Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.24680

Research on 3D Object Detection Technology Based on Multimodal Fusion

Shijie Lyu

Georgia Institute of Technology, Atlanta, USA
slyu41@gatech.edu

Abstract. To address the challenge of missed detections of long-distance targets in autonomous driving, this study proposes an enhanced 3D object detection model based on the CenterFusion framework, integrating camera and millimeter-wave radar data. An early fusion strategy is employed to project radar data onto the image plane, combining it with image data to form a multi-channel input, thereby enhancing the model's robustness against interference. Additionally, an attention mechanism is incorporated post-feature fusion to prioritize the extraction of critical information from the fused feature map, significantly improving detection accuracy. The loss function is optimized to mitigate the imbalance between positive and negative samples. Comparative and ablation experiments conducted on the nuScenes dataset demonstrate that the proposed model achieves a 1.5% improvement in average detection accuracy and a 2.1% increase in nuScenes Detection Score (NDS) compared to the baseline CenterFusion model, effectively enhancing long-distance target detection capabilities.

Keywords: Autonomous Driving, Sensor Fusion, 3D Object Detection, Early Fusion, Attention Mechanism

1. Introduction

Three-dimensional (3D) object detection is a cornerstone of autonomous driving systems, enabling vehicles to perceive and interpret their surroundings accurately. Unlike traditional two-dimensional (2D) detection, 3D object detection provides spatial information, including depth, orientation, and scale, which are critical for safe navigation in complex environments. However, single-sensor modalities, such as cameras or radar, often face limitations. Camera-based systems excel in capturing rich semantic and textural information but struggle with depth estimation and perform poorly under adverse lighting or weather conditions [1]. Conversely, radar systems provide reliable depth and velocity data but suffer from sparse point clouds and limited semantic detail [2]. These limitations highlight the need for multimodal fusion to leverage complementary sensor data for robust 3D object detection.

Recent advancements in 3D object detection have focused on both single-sensor and multi-sensor approaches. Single-sensor methods, such as those utilizing radar point clouds, often employ deep learning architectures like PointNet or Transformer-based models to process sparse data [3]. However, these methods are sensitive to occlusions and signal interference, leading to incomplete detections. Multi-sensor fusion approaches, which combine data from cameras, radar, and LiDAR, have shown promise in improving detection robustness [4]. For instance, fusion models like the Multi-View Projection (MVP) model [5] and Camera-Radar Network (CRN) [6] integrate 2D image features with 3D point clouds to enhance detection performance. Despite these advances, challenges remain, including quantization errors during point cloud voxelization, loss of spatial information in data transformations, and inaccuracies in depth estimation for distant objects [7].

To address these issues, particularly the missed detection of long-distance targets, this study proposes an improved 3D object detection model based on the CenterFusion framework [8]. The proposed model introduces an early fusion strategy to align radar and camera data, an attention mechanism to enhance feature extraction, and an optimized loss function to balance sample distributions. Experiments on the nuScenes dataset validate the model's superior performance in detecting distant objects compared to existing methods.

2. Preliminary knowledge

2.1. Multimodal fusion in 3D object detection

Multimodal fusion integrates data from multiple sensors to enhance the robustness and accuracy of 3D object detection. Common sensors include cameras, which provide high-resolution RGB images, and millimeter-wave radar, which generates sparse point clouds with depth and velocity information. Fusion strategies are typically categorized into early, late, and deep fusion. Early fusion combines raw or pre-processed sensor data at the input stage, enabling the model to learn joint representations. Late fusion processes each modality independently before combining high-level features. Deep fusion integrates features at multiple network layers, balancing computational complexity and feature interaction [1].

