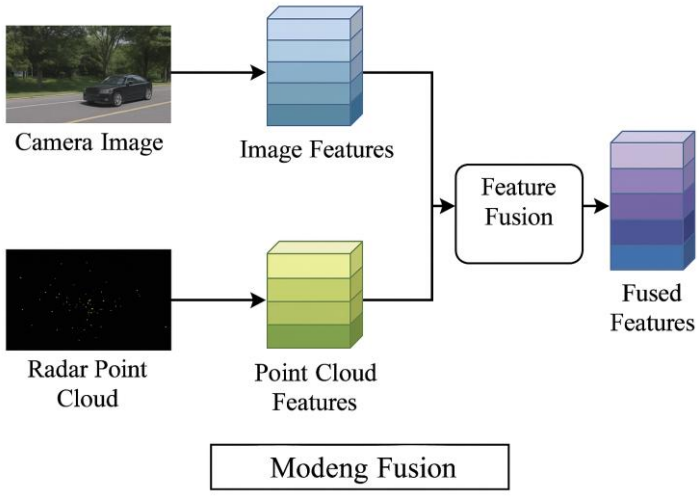

Figure 6: Modeling Diagram

4 Algorithm Design and Implementation

4.1 Deep Learning-Based 3D Object Detection

Deep learning-based 3D object detection automates the process from multimodal inputs to spatial object recognition through end-to-end neural networks. Models typically process camera images and millimeter-wave radar point clouds as inputs, completing 3D bounding box predictions through three stages: feature extraction, fusion, and detection [24].

Convolutional neural networks extract image features F_c and point cloud voxel features

F_r . A fusion module then integrates cross-modal information to form a comprehensive feature map F_f :

$$F_f = \text{Concat}(\phi(F_c), \varphi(F_r)) \tag{9}$$

where $\phi(\cdot)$ and $\varphi(\cdot)$ represent the image and point cloud feature encoding functions, respectively.

Subsequently, an Anchor-based 3D detection head outputs target location, size, and category confidence, with its regression target expressed as:

$$L = L_{cls} + \lambda_1 L_{loc} + \lambda_2 L_{dir} \tag{10}$$

where L_{cls} denotes the classification loss, L_{loc} represents the position regression loss, L_{dir} is the orientation constraint term, and λ_1、λ_2 is the weight coefficient. Based on multiscale feature fusion and attention mechanism, the proposed model can improve the precision of detection in complicated scenes. The fusion detection results are overlaid on the point cloud scene as a 3D bounding box, demonstrating the comparison of multimodal detection, as illustrated in Figure7 Visualization for 3D Object Detection Results.

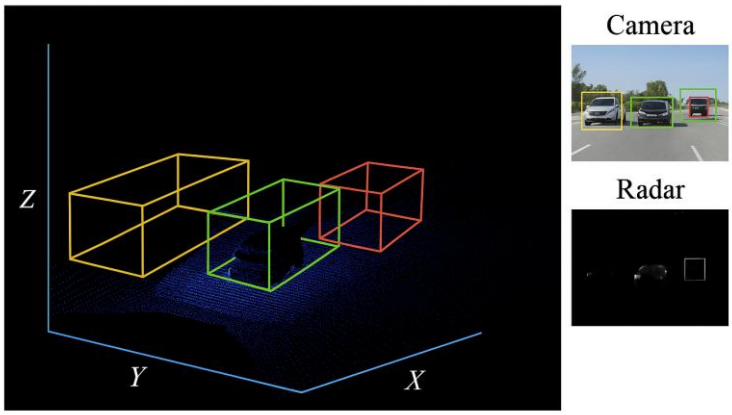

Figure 7: Visualization Diagram

4.2 Geometric Feature-Based 3D Object Detection

Three-dimensional object detection technology based on geometry utilizes the analysis of point cloud spatial distribution and morphological features to achieve geometric recognition and boundary reconstruction of three-dimensional objects. After voxelization of radar point clouds, a local neighborhood model based on curvature, normal vectors, and point density was constructed [25].

