IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

AttBEV: Enhancing Multi-Modal 3D Object Detection with CBAM Attention in BEVFusion for Autonomous Driving

Na Zhang, Edmundo Guerra, and Antoni Grau

***Abstract*—Multimodal fusion has an important research value in environmental perception for autonomous driving. Among them, BEVFusion has become one of the mainstream framework for LiDAR camera fusion by unifying multimodal features in the bird's-eye view (BEV) space. However, its performance is limited by inefficient cross-modal interaction and information loss during BEV projection, especially for dynamic objects and edge cases. To address these limitations, we propose AttBEV, an advanced fusion architecture that introduces a CBAM at the feature fusion layer: a lightweight attention mechanism that improves the model's ability to capture key information through dynamic feature calibration of channel and spatial dimensions. Extensive experiments on the nuScenes dataset demonstrate that AttBEV achieves superior performance compared to BEVFusion on most evaluation metrics. NDS reaches 0.6795, which is 2.63% higher than BEVFusion's 0.6532, and mAP reaches 0.6426, which is 1.79% higher than BEVFusion's 0.6247. In general, AttBEV outperforms existing methods in both model accuracy and generalization ability and significantly improves the performance of 3D object detection in autonomous driving scenarios.**

***Index Terms*—Intelligent Transportation Systems, Sensor Fusion, Computer Vision for Transportation.**

I. INTRODUCTION

THE environmental perception systems serve as the "eyes" of autonomous vehicles, enabling independent operation with minimal human intervention. Recent hardware and software breakthroughs have accelerated the development of autonomous driving in intelligent transportation, delivery services, and port logistics. Hardware components like LiDAR, radar systems, and high-definition cameras provide critical real-time environmental data, while advanced computer vision algorithms enhance detection efficiency and accuracy. Unlike traditional 2D detection methods that only identify object categories and 2D bounding box positions, 3D detection outputs complete 3D bounding boxes (x, y, z, length, width, height, yaw), directly providing accurate collision time prediction, space occupancy information for path planning, and target motion state estimation such as lane-cutting detection. This 3D capability is crucial for obstacle avoidance and path planning in autonomous driving systems.

This paper proposes an efficient 3D object detection method AttBEV, which fuses a LiDAR point cloud and camera image to overcome the limitations of single-sensor methods. To achieve the best performance, we designed a CBAM-Fuser feature fusion module, replaced the original feature fusion module with the proposed the CBAM [1] attention mechanism, applied SwinTransformer [2] for 2D feature extraction, SECOND [3] and DynamicSimpleVFE [4] for 3D point cloud processing; thereby improving the accuracy of 3D detection and ensuring real-time performance.

This paper uses the CBAM-Fuser feature fusion module and the DynamicSimpleVFE network to ensure real-time performance and accuracy. The main contributions are as follows:

- We propose AttBEV, a novel multi-modal 3D detection framework that enhances detection accuracy through attention-based fusion of LiDAR and camera data.
- We design a CBAM-Fuser module that uses channel and spatial attention to dynamically enhance the fusion of LiDAR and camera features.
- We introduce DynamicSimpleVFE to enhance computational efficiency, enabling real-time performance.
- Experiments on nuScenes demonstrate AttBEV's effectiveness and efficiency.

The rest of this paper is organized as follows. Section II reviews related work. Section III introduces the proposed detection method. Section IV shows the experimental results of the nuScenes dataset and evaluates the performance of the proposed AttBEV method. Finally, Section V concludes this paper.

II. RELATED WORK

A. CAMERA-ONLY 3D Object Detection

Contemporary multiview 3D object detection has emerged as a transformative technology in autonomous perception systems. By leveraging surround-view cameras positioned around vehicles, these systems generate comprehensive 360-degree environmental awareness. The core innovation lies in aggregating visual information from multiple perspectives and projecting extracted features into a unified bird's-eye view (BEV) representation. This consolidated spatial framework enables downstream applications including object localization, semantic understanding, and environmental mapping.

Current methodologies can be categorized into two primary architectural paradigms based on their feature processing strategies.

1) Depth-Prediction Based Framework: The LSS methodology [5] established the foundational approach through a sequential pipeline, depth probability estimation for multi-perspective features creates volumetric frustums, followed by geometric transformation into BEV coordinates using intrinsic camera parameters, culminating in task-specific processing within the unified spatial representation. This paradigm has been extensively refined by subsequent research [6].

This work was supported by Chinese Scholarship Council (CSC) under grant (202408440115).

The authors are with the School of Industry Engineering, Polytechnic University of Catalonia Barcelona, Barcelona, Spain. na.zhang@upc.edu

2) Attention-Driven Query Architecture: Drawing inspiration from transformer-based detection frameworks [7], recent innovations [8], [9] utilize learnable object queries that establish cross-modal correspondences with visual features through spatially-aware attention mechanisms. These adaptive query representations undergo iterative refinement and directly generate 3D bounding box predictions and classification outputs through lightweight multilayer perceptrons, bypassing explicit geometric view transformation procedures.