The CenterFusion model, a representative multimodal fusion framework, combines camera images and radar point clouds for 3D object detection [8] as show in Figure 1. It consists of three main components: an image-based detection branch, a radar-based voxel network branch, and a secondary regression feature fusion network. The image branch employs an improved Deep Layer Aggregation (DLA) network within the CenterNet architecture to extract 2D features. The radar branch processes point clouds into voxel grids, capturing local features. The fusion network aligns and integrates these features to produce 3D bounding boxes. However, CenterFusion struggles with detecting distant objects due to sparse radar data and limited feature interaction at long ranges.

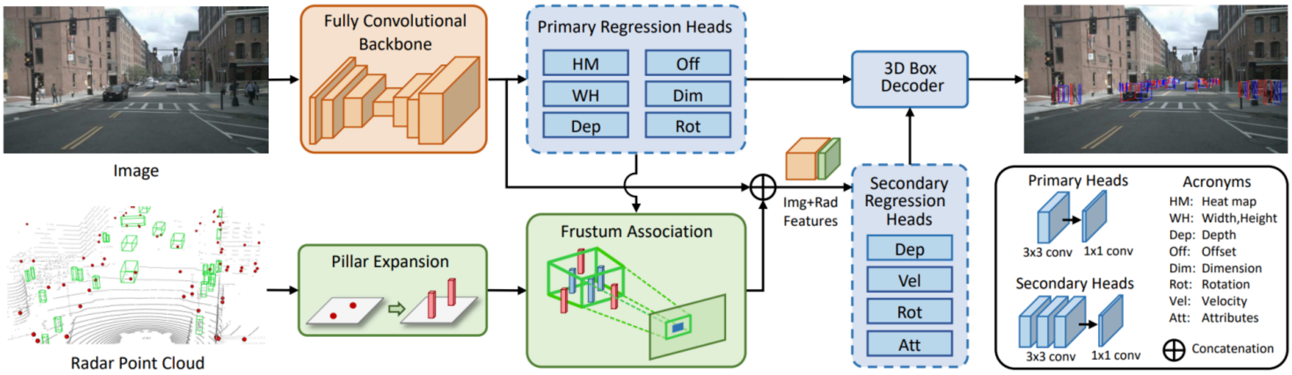


Figure 1: CenterFusion network architecture. Preliminary 3D boxes are first obtained using the image features extracted by the backbone. The frustum association module uses the preliminary boxes to associate radar detections to objects and generate radar feature maps. The image and radar features maps are then concatenated and used to refine the preliminary detections by recalculating depth and rotation as well as estimating objects' velocity and attributes

2.2. Attention mechanisms

Attention mechanisms enhance neural networks by focusing on relevant features while suppressing noise. In 3D object detection, attention mechanisms, such as those based on Transformer architectures, capture global contextual information and improve feature representation [4]. For instance, the Pyramid Squeeze Attention (PSA) block, proposed by Zhang et al. [9], efficiently aggregates multi-scale features, making it suitable for enhancing fusion networks. Integrating attention mechanisms into multimodal fusion models can prioritize critical spatial and semantic information, addressing challenges like sparse data and occlusion. The improved network architecture of the CenterFusion model as show in Figure 2.