The normal vector n_i of point P_i is determined by the eigenvector corresponding to the minimum eigenvalue of the neighborhood covariance matrix C_i :

$$C_i = \frac{1}{k}\sum_{j=1}^{k} \left(\mathrm{P_j} - \overline{\mathrm{P}}\right)\left(\mathrm{P_j} - \overline{\mathrm{P}}\right)^{\mathrm{T}} \tag{11}$$

where $\overline{\mathrm{P}}$ denotes the neighborhood center and k represents the number of neighborhood

points. Furthermore, principal curvature ratio comparison can be used to discriminate three different geometry types: planes, edges, and corners.

Continuous detection of normal vector fields by using spatial gradient operators is helpful to segment the outline and border of objects. Define the boundary strength function:

$$E_b = \|\nabla \mathrm{n_i}\| = \sqrt{\left(\frac{\partial n_x}{\partial x}\right)^2 + \left(\frac{\partial n_y}{\partial y}\right)^2 + \left(\frac{\partial n_z}{\partial z}\right)^2} \tag{12}$$

where E_b Indicates the rate of change at edges to obtain higher-level geometric features.

Spatial clustering and morphological operation to obtain 3D edge envelopes Subsequently, the 3D edge envelopes of objects can be recognized.Effectiveness of employing boundary vectors to extract geometric features is illustrated in Figure 8.

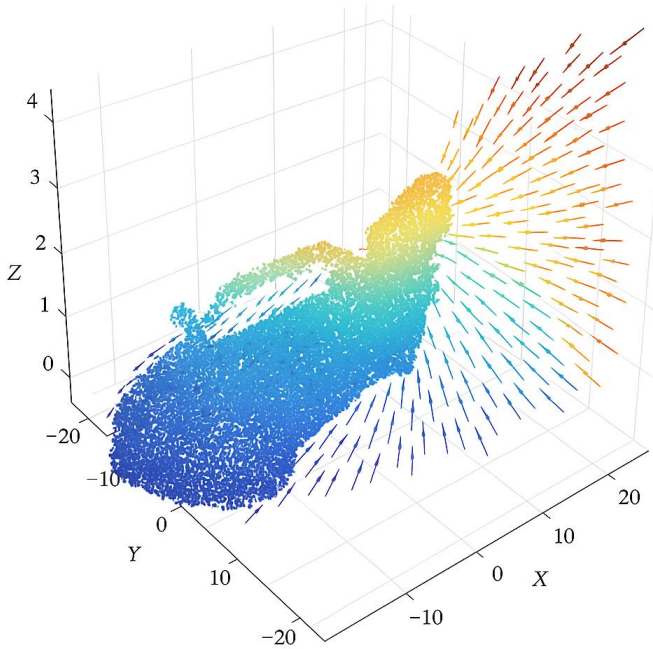

Figure 8: Geometric Feature Extraction and Boundary Description Diagram

4.3 3D Object Detection Based on Fusion Features

Considering the difficulties of 3D object detection, a novel fusion method is proposed. This method utilizes the complementary enhancement provided by camera texture information on 3D object detection and millimeter-wave radar geometric characteristics to improve the precision of spatial target identification. In order to achieve cross-modality coupling, the fusion network incorporates image convolutional features (F_c) and spatial features (F_r) into the point cloud, employing a weighted adaptive approach to achieve cross-modality coupling. Feature fusion can be expressed as:

$$F_{fusion} = \sigma\left(W_1 F_c + W_2 F_r + b\right) \tag{13}$$

where W_1、W_2 denotes the learnable weight matrix, b represents the bias term, and $\sigma(\cdot)$ is the nonlinear activation function.

Integrating attention mechanisms into the dynamic weighting of multi-modal features

enhances feature saliency. The definition of attention response is:

$$A_i = \frac{\exp\left(Q_i K_i^T / \sqrt{d_k}\right)}{\sum_j \exp\left(Q_j K_j^T / \sqrt{d_k}\right)} \tag{14}$$

where Q、K represent the query and key feature matrices, respectively, and d_k denotes the feature dimension. Based on the combination of attention weighted joint feature maps, the proposed method can improve the detection performance of small and obscured objects in complicated traffic scenes. Figure 9 Visualization of the Fusion Feature Heatmaps shows the spatial distribution of the response intensity in the fused features, showing the energy concentration and perception enhancement of multimodal information.