B. LIDAR-ONLY 3D Object Detection

LiDAR technology operates through precise time-of-flight measurements, emitting laser pulses and analyzing return signals to generate high-fidelity 3D point cloud representations. Despite higher costs than vision-based alternatives, LiDAR delivers exceptional advantages: sub-millimeter accuracy and consistent performance across challenging illumination conditions, including complete darkness.

Contemporary point cloud analysis frameworks are classified into four computational paradigms. Early projection approaches [10] flatten 3D clouds onto 2D planes for computational efficiency, enabling mature 2D convolutional architectures but relying on manually crafted features. Voxelization techniques [11] partition clouds into regular 3D grids while employing PointNet-derived architectures [12] for automatic feature learning. Direct point processing methods [13], [14] operate on raw unordered point sets, utilizing hierarchical architectures like PointNet and PointNet++ for efficient aggregation through strategic sampling, reducing computational overhead while preserving geometric fidelity. Recent hybrid innovations [15] integrate multiple paradigms within unified frameworks, capitalizing on synergistic benefits of point-level granularity and volumetric organization to achieve superior performance by combining geometric precision with computational efficiency.

C. MULTI-MODAL 3D Object Detection

Contemporary autonomous systems integrate heterogeneous sensors, combining visually interpretable data with precise LiDAR measurements. This fusion enhances perception beyond individual modalities and can occur at early (input), intermediate (feature), or late (decision) stages.

Early-stage fusion techniques often face significant computational overhead when generating dense virtual point representations from images. To address this, VirConv [16] introduces depth-guided voxel sampling and noise-resilient convolution operations, which effectively enlarge the 2D receptive fields while mitigating artifacts caused by depth estimation uncertainties. As an alternative paradigm under early fusion, data enrichment strategies such as AVFP-MVX [17] have been shown to enhance 3D object localization performance. Intermediate-stage fusion has attracted considerable research interest due to its favorable trade-off between computational cost and performance. For instance, LoGoNet [18] proposes a hierarchical local-to-global architecture that projects structured 3D region proposals onto image planes to facilitate cross-modal feature aggregation. In the context of Radar-LiDAR integration, Bi-LRFusion [19] tackles the sparsity of radar data by enhancing radar features with LiDAR information before consolidating them in BEV space. Meanwhile, MetaBEV [20] improves system robustness against sensor failures through an alternating modality training strategy.

The AttBEV method proposed in this paper belongs to mid-level fusion. Although the BEVFusion method effectively integrates the image and LiDAR features, its accuracy still needs to be further improved to ensure the safety of autonomous driving. Therefore, this paper attempts to improve the accuracy of 3D detection while ensuring speed, promote the development of multimodal fusion technology in the perception of the autonomous driving environment, and fully ensure the safety of autonomous driving.

III. PROPOSED METHOD

BEVFusion [21] is a mid-level fusion multi-modal 3D object detection framework that independently processes LiDAR point clouds and multi-view camera images before projecting them into a shared BEV space. The LiDAR branch uses HardVFE for voxel feature encoding and SECOND for sparse 3D feature extraction, generating 256-channel BEV features. The camera branch employs a Swin Transformer and DepthLSSTransform to produce 80-channel BEV features. The two modalities are fused by a simple ConvFuser that concatenates and compresses the features back to 256 channels, but this linear design limits the modeling capability of inter-modal interactions.

To address this issue, we propose AttBEV, which enhances feature complementarity through a lightweight CBAM-Fuser module, enabling more effective interaction between LiDAR and camera features. Meanwhile, HardVFE is replaced by DynamicSimpleVFE to improve efficiency on sparse point clouds and reduce unnecessary computation, thus achieving better real-time performance.

A. AttBEV

The overall architecture of our proposed AttBEV framework is depicted in Fig. 1, which builds upon but critically enhances the BEVFusion pipeline. The model first processes LiDAR point clouds and multi-view RGB images through separate branches. In the LiDAR branch, the raw point cloud is structured through voxelization using the DynamicSimpleVFE introduced to replace HardVFE, followed by efficient 3D feature extraction with a 3D sparse convolutional network (SECOND) and a feature pyramid network (SECONDFPN). The resulting 3D features are then projected downward to form the LiDAR BEV features, capturing precise geometric information. Simultaneously, the camera branch takes six surrounding views of RGB images and extracts rich 2D features using a SwinTransformer backbone. These perspective-view features are then transformed into a top-down representation through a view transformation module (DepthLSSTransform) to generate the camera BEV features, which carry strong semantic cues.