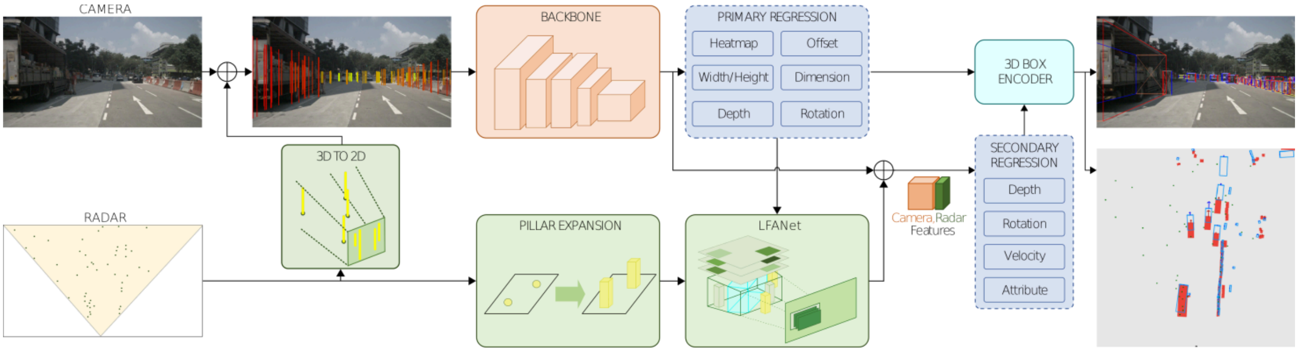


Figure 2: Improved network architecture of the CenterFusion Model

2.3. Evaluation metrics

The nuScenes dataset, a widely used benchmark for autonomous driving, evaluates 3D object detection using the nuScenes Detection Score (NDS) [10]. NDS combines the mean Average Precision (mAP) with five True Positive (TP) metrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE). The NDS is computed as:

$$NDS = \frac{1}{10}\left[5 \cdot mAP + \sum_{mTP \in \{mATE, mASE, mAOE, mAVE, mAAE\}} (1 - \min(1, mTP))\right]$$

This metric provides a comprehensive assessment of detection accuracy, localization, and attribute estimation, making it ideal for evaluating the proposed model’s performance.

3. Experiments

3.1. Experimental setup

To evaluate the performance of the proposed improved 3D object detection model based on CenterFusion, experiments were conducted using the nuScenes dataset, a comprehensive multimodal benchmark for autonomous driving [11]. The nuScenes dataset includes data from cameras, millimeter-wave radar, and LiDAR, capturing diverse urban driving scenarios. The dataset was split into training, validation, and test sets as per the official protocol, with the validation set used for performance evaluation.

The proposed model integrates an early fusion strategy, an attention mechanism, and an optimized loss function. The early fusion module maps radar point clouds onto the image plane, creating a multi-channel input alongside RGB images. The attention mechanism, inspired by the Pyramid Squeeze Attention (PSA) block [9], enhances feature extraction post-fusion. The loss function was adjusted to address the imbalance between positive and negative samples, improving detection stability. The baseline CenterFusion model [8] and other state-of-the-art methods were used for comparison.

Training was performed on a high-performance computing platform with NVIDIA GPUs, using the Adam optimizer and a learning rate of 1e-4. The model was trained for 50 epochs, with batch size and other hyperparameters tuned based on validation performance. Evaluation metrics included the nuScenes Detection Score (NDS) and mean Average Precision (mAP), as defined in Section 2.3. NDS combines mAP with five True Positive (TP) metrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE).

3.2. Comparative experiments

Comparative experiments were conducted to assess the proposed model against the baseline CenterFusion model and other multimodal fusion approaches, such as the Multi-View Projection (MVP) model [5] and Camera-Radar Network (CRN) [6]. The results, evaluated on the nuScenes validation set, are summarized as follows:

· **Average Detection Accuracy**: The proposed model achieved a 1.5% higher mAP compared to the baseline CenterFusion model, demonstrating improved detection precision across object categories.

· **NuScenes Detection Score (NDS)**: The proposed model attained an NDS of 2.1% higher than CenterFusion, reflecting enhanced performance in localization, scale, orientation, velocity, and attribute estimation.

· **Long-Distance Target Detection**: Qualitative results, as shown in Figure 8 of the original paper, highlight the proposed model’s superior ability to detect distant objects. For instance, in visualization results (Figure 8(a2) and 8(b2)), CenterFusion failed to detect a white car at a long distance and two white cars in a complex road scenario, respectively. In contrast, the proposed model accurately detected these targets, validating its robustness for long-range detection.