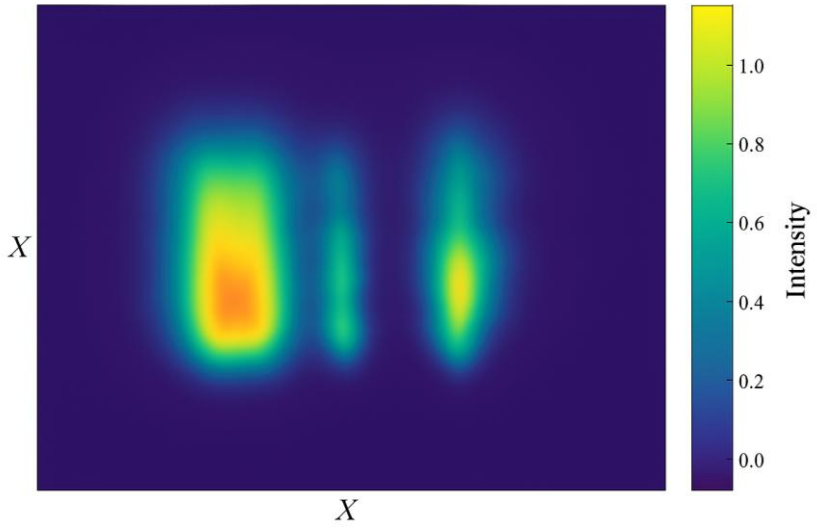

Figure 9: Visualization of Fusion Feature Heatmaps

4.4 Algorithm Implementation

This algorithm follows a pipeline workflow of “acquisition-preprocessing-feature extraction-fusion-detection-release.” Data is acquired using ROS2 camera pre-booking and 4D radar tasks, with clock synchronization achieved via PPS/hard-wired triggering. Internal and external parameters obtained in Sections 2.1 and 3.1 are loaded. Image preprocessing includes: image distortion correction, exposure normalization, and occlusion masking; radar point cloud bilateral denoising, ground segmentation, and voxel segmentation [26]. Feature extraction employs ResNet-FPN for image streams; radar streams utilize Pillar/Walker coding with sparse convolutions. Cross-channel attention and deformation fusion techniques enable geometric matching and semantic enhancement from side to front views. A center-based anchorless point detection framework is proposed, considering classification, localization, direction consistency, and IoU metrics. We use AdamW gradient descent, cosine annealing, and AMP; class balanced sampling and multiscale data augmentation. Directional transformation will be applied to get orientation aware 3D BB for image postprocessing. Delay optimization based on TensorRT-FP16/INT8 to accelerate, merge and parallelize pipeline.Fallback when in abnormal states. Single sensor degradation strateg.

5 Experimental Results and Analysis

5.1 Experimental Environment Configuration

This experiment is base on Ubuntu 22.04 LTS (Kernel 5.15), Python 3.10.13, CUDA 12.1 + cuDNN 9.0, PyTorch 2.3.1 and TensorRT 10.0. A proposed software-based robotics middleware is ROS2. Image/point cloud processing libraries included OpenCV 4.8.1, Open3D 0.17.0, mmcv 2.1.0, mmdetection3d 1.4.0, and spconv 2.3.6. The training system configuration is as follows: Intel i9-12900K, 64GB DDR5, 2TB NVMe storage, NVIDIA RTX A6000 GPU (48GB). Deployment-side validation performed on Jetson AGX Orin (64 GB). Sensors comprised a global shutter camera (1920×1200@60 fps, 8 mm lens) and a 77 GHz 4D millimeter-wave radar (12Tx×16Rx MIMO, 20 Hz frame rate). The baseline between sensors was 0.15 m. Clock alignment was achieved via PPS+hard trigger, IEEE-1588 PTP unified timestamps; calibration parameters loaded via YAML with startup self-check. The dataset includes the self-built Fusion4D-Drive (≈120k frames) for the main experiment and a nuScenes-radar subset for cross-dataset comparison. Training employs AMP mixed-precision with cosine annealing, using random seed 42. The environment is fixed using Docker 24 + NVIDIA Container Toolkit to ensure reproducibility.