The key distinction between AttBEV and BEVFusion is the fusion stage. We replace direct feature summation/concatenation with a CBAM-Fuser module, which uses

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

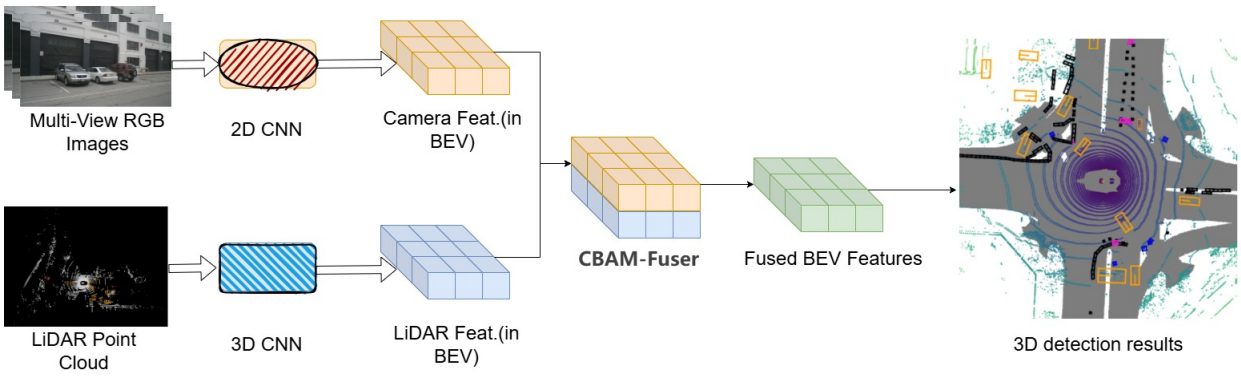


Fig. 1: AttBEV structure

channel and spatial attention to recalibrate LiDAR and camera BEV features. This selective fusion enhances complementary information and suppresses noise, generating superior fused features for the detection head and yielding significant performance gains.

B. CBAM-Fuser

This work introduces the CBAM to address key limitations in BEVFusion's fusion strategy. The design directly tackles two core issues: the inability to dynamically balance cross-modal features in the channel dimension and the failure to resolve spatial misalignments from depth errors. CBAM's sequential channel-spatial attention provides a targeted solution to these dual challenges.

Compared to alternative attention mechanisms, CBAM offers a uniquely efficient and appropriate solution. While SE attention [22] only addresses channel reweighting and self-attention [23] suffers from prohibitive computational complexity on high-resolution BEV maps, CBAM delivers both channel recalibration and spatial refinement in a computationally efficient manner, as seen in Equation 1. This enables the transformation from a static fusion process to a dynamic, adaptive one, achieving precise feature calibration while maintaining minimal computational overhead. The selection of CBAM is therefore justified by its structural suitability for addressing the specific weaknesses in the BEVFusion pipeline, rather than mere performance considerations.

$$F'' = Ms(F') \otimes (Mc(F) \otimes F) \tag{1}$$

Among them:

- Input feature map F,
- output of CBAM F'' ,
- $Mc(F)$: Channel Attention Weighting,
- $Ms(F')$: Spatial Attention Weighting,
- $\otimes$: Element-wise multiplication.

As shown in Fig.2, this module (CBAM-Fuser) is an efficient feature enhancement module that combines multimodal feature fusion with the dual attention mechanism. Its core process is divided into three stages: feature splicing, convolution fusion and attention optimization. First, the module receives multimodal input (such as features from different sources such as images and point clouds) and integrates information in the channel dimension through splicing (torch.cat). Subsequently, the basic convolutional layer (Conv2d + BN + ReLU) is used to nonlinearly fuse the spliced features, unify the dimensions and extract common features. Finally, channel attention (ChannelAttention) and spatial attention (SpatialAttention) are applied in sequence to dynamically calibrate feature weights from the channel dimension and spatial dimension respectively. Channel attention learns the importance between channels through global average pooling and fully connected layers, while spatial attention aggregates the global average and maximum response of features and generates spatial weight masks through convolution. This staged processing enables the module to gradually refine feature expression and significantly improve the fusion effect of multimodal data.

Based on the CBAM module diagram shown in Fig.3, the innovation of this module is primarily reflected in its dual-branch attention mechanism design and efficient feature refinement strategy. The Channel Attention branch processes input features through parallel MaxPool and AvgPool operations, followed by a shared MLP (Multi-Layer Perceptron) network. This design efficiently models inter-channel dependencies through lightweight fully connected layers with dimensionality reduction (reduction ratio = 16), avoiding computationally expensive operations while maintaining feature representation capability. The Spatial Attention branch takes the channel-refined features and applies both MaxPool and AvgPool operations along the channel dimension, then concatenates these dual-path features. A convolutional layer with a 7×7 kernel captures spatial contextual information across a wide receptive field, generating spatial attention weights Ms that highlight important spatial regions. The sequential application of these two attention mechanisms creates a "channel-first, spatial-second" attention cascade, enabling the model to adaptively focus on both important feature channels and critical spatial locations. This collaborative design allows for comprehensive feature enhancement - first identifying which feature channels are most relevant, then determining where in the spatial domain to focus attention. The modular design offers flexibility in implementation, supporting different configurations based on computational constraints and application requirements. Compared to traditional feature fusion approaches, this CBAM-based architecture achieves a good balance between computational efficiency and feature discriminability, making it particularly well-suited for real-time applications such as autonomous