Table 1: Comparison of detection accuracy for different object categories by the algorithm

Models	Detection Accuracy for Different Object Categories									
	Car	Truck	Bus	Trailer	Const.	Pedest.	Motor.	Bicycle	Traff.	Barrier
CenterNet	0.461	0.237	0.327	0.135	0.035	0.364	0.249	0.233	0.551	0.452
CenterFusion	0.525	0.265	0.368	0.148	0.054	0.388	0.303	0.227	0.563	0.471
Ours	0.534	0.269	0.371	0.156	0.065	0.421	0.345	0.242	0.576	0.479

3.3. Ablation studies

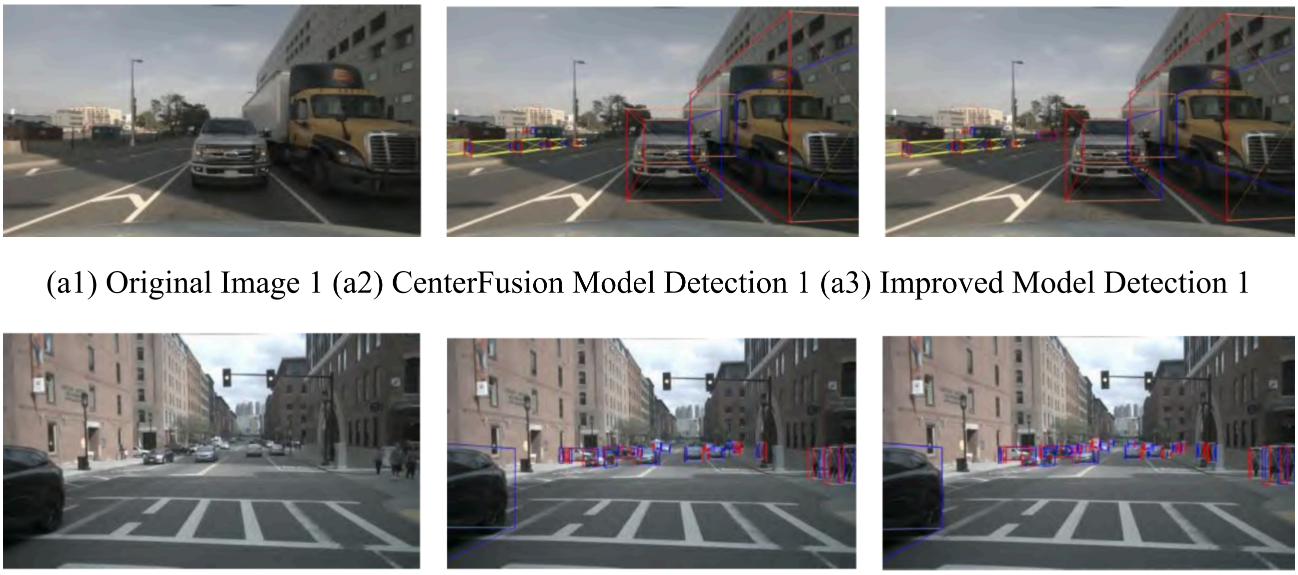
Ablation experiments were conducted to evaluate the contributions of each component in the proposed model. The following configurations were tested:

· **Baseline CenterFusion**: The original CenterFusion model without modifications.

· **Early Fusion Only**: Adding the early fusion strategy to map radar data onto the image plane.

· **Early Fusion + Attention Mechanism**: Incorporating the attention mechanism post-feature fusion.

· **Full Model**: Combining early fusion, attention mechanism, and optimized loss function.