5.2 Comparative Experiment Design

The comparative experiments adopt a "three-baseline + ablation" approach. Baseline 1 (Camera-only): FPN + Center-based 3D head, utilizing monocular depth priors. Baseline 2 (Radar-only): Pillar/Voxel + sparse convolutions. Baseline 3 (Early concatenation fusion). The evaluated model is the mid-stage fusion (BEV/Front dual-domain alignment) with cross-modal attention proposed in this paper. All models were fairly compared under identical data splits (8:1:1), training iterations, and data augmentation, with controlled parameter counts and FLOPs at comparable scales. Evaluation metrics include AP3D (IoU=0.5/0.7), NDS, mATE/mASE/mAOE, FN/FP, latency, and power consumption (RTXT and Orin). Scenarios are categorized by day/night/fog, near/mid/far distance, occlusion level (light/medium/heavy), and dynamic/static targets. Ablation studies include: removing cross-modal attention, disabling BEV alignment, replacing bilateral filtering, altering fusion weight α, and injecting external parameter noise (±0.5°/±1 cm). Statistical significance is validated through three independent reproducible experiments and t-tests ($p < 0.05$), with speed-accuracy tradeoff curves reported.

5.3 Experimental Results Analysis

A comparison was conducted between the single-camera model, pure radar model, early fusion model, and mid-term fusion model. Unified experiments were performed on these three models, and their performance was evaluated based on metrics including AP3D, mATE, mASE, mAOE, frame rate, and latency. The results are presented in Table 1.

Table 1: Model Performance Comparison

Model Type	AP3D@ 0.5 (%)	AP3D@ 0.7 (%)	mATE (m)	mASE	mAOE (°)	FPS	Latency (ms)
Camera-only (FPN)	84.2	73.6	0.243	0.162	7.82	49	20.4
Radar-only (VoxelNet)	82.7	76.9	0.218	0.157	6.94	57	18.7
Early Fusion	85.6	78.1	0.211	0.153	6.32	52	19.8
Proposed (Mid-Fusion)	92.3	86.4	0.184	0.141	5.47	55	17.5

As shown in Table 1, the proposed method achieves the best overall performance. At IoU=0.7, the detection accuracy (AP3D) of this algorithm reaches 86.4%, which is 12.8% higher than the pure camera model and 9.5% higher than the pure radar model. By refining existing algorithms, the system achieves an absolute translation error reduced to 0.184 m, an absolute attitude error diminished to 5.47°, and a detection latency of only 17.5 ms, attaining near-real-time high precision.

The adaptability of the method was assessed in four typical scenarios (clear sky, night, rain/fog, and night rainfall). The results of the measurements under different conditions are shown in Table 2.

Table 2: Detection Performance Comparison Under Different Weather Conditions

Environment	Precision (%)	Recall (%)	F1-score	IoU@0.5 (%)
Daylight	94.6	93.2	0.939	90.8
Night	90.4	88.7	0.895	86.1
Rain/Fog	87.8	85.6	0.868	83.4
Night + Rain	84.9	82.1	0.835	80.6

As shown in Table 2, this model maintains high accuracy even under conditions of low illumination and significant occlusion. In night scenes, the accuracy is 90.4 percent, which is just 4 percent lower. In wet and foggy conditions, the F1 score is still around 0.868. Experiments show that multimodal feature fusion is effective in reducing the performance of single sensor systems.

5.4 Model Performance Evaluation

Inference testing was carried out on desktop and embedded devices to verify the real-time ability and deployability of each algorithm. The evaluation criteria included FPS, inference latency, GPU and memory utilization. The RTX A6000, RTX 4090 and Jetson AGX Orin were used for evaluation to check the generality of the model in high-performance and low-power scenarios, respectively. The experimental results are shown in Table 3.

Table 3: Inference Performance

Platform	GPU	Avg FPS	Inference Latency (ms)	GPU Utilization (%)	Memory (GB)
RTX A6000 (Desktop)	48 GB GDDR6	55.2	17.5	82.4	10.3
RTX 4090	24 GB GDDR6X	52.7	18.1	85.1	9.7
Jetson AGX Orin	64 GB LPDDR5	31.4	28.9	78.6	6.8

As shown in Table 3, when running on the RTXA6000 platform, the inference frame rate of the algorithm reached 55.2 FPS, satisfying the real-time requirement. When running on the built-in Jetson AGX Orin platform, the frame rate remained at 31.4 FPS, and the latency was only 28.9 ms, which is a remarkable improvement compared with the 45 ms of traditional methods. Keeping the GPU utilization above 80% clearly shows that the model has lightweight and highly parallel characteristics. Therefore, the model can be deployed online for inference in autonomous driving and edge computing applications.