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

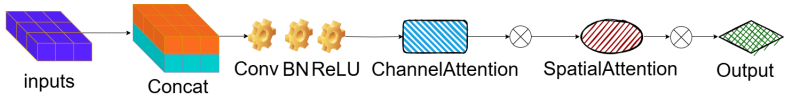
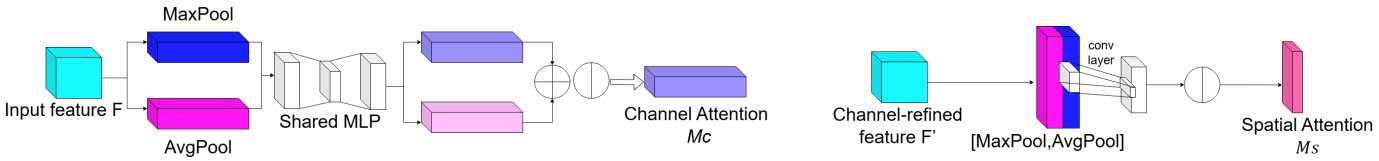


Fig. 2: CBAM-Fuser



(a) ChannelAttention (b) SpatialAttention

Fig. 3: The structure of CBAM

driving and multi-sensor fusion systems that demand efficient processing of heterogeneous data streams.

C. DynamicSimpleVFE

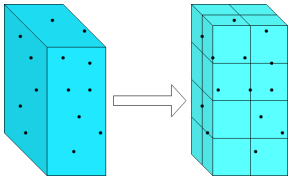


Fig. 4: DynamicSimpleVFE LAYER

DynamicSimpleVFE represents an improvement of the original Voxel Feature Encoding (VFE) methodology introduced in PointPillars [4]. This advancement addresses the fundamental challenge of transforming irregular, sparse point cloud data into structured representations suitable for neural network architectures. As illustrated in Fig.4, the technique operates within the 3D CNN module of our LiDAR processing pipeline, systematically converting unorganized point distributions into regularized voxel grids or BEV feature maps that facilitate downstream computational processing.

The enhanced framework delivers three main advantages over conventional approaches. First, the architecture streamlines the complexity of traditional voxel encoding, making it deployable on resource-constrained embedded systems such as NVIDIA Drive platforms. This computational efficiency enables real-time processing while maintaining detection accuracy. Second, the system dynamically modulates voxel dimensions based on distance-dependent requirements, employing fine-grained voxels for detailed near-field analysis while utilizing coarser discretization to minimize computational burden in distant regions. This adaptive spatial resolution optimizes the trade-off between computational resources and spatial accuracy. Third, statistical aggregation of points within individual voxels effectively suppresses outlier influences and single-point anomalies, improving overall signal integrity. This robust noise mitigation improves the stability of the features and the detection performance under challenging environmental conditions.

IV. EXPERIMENT AND ANALYSIS

A. Hardware Platform

The experimental platform runs Ubuntu 24.04 operating system, the hardware configuration is 8 GPU NVIDIA GeForce RTX 4090 model, GPU memory of 24 GB graphics card, Python version 3.8.20, Pytorch version 2.0.1, CUDA version 11.8, CuDNN version 8.7.

B. Dataset and Evaluation Indicators

The nuScenes [24] dataset is a large-scale multimodal dataset widely used in the field of autonomous driving. It was released by the Motional team in 2019. It is known for its rich sensor configuration, fine annotations, and diverse scenes. It is mainly used for the research and development and evaluation of tasks such as 3D target detection, tracking, and prediction. Among them, the sensor configuration in the multimodal data: 1 32-line laser radar (LiDAR), 6 cameras (covering 360° field of view, front/back/left/right/front left/front right), 5 millimeter-wave radars (Radar), inertial measurement unit (IMU) and GPS, all sensor data are strictly time synchronized (time alignment with high accuracy).

This paper uses the labelled part of nuScenes dataset, which covers 850 complete driving scenarios, and conducts multimodal fusion experiments using both image and LiDAR datasets. The data is divided into a training set consisting of 700 scenes with 28,130 frames, each corresponding to a LiDAR scan; and a test set consisting of 150 scenes with 6,019 frames, each corresponding to a LiDAR scan.

The model was trained using Adam Optimizer. The AdamW optimizer's initial learning rate was set to 0.0002. Average Precision (AP) is an object detection evaluation metric. This formula is given by equation (2):

$$AP_a = \frac{1}{N_a} \sum_{r_a \epsilon R_a} p(r_a) \tag{2}$$

In the paper, mean Average Precision (mAP) serves as the evaluation metric for three object detection categories. This metric is defined by equation (3):

$$mAP = \frac{1}{N} \sum AP_a \tag{3}$$

C. Comparison and Analysis of Results

The nuScenes dataset evaluates 3D detection using the nuScenes Detection Score (NDS) as the primary metric, which integrates mean average precision (mAP) and five error metrics: mATE for localization, mASE for scale, mAOE for orientation, mAVE for velocity, and mAAE for attributes.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: Comparison of overall performance indicators on the nuScenes validation set

METHOD	Sensors	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
Pointpillars [4]	Lidar	0.49078	0.34327	0.4239	0.2844	0.5293	0.3773	0.1936
SSN [25]	Lidar	0.49761	0.35167	0.42438	0.28495	0.49674	0.38269	0.19396
PGD [26]	Camera	0.39335	0.31736	0.7636	0.2668	0.4572	1.2849	0.1658
PointPainting [27]	Lidar+Camera	0.5810	0.4640	0.3877	0.2712	0.4958	**0.2466**	**0.1114**
3D-CVF [28]	Lidar+Camera	0.4978	0.4217	0.3001	**0.2455**	0.4576	0.2795	0.12225
BEVFusion [21]	Lidar+Camera	0.6532	0.6247	0.3014	0.2722	0.4522	0.3755	0.1909
AttBEV	**Lidar+Camera**	**0.6795**	**0.6426**	**0.2822**	0.2582	**0.3882**	0.2990	0.1901

Higher NDS and mAP values indicate better performance, while lower error metric values are preferred.