(a1) Original Image 1 (a2) CenterFusion Model Detection 1 (a3) Improved Model Detection 1

(b1) Original Image 2 (b2) CenterFusion Model Detection 2 (b3) Improved Model Detection 2

Figure 3: Detection result figure

Table 2: Ablation experiments under different improvement schemes

Improvement Schemes	Early Fusion	LEPS A	Loss Function	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
1	×	×	×	0.452	0.331	0.649	0.263	0.534	0.543	0.143
2	√	×	×	0.459	0.338	0.633	0.261	0.528	0.531	0.135
3	√	×	√	0.463	0.339	0.617	0.258	0.527	0.529	0.133
4	√	√	×	0.470	0.345	0.609	0.251	0.524	0.505	0.131
5	√	√	√	0.473	0.346	0.593	0.249	0.523	0.498	0.128

Results demonstrated that each component incrementally improved performance. The early fusion strategy enhanced robustness against sparse radar data, the attention mechanism improved feature prioritization, and the optimized loss function mitigated sample imbalance. The full model achieved the highest mAP and NDS, confirming the synergistic effect of the proposed modifications.

4. Conclusion

This study proposes an enhanced 3D object detection model based on the CenterFusion framework, addressing the challenge of missed detections of long-distance targets in autonomous driving. By introducing an early fusion strategy, the model effectively integrates camera and radar data at the input stage, enhancing robustness against interference. The incorporation of an attention mechanism, inspired by advanced feature aggregation techniques, prioritizes critical information in the fused feature map, improving detection accuracy. Additionally, an optimized loss function mitigates the imbalance between positive and negative samples, further stabilizing training.

Experimental results on the nuScenes dataset demonstrate the proposed model's superiority over the baseline CenterFusion model and other state-of-the-art methods. The model achieves a 1.5% improvement in mean Average Precision (mAP) and a 2.1% increase in nuScenes Detection Score (NDS), with significant gains in detecting distant objects. Ablation studies confirm the effectiveness of each component, highlighting the importance of early fusion, attention mechanisms, and loss function optimization.

Future work will focus on addressing remaining limitations, such as sensitivity to extreme sparsity in radar data and computational efficiency for real-time applications. Additionally, integrating LiDAR data and exploring advanced Transformer-based architectures could further enhance detection performance in complex driving scenarios.

References

[1] ARNOLD E, AL-JARRAH O Y, DIANATI M, et al.A survey on 3D object detection methods for autonomous driving applications [J].IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[2] ZHANG Y P, LU J W, ZHOU J.Objects are different: flexible monocular 3D object detection [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, June 20-25, 2021: 3288-3297.

[3] ZHENG Z, YUE X, KEUTZER K, et al.Scene-aware learning network for radar object detection [C]//Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, China, August 21-24, 2021: 573-579.

[4] YANG H, WANG W, CHEN M, et al.PVT-SSD: single-stage 3D object detector with point-voxel transformer [C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, June 18-22, 2023: 13476-13487.

[5] YIN T W, ZHOU X Y, KRÄHENBÜHL P.Multimodal virtual point 3D detection [J].Advances in Neural Information Processing Systems, 2021, 34: 16494-16507.

[6] KIM Y, SHIN J, KIM S, et al.CRN: camera radar net for accurate, robust, efficient 3D perception [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, October 1-6, 2023: 17569-17580.

[7] NABATI R, QI H R.CenterFusion: center-based radar and camera fusion for 3D object detection [C]//2021 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, January 3-8, 2021: 1526-1535.

[8] NOBIS F, GEISSLINGER M, WEBER M, et al.A deep learning-based radar and camera sensor fusion architecture for object detection [C]//2019 Sensor Data Fusion: Trends, Solutions, Applications, Bonn, Germany, October 15-17, 2019: 1-7.

[9] ZHANG H, ZU K, LU J, et al.EPSANet: an efficient pyramid squeeze attention block on convolutional neural network [C]//Proceedings of the Asian Conference on Computer Vision, Macao, China, December 4-8, 2022: 1161-1177.

[10] GEIGER A, LENZ P, URTASUN R.Are we ready for autonomous driving?The KITTI vision benchmark suite [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, June 16-21, 2012:

3354-3361.

[11] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al.Scalability in perception for autonomous driving: waymo open dataset [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, June 13-19, 2020: 2443–2451.