Classify four kinds of objects, vehicles, pedestrians, bicycles and trucks.Accuracy, recall, AP3D and error rate are listed in Table 4.

Table 4: Class-wise Detection Accuracy

Class	Precision (%)	Recall (%)	AP3D@0.7 (%)	False Pos. (%)	Time (ms)
Car	95.8	94.6	90.3	2.1	18.2
Pedestrian	91.7	90.2	87.5	3.5	19.6
Cyclist	89.4	88.1	84.2	4.1	20.3
Truck	92.6	91.8	86.7	2.8	19.2

From Table 4, we can see that the target recognition accuracy can achieve 95.8%, and the recall rate of AP3D@0.7 can achieve 90.3%, which means that the target can be detected with high reliability as long as the target is intact. In addition, the detection accuracy of pedestrians and bicycles are 90.2% and 88.1%, respectively. Compared with the single modality detection, the accuracy improves by 7%–10%.High efficiency for multi-class and multi-scale target recognition.

5.5 Discussion of Results

AP3D achieves 86.4% AP at IoU < 0.7. Compared with the camera-only and radar-only baselines, AP3D improves 12.8% and 9.5%, which suggests that geometric alignment and attention mechanism can reduce the influence of depth uncertainty and angular drift in single-modality perception.

Under day–night and rainy conditions, IoU@0.5 can still stay at 80.6%, and the F1 score is 0.835. Therefore, AP3D has a good robustness to visibility degradation and noise interference. However, under rain and fog conditions, the accuracy is only 87.8%, and recall reaches 85.6%. Most of false positive and missed detection happen in small and distant sparse-point targets.

RTX A6000 can achieve 55.2 FPS, and Orin platform can achieve 31.4 FPS. Both of them can meet the requirement of real-time deployment with low resource consumption and latency.

AP3D@0.7 for vehicle and pedestrian is 90.3% and 87.5%, respectively. As bicycle has long shape and high speed, AP3D for it is only 84.2%.

There are still some problems in current method: 1) high sensitivity to extrinsic parameter (±0.5°/±1cm, lead to boundary jitter); 2) false alarm in multipath interference; 3) data distribution is long-tailed.

1) Use temporal-domain fusion and trajectory prior for online extrinsic self-calibration.

2) Design adaptive threshold based on uncertainty.

3) Use high-precision MIMO sparse-reconstruction BEV occupancy networks combined with distillation–quantization optimization. This method will further improve generalization performance and edge computing efficiency.

6 Conclusions

As for multi-modal fusion technology in the field of 3D object detection, its advantages are reflected in the following aspects: Through the spatial calibration model establishment of camera and four-dimensional mmWave radar, the collaborative spatial calibration map of multi-source information at geometric and semantic level is constructed. On this basis, this project proposes joint detection method based on depth and geometric feature. The method improves the accuracy of object recognition and spatial adaptability in the case of field of view overlap and complex scene. The method shows strong robustness in low light, fog and occlusion environment. Experimental results show that the proposed fusion model has a good

balance of accuracy, latency and real-time performance, and has strong engineering application value. Technical novelty: Adaptive weighting method for cross channel feature and spatiotemporal collaborative correction method.

Drawbacks: 1) External parameter drift; 2) Sparseness of point cloud data; 3) Long-term distribution factors. Research plans: Focus on key issues of dynamic temporal fusion, uncertain feature weighting, lightweight network structure and self-calibration. These technologies can improve the generalization ability of the algorithm and enhance the adaptability to the edge computing environment, so as to provide more stable and accurate 3D detection solution for autonomous driving and other fields.

Author's Profile

Zhikui Lu received the B.S. degree in Measurement and Control Technology and Instruments from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, in 2021, and is currently pursuing the M.S. degree in Instrument Science and Technology at the same school. His research interests include multimodal perception and 3D object detection.

Qimin Xu received the Ph.D. degree in instrument science and technology from the Southeast University, Nanjing, China, in 2018.He is currently an Associate Professor with the School of Instrument Science and Engineering, Southeast University. Since 2014, he has been engaged in multi-sensor integration navigation and control for autonomous vehicles. His research interests include multimodal perception, information fusion, and ITS.