To study global performance AttBEV was compared with single-modality methods including PointPillars, SSN, and PGD, as well as multimodal fusion methods including PointPainting, 3D-CVF, and BEVFusion, demonstrating superior performance across core metrics, as shown on Table I. Its NDS achieved 0.6795 and its mAP reached 0.6426, significantly outperforming all compared methods. Compared to the BEVFusion baseline method, AttBEV improved NDS by 0.0263 and mAP by 0.0179. Compared to other multimodal methods, AttBEV outperformed PointPainting by 0.0985 in NDS and 0.1786 in mAP, and exceeded 3D-CVF by 0.1817 in NDS and 0.2209 in mAP. AttBEV also significantly exceeded all single-modality methods, with improvements of at least 0.17 in NDS and 0.29 in mAP over the best single-modality approach.

AttBEV also demonstrates superior performance in geometric estimation compared to BEVFusion, achieving reductions of 0.0192 in mATE and 0.014 in mASE. Most notably, AttBEV yields substantial improvements in orientation and velocity estimation, with reductions of 0.064 in mAOE and 0.0765 in mAVE. We attribute the improved orientation accuracy to the model's enhanced capability to suppress noise and extract cleaner spatial features, leading to more precise object boundaries and thus better orientation estimates. The improvement in velocity, while challenging to pinpoint without explicit temporal modeling, may stem from a more stable and accurate per-frame detection. The enhanced BEV representation and attention-based fusion strengthen spatial localization and temporal consistency, leading to more precise trajectory estimation in dynamic scenarios. AttBEV achieves mATE of 0.2822 and mASE of 0.2582, both ranking among the best results. These improvements validate the effectiveness of the CBAM-Fuser module in multimodal fusion for robust 3D object detection.

To study the positioning accuracy of our method Table II presents the AP of different object detection models across five object categories (Car, Pedestrian, Truck, Bus, and Traffic Cone) at varying detection distances. Compared with the strong baseline BEVFusion, our proposed AttBEV consistently demonstrates superior or highly competitive performance across almost all categories and ranges. For instance, in Car detection, AttBEV achieves an AP of 0.9156 and 0.9228 at 2.0 m and 4.0 m, respectively, which is very close to BEVFusion (0.9188 and 0.9289) while maintaining lightweight characteristics. More importantly, in Pedestrian detection, AttBEV surpasses BEVFusion at 0.5m ranges, reaching 0.8537 at 0.5 m compared to 0.8509, while achieving high performance at long distances.

A particularly notable advantage of AttBEV is observed

TABLE II: Average Precision for different object categories at various detection distances

Categories	Model	0.5m	1.0m	2.0m	4.0m
Car	PointPillars	0.6763	0.7949	0.8374	0.8525
	SSN	0.6773	0.7957	0.8335	0.8516
	PGD	0.1327	0.3918	0.6675	0.8168
	PointPainting	0.6640	0.7900	0.8220	0.8410
	3D-CVF	0.7420	0.8390	0.8630	0.8750
	BEVFusion	**0.7983**	**0.8903**	**0.9188**	**0.9289**
	AttBEV	0.7951	0.8873	0.9156	0.9228
Pedestrian	PointPillars	0.5729	0.5872	0.6076	0.6331
	SSN	0.5853	0.6000	0.6187	0.6410
	PGD	0.1003	0.3239	0.5613	0.7136
	PointPainting	0.6420	0.7220	0.7690	0.7960
	3D-CVF	0.7200	0.7350	0.7480	0.7660
	BEVFusion	0.8509	0.8649	**0.8749**	**0.8858**
	AttBEV	**0.8537**	**0.8649**	0.8747	0.8844
Truck	PointPillars	0.1936	0.3738	0.4673	0.5122
	SSN	0.2045	0.3766	0.4736	0.5122
	PGD	0.0049	0.0918	0.3054	0.5023
	PointPainting	0.2070	0.3530	0.4250	0.4470
	3D-CVF	0.2890	0.4500	0.5210	0.5400
	BEVFusion	0.2459	0.4084	0.5005	0.5560
	AttBEV	**0.3691**	**0.5531**	**0.6526**	**0.6959**
Bus	PointPillars	0.2140	0.5245	0.6600	0.6866
	SSN	0.2306	0.5076	0.6539	0.6873
	PGD	0.0058	0.1115	0.3952	0.6511
	PointPainting	0.1470	0.3620	0.4520	0.4860
	3D-CVF	0.2840	0.5150	0.5670	0.5870
	BEVFusion	0.3367	0.5997	0.7244	0.7782
	AttBEV	**0.4762**	**0.7069**	**0.8187**	**0.8501**
Traffic Cone	PointPillars	0.1414	0.1563	0.1836	0.2378
	SSN	0.1487	0.1665	0.1952	0.2432
	PGD	0.2596	0.5176	0.6738	0.7528
	PointPainting	0.5550	0.6080	0.6400	0.6930
	3D-CVF	0.6090	0.6190	0.6300	0.6590
	BEVFusion	0.7273	0.7415	0.7594	0.7838
	AttBEV	**0.7549**	**0.7650**	**0.7837**	**0.8051**

in large-object categories such as Truck and Bus. At 2.0 m, AttBEV yields 0.6526 for Truck and 0.8187 for Bus, substantially outperforming BEVFusion (0.5005 and 0.7244, respectively). This margin becomes even more significant at longer ranges (4.0 m), where AttBEV improves to 0.6959 (Truck) and 0.8501 (Bus), exceeding BEVFusion by over 13% in both cases. Furthermore, for Traffic Cone detection, AttBEV achieves consistently higher accuracy across all distances, reaching 0.8051 at 4.0 m compared with BEVFusion's 0.7838. Overall, AttBEV excels in challenging categories like Truck, Bus, and Traffic Cone while maintaining competitive performance on Car and Pedestrian. These improvements enhance perception robustness for autonomous driving systems.

Focusing on the AP scores at 0.5m range presented in Fig. 5, AttBEV demonstrates superior performance over the BEVFusion baseline across five out of ten object categories, substantiating the effectiveness of its attention mechanism in 3D object detection tasks. Specifically, AttBEV exhibits pronounced advantages in large vehicle detection: it achieves

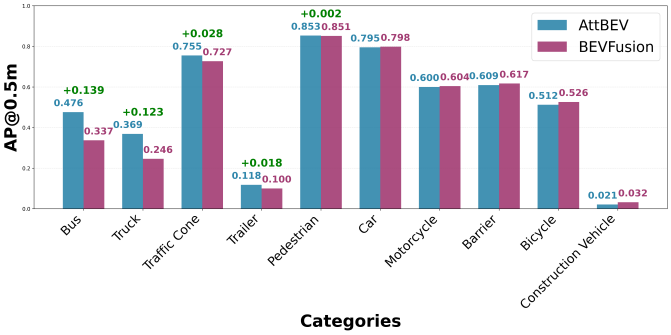


Fig. 5: Categories with 0.5m-range AP scores

an AP of 0.476 in the Bus category, surpassing BEVFusion's 0.337 by 0.139, representing a relative improvement of 41.2%; in the Truck category, it attains an AP of 0.369 compared to BEVFusion's 0.246, yielding a substantial relative gain of 50.0%. Furthermore, AttBEV outperforms in the Traffic Cone category by 0.028 (0.75 vs 0.727), maintains a marginal lead of 0.018 in the challenging Trailer category (0.118 vs 0.100), and achieves a competitive AP of 0.795 in the Car category. The attention mechanism is shown to make AttBEV focus more effectively on salient spatial features, improving representation of complex geometries and multi-scale objects. By adaptively allocating computational resources, particularly for large vehicles, AttBEV achieves significant accuracy gains while maintaining efficiency.

TABLE III: Comparison of model size and inference speeds obtained from various architectures

METHOD	Sensors	MEMORY(MB)	FPS(sample/s)
Pointpillars [4]	Lidar	**24**	7.8
SSN [25]	Lidar	24	8.7
PGD [26]	Camera	215	**15.1**
BEVFusion [21]	Lidar+Camera	468	6.7
AttBEV	**Lidar+Camera**	468	6.9

In terms of costs, we can conclude, based on the model size and inference speed comparison results presented in Table III, AttBEV achieves superior inference performance under the same memory footprint as BEVFusion, fully validating the effectiveness and efficiency of its attention mechanism. AttBEV attains an inference speed of 6.9 FPS compared to BEVFusion's 6.7 FPS, representing approximately 3% improvement that holds practical significance for real-time detection applications. More importantly, this performance gain is achieved under identical hardware resource constraints, with both methods consuming 468 MB of memory, indicating that AttBEV's attention module enhances processing efficiency through optimized feature extraction and fusion pipelines without introducing additional computational burden. Compared to unimodal methods, although multimodal fusion architectures require significantly higher memory consumption with Pointpillars and SSN requiring only 24 MB and PGD requiring 215 MB, AttBEV obtains more comprehensive environmental understanding through collaborative perception of LiDAR and camera, an advantage that is fully reflected in detection accuracy. It is noteworthy that while PGD achieves 15.1 FPS in inference speed, substantially outperforming multimodal methods, AttBEV's multi-sensor fusion capability acquired through partial speed sacrifice provides stronger robustness and accuracy in complex scenarios. Overall, AttBEV achieves a good balance among accuracy, speed, and resource utilization, offering a practical solution for autonomous driving systems.