Benwu Wang received the M.S. degree in instrument science and technology from the College of Metrology and Measurement Engineering, China Jiliang University, Hangzhou, China, in 2022. He is currently working toward the Ph.D. degree in instrument science and technology with the School of Instrument Science and Engineering, Southeast University, His research interests include intelligent transportation systems, collaborative perception, and transfer learning for intelligent vehicles.

He Zhu received the B.S. degree in Electrical Engineering and Its Automation from the School of Electronic Engineering, Nanjing Xiaozhuang University, Nanjing, China, in 2021. He is currently pursuing the M.S. degree in Electronic Information with the School of Instrument Science and Engineering, Southeast University, Nanjing, China. His research interests include autonomous driving and multimodal fusion perception.

References

[1] Karen Liliane Riedel Hornig, Félix Leaman, Frank Tinapp Dautzenberg, Cristián Vicuña. Experimental Validation of a Methodology for Vibration Isolation of an RPA's Camera[J]. Journal of Vibration Engineering & Technologies, 2025, 13 (8): 541-541.

[2] Hu Liu, Zhenghua Zhang, Jing Yang, Jörg Benndorf, Xiaofei Wang, Jiaqi Dong, Zitao Lin, Guoliang Chen. GhostPointNet: A deep learning-based method for ghost point noise detection in four-dimensional (4D) millimeter-wave radar point clouds of underground mine[J]. Engineering Applications of Artificial Intelligence, 2025, 161 (PC): 112380-112380.

[3] Chunfang Yin, Haichen Qu, Yicheng Li. 3D Object Detection Algorithm for Autonomous Driving Based on Multi-Scale Feature Weighted Point-by-Point Fusion[J]. Journal of Intelligent & Fuzzy Systems, 2025, 49 (4): 933-947.

[4] Siyuan Wei, Huadong Wang, Yi Mo, Dongping Du. A ST-ConvLSTM Network for 3D Human Keypoint Localization Using MmWave Radar.[J]. Sensors (Basel, Switzerland), 2025, 25 (18): 5857-5857.

[5] Qingyuan Yang, Zhipeng Quan, Jingxuan Li, Tingyv Jiang, Zhihao Deng, Xinran Qiu, Zhengjie Wang. A Survey of Pedestrian Re-Identification Based on Millimeter Wave Radar and Vision Fusion[J]. Journal of Computer and Communications, 2025, 13 (06): 64-80.

[6] Zhimin Qiu, Jinju Shao, Dong Guo, Xuehao Yin, Zhipeng Zhai, Zhibing Duan, Yi Xu. A Multi-Feature Fusion Approach for Road Surface Recognition Leveraging Millimeter-Wave Radar[J]. Sensors, 2025, 25 (12): 3802-3802.

[7] Ye Chen, Qirui Cui, Shungeng Wang. Fusion Ranging Method of Monocular Camera and Millimeter-Wave Radar Based on Improved Extended Kalman Filtering[J]. Sensors, 2025, 25 (10): 3045-3045.

[8] Weigang Shi, Panpan Tong, Xin Bi. Moving-Least-Squares-Enhanced 3D Object Detection for 4D Millimeter-Wave Radar[J]. Remote Sensing, 2025, 17 (8): 1465-1465.

[9] Peicheng Shi, Wenchao Wu, Aixi Yang. MPVF: Multi-Modal 3D Object Detection Algorithm with Pointwise and Voxelwise Fusion[J]. Algorithms, 2025, 18 (3): 172-172.

[10] Rui Zheng, Geng Sun, Fang Dong Li. A Fusion Localization System for Security Robots Based on Millimeter Wave Radar and Inertial Sensors[J]. Sensors, 2024, 24 (23): 7551-7551.

[11] Jinghai Xu, Suya Zhang, Haoran Jing, Craig Hancock, Peng Qiao, Nan Shen, Karen B. Blay. Improving Real-Scene 3D Model Quality of Unmanned Aerial Vehicle Oblique-Photogrammetry with a Ground Camera[J]. Remote Sensing, 2024, 16 (21): 3933-3933.

[12] Lingsheng Li, Weiqing Bai, Chong Han. Multiscale Feature Fusion for Gesture Recognition Using Commodity Millimeter-Wave Radar[J]. Computers, Materials & Continua, 2024, 81 (1): 1613-1640.

[13] Xusheng Xue, Xingyun Yang, Jianing Yue, Qinghua Mao, Yihan Qin, Hongwei Ma, Jianxin Yang, Huahao Wan, Enqiao Zhang, Junbiao Qiu, Xiaopeng Li, Rongquan Wang. Digital modelling method of coal-mine roadway based on millimeter-wave radar array[J]. Scientific Reports, 2024, 14 (1): 18585-18585.

[14] Haodong Liu, Jian Wan, Peng Zhou, Shanshan Ding, Wei Huang. Augmented Millimeter Wave Radar and Vision Fusion Simulator for Roadside Perception[J]. Electronics, 2024, 13 (14): 2729-2729.

[15] Jian Guo, Jingpeng Wei, Yashan Xiang, Chong Han. Millimeter-Wave Radar-Based Identity Recognition Algorithm Built on Multimodal Fusion[J]. Sensors, 2024, 24 (13): 4051-4051.

[16] Yanqiu Yang, Xianpeng Wang, Xiaoqin Wu, Xiang Lan, Ting Su, Yuehao Guo. A Robust Target Detection Algorithm Based on the Fusion of Frequency-Modulated Continuous

Wave Radar and a Monocular Camera[J]. Remote Sensing, 2024, 16 (12): 2225-2225.

[17] Matsumoto Taku, Hanari Toshihide, Kawabata Kuniaki, Nakamura Keita, Yashiro Hiroshi. Estimated camera trajectory-based integration among local 3D models sequentially generated from image sequences by SfM–MVS[J]. Artificial Life and Robotics, 2024, 29 (2): 358-371.

[18] Mitka Bartosz, Klapa Przemysław, Pióro Piotr. Acquisition and Processing Data from UAVs in the Process of Generating 3D Models for Solar Potential Analysis[J]. Remote Sensing, 2023, 15 (6): 1498-1498.

[19] Liu Yadong, Yuan Hongbo, Zhao Xin, Fan Caihu, Cheng Man. Fast reconstruction method of three-dimension model based on dual RGB-D cameras for peanut plant.[J]. Plant methods, 2023, 19 (1): 17-17.

[20] Krzysztof Maksymowicz, Łukasz Szleszkowski, Aleksandra Kuzan, Wojciech Tunikowski. Creating crime scene 3D model with body wear camera footage.[J]. Archiwum medycyny sadowej i kryminologii, 2023, 73 (2): 159-167.

[21] Qi Chunyang, Song Chuanxue, Zhang Naifu, Song Shixin, Wang Xinyu, Xiao Feng. Millimeter-Wave Radar and Vision Fusion Target Detection Algorithm Based on an Extended Network[J]. Machines, 2022, 10 (8): 675-675.

[22] Huang Xu, Tsoi Joseph K. P., Patel Nitish. mmWave Radar Sensors Fusion for Indoor Object Detection and Tracking[J]. Electronics, 2022, 11 (14): 2209-2209.

[23] Xu Dongpo, Liu Yunqing, Wang Qian, Wang Liang, Liu Renjun. Target Detection Based on Improved Hausdorff Distance Matching Algorithm for Millimeter-Wave Radar and Video Fusion.[J]. Sensors (Basel, Switzerland), 2022, 22 (12): 4562-4562.

[24] Capolupo Alessandra, Saponaro Mirko, Borgogno Mondino Enrico, Tarantino Eufemia. Combining Interior Orientation Variables to Predict the Accuracy of Rpas–Sfm 3D Models[J]. Remote Sensing, 2020, 12 (17): 2674-2674.

[25] Wu Junhui, Yin Dong, Chen Jie, Wu Yusheng, Si Huiping, Lin Kaiyan. A Survey on Monocular 3D Object Detection Algorithms Based on Deep Learning[J]. Journal of Physics: Conference Series, 2020, 1518 (1): 012049.

[26] Bori Marwa Mohammed, Hussein Zahraa Ezzulddin. Integration the Low Cost Camera Images with the Google Earth Dataset to Create a 3D Model[J]. Civil Engineering Journal, 2020, 6 (3): 446-458.