D. Ablation Experiment

Based on the results of the ablation study in Table IV, the proposed CBAM-Fuser module brings significant performance enhancements while maintaining high efficiency.

In terms of detection accuracy, the complete CBAM-Fuser model achieves an NDS of 0.6795 and an mAP of 0.6426, substantially outperforming the BEVFusion baseline with NDS of 0.6532 and mAP of 0.6247. Through incremental module addition, the DynamicSimpleVFE demonstrates the largest contribution with an NDS improvement of 1.16% and FPS improvement of 0.2, while using channel or spatial attention modules individually shows limited or even slightly degraded performance, indicating that channel and spatial attention mechanisms need to work synergistically to achieve optimal results. Regarding category-specific performance, the complete model exhibits the most substantial improvements in detecting large vehicles such as Truck and Bus with improvements of 14.0% and 7.19% respectively, advancing from 0.5560 to 0.6959 for truck and from 0.7782 to 0.8501 for bus. Meanwhile, performance for Cars at 0.9228 and Pedestrians at 0.8844 remains highly competitive, and Traffic Cone detection improves from 0.7838 to 0.8051. These results validate the effectiveness of the attention-based fusion strategy in improving perceptual capacity, particularly for challenging large vehicle detection.

From a computational perspective, the integration of CBAM-Fuser brings some efficiency improvements, increasing the inference speed from 6.7 to 6.9 FPS while maintaining stable memory usage at 468MB. AttBEV balances detection performance with computational efficiency, improving the accuracy and real-time capabilities for autonomous driving applications.

E. Visualization of Test Results

To validate the trained AttBEV models, we visualize their detection results through testing, which include objects such as cars, pedestrians, and some distant and occluded objects.

Fig.6 shows the comparison of the detection results of BEVFusion and AttBEV in actual scenes. The picture is captured by 6 cameras. From the visualization effect, AttBEV shows obvious detection advantages. In particular, in the area marked by the red circle, AttBEV successfully identified and accurately located the pedestrian target in the distance, while BEVFusion failed to detect the target effectively. At the same time, in the area pointed by the green arrow, AttBEV provides a more accurate target bounding box, which is more consistent with the contour of the actual object. In complex environments (such as the urban street scene in the lower left corner), AttBEV's detection of multiple long-distance vehicles is also more stable, and the bounding box is more in line with the actual size of the target. In addition, from the

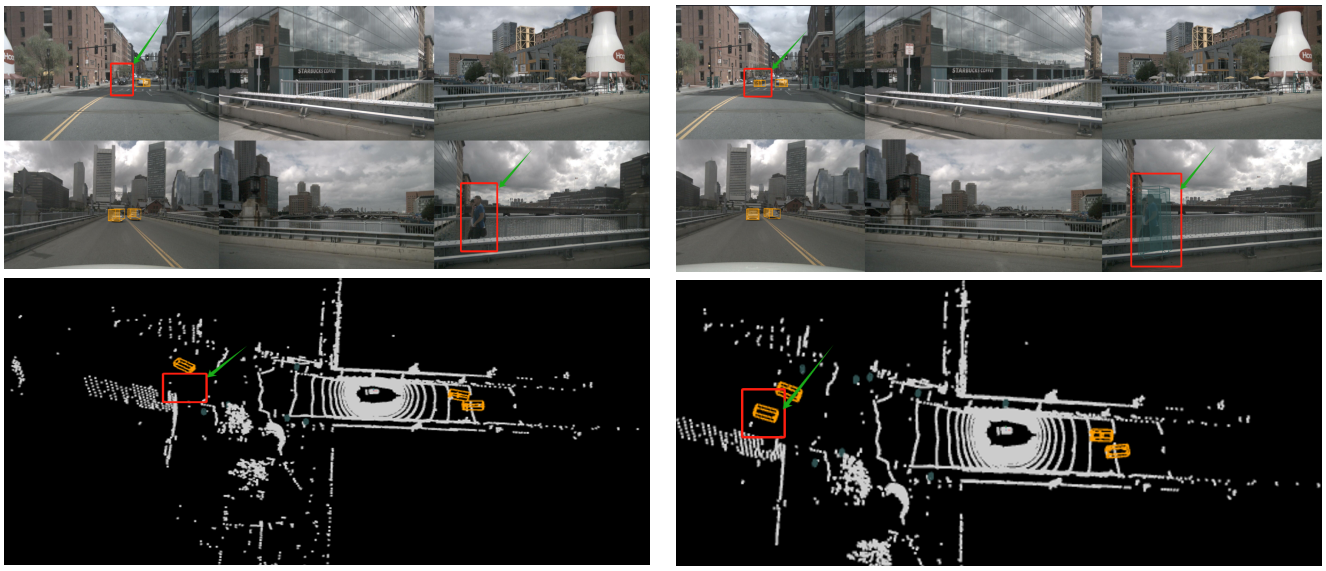
IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV: Impact of analysis modules on model performance

METHOD	NDS	mAP	Car	Pedestrian	Truck	Bus	Traffic Cone	MEMORY(MB)	FPS(sample/s)
BEVFusion [21]	0.6532	0.6247	**0.9289**	**0.8858**	0.5560	0.778	0.7838	468	6.7
+DynamicSimpleVFE	0.6648	0.6357	0.9265	0.8853	0.5893	0.829	0.805	**465**	6.9
+channel	0.6547	0.6342	0.9278	0.8852	0.6280	0.855	0.8021	468	6.9
+spatial	0.6519	0.6209	0.9279	0.8811	0.6029	0.854	0.7911	468	6.9
+CBAM-Fuser	**0.6795**	**0.6426**	0.9228	0.8844	**0.6959**	**0.855**	**0.805**	468	**6.9**

(a) BEVFusion (b) AttBEV

Fig. 6: BEVFusion and AttBEV detection results



(a) BEVFusion (b) AttBEV

Fig. 7: Predictions from BEVFusion and AttBEV in the nuScenes validation set.

overall detection results, AttBEV reduces the false detection rate while maintaining a high recall rate, and there is almost no bounding-box offset or inaccurate size that may exist in BEVFusion. These visualization results intuitively verify the superior performance of AttBEV in long-distance target detection, small target recognition, and precise positioning, and corroborate the aforementioned quantitative evaluation results, fully demonstrating that AttBEV has effectively improved its perception capabilities in complex scenarios through an improved multimodal fusion strategy.

The detection results in camera view and LiDAR view for BEVFusion and AttBEV are given in Fig.7, Fig.7(a) for the BEVFusion baseline model , and Fig. 7(b) for the AttBEV model. In particular, Fig. 7(b) accurately identifies cars and pedestrians in the image, with cars marked by orange 3D boxes and pedestrians marked by two blue boxes, and provides a more accurate 3D bounding box in the corresponding LiDAR point cloud view below of Fig. 7. In contrast, the car and the two pedestrians are not detected by the BEVFusion benchmark model in the red boxes of Fig. 7(a). It is particularly noteworthy that AttBEV shows richer detection results in the LiDAR point cloud view, successfully identifying and classifying multiple vehicle targets in the left region, including some targets in the sparse region of the point cloud, while BEVFusion's detection results in the same region are obviously insufficient and have missed detections.

In addition, AttBEV performs more balanced when dealing with targets at different distances, maintaining high detection accuracy and localization accuracy for both nearby pedestrians and distant vehicles. These visualization results strongly demonstrate that AttBEV effectively integrates the rich semantic information of the camera and the precise spatial information from the LiDAR through the improved multimodal feature alignment and fusion strategy, providing a more reliable technical solution for autonomous driving environment sensing.

V. Conclusion

This paper reviews recent advances in 3D object detection for autonomous driving and proposes AttBEV, a novel framework featuring a CBAM-Fuser module that enhances LiDAR-camera fusion through channel and spatial attention mechanisms. Experiments on nuScenes demonstrate that AttBEV achieves 67.95% NDS and 64.26% mAP, outperforming BEVFusion by 2.63% in NDS and 1.79% in mAP. The method shows improvements category-specific detection, particularly for challenging objects like bus, truck and traffic cones. These results are achieved while slightly improving reference performance at 6.9 FPS with 468MB memory, demonstrating an advance towards practical viability for autonomous driving applications. Despite these promising results, the model's computational footprint limits deployment on resource-constrained platforms. Future work will focus on model compression, temporal fusion to leverage motion cues, e.g.Spatial-Temporal Transformer, and robustness enhancement for diverse environmental conditions, aiming to advance AttBEV toward a production-ready system for real-world autonomous driving.

References

[1] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[3] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[4] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[5] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020.

[6] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[8] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[9] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[10] Martin Simony, Stefan Milzy, Karl Amendey, and Horst-Michael Gross. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[11] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *European Conference on Computer Vision*, pages 35–52. Springer, 2022.

[12] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[13] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11040–11048, 2020.

[14] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7463–7472, 2021.

[15] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023.

[16] Yang F Jin S, Li X P and Zhang W G. 3d object detection in road scenes by pseudo-lidar point cloud augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21653–21662, 2023.

[17] Xiaohan Wang, Jinhui Lan, Bingxu Wang, Chengkai Chen, and Shuai Chen. Avfp-mvx: Multimodal voxelnet with attention mechanism and voxel feature pyramid. *IEEE Sensors Journal*, 23(6):6139–6149, 2023.

[18] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023.

[19] Yingjie Wang, Jiajun Deng, Yao Li, Jinshui Hu, Cong Liu, Yu Zhang, Jianmin Ji, Wanli Ouyang, and Yanyong Zhang. Bi-lrfusion: Bi-directional lidar-radar fusion for 3d dynamic object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13394–13403, 2023.

[20] Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for bev detection and map segmentation. *arXiv preprint arXiv:2304.09801*, 2023.

[21] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781, 2023.

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[23] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[24] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020.

[25] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 581–597. Springer, 2020.

[26] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022.

[27] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020.

[28] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *European conference on computer vision*, pages 720–736. Springer, 2020.