CL-FusionBEV: A Cross-Attention Based Fusion Model for Camera and LiDAR in Bird's Eye View Perception

S. P. Samyuktha[1,*], S. Renuka[1], R. Shakthi Priyaa[1], Angel Meriba D. S.[1], Maheshwari M.[1], Megavarshini M.[1], S. Malathi[2,*]

[1] UG Scholar, Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, India

[2]Professor, Panimalar Engineering College, Chennai, India

Emails: spsamyu516@gmail.com; renu74496@gmail.com; shakthipriyaa16@gail.com; angelmeriba17@gmail.com; maheshwari23maniveeran@gmail.com; varshinimega14@gmail.com; malathi.raghuram@gmail.com

Abstract

In autonomous navigation, the ability to detect 3D objects from a Bird's-Eye View (BEV) perspective is essential. Nevertheless, many obstacles remain before LiDAR and camera data can be effectively combined. We propose CL-FusionBEV, a novel framework for sensor fusion that enhances Three-dimensional object recognition in the BEV domain. This method structures LiDAR point clouds for improved spatial feature extraction while converting camera data into BEV format via an implicit learning technique. An implicit fusion network and a multi-modal cross-attention mechanism facilitate seamless sensor interaction, ensuring comprehensive feature integration. Additionally, a self-attention mechanism of BEV enhances broad-scale reasoning and data extraction, improving the detection of occluded and distant objects. By efficiently synchronising data from several sensors, the suggested method improves feature uniformity and resolves spatial inconsistencies. It further leverages adaptive feature selection to enhance robustness against sensor noise and varying conditions. We evaluate CL-FusionBEV on the nuScenes dataset, achieving achieved a 73.3% mAP and a 75.5% NDS on the nuScenes benchmark, with vehicle and pedestrian detection accuracies of 89% and 90.7%, respectively. Our model demonstrates superior robustness in challenging conditions such as low visibility and dense urban environments. CL-FusionBEV maintains high efficiency with real-time inference, making it suitable for deployment in autonomous systems. Extensive experiments show our strategy routinely beats cutting-edge techniques, especially in detecting small and distant objects. By addressing key sensor fusion challenges in the BEV domain, CL-FusionBEV offers a notable advancement in Three-dimensional object recognition, ensuring high accuracy, efficiency, and reliability for real-world driving scenarios.

Keywords: BEV-based vision; Three-dimensional object recognition; Attention-based model; Self-drivin

1. Introduction

In the field of self-driving cars, for navigation to be both safe and operationally efficient, a precise sense of the environment is necessary. This process involves converting sensor inputs into semantic insights, such as identifying, localizing, and tracking road entities, including motorized vehicles, bicycles, and pedestrians. Among the several perceptual techniques, 3D object identification is a crucial step in the analysis and interpretation of complex 3D environments. The precise determination of an object's size and distance is vital for minimizing traffic incidents and improving proactive safety measures. Different techniques to 3D object identification using multiple sensor modalities have evolved as sensor technology advances. These include multimodal fusion approaches that combine many sensors to improve detection accuracy, LiDAR-based techniques that use point cloud processing for accurate location, and camera-based techniques that use image data. Additionally, the goal of Bird's Eye View (BEV) multimodal fusion techniques is to increase computing efficiency and accuracy by projecting data into the BEV perspective. These many methods provide a wide range of answers to the problems of 3D object recognition. In order to recover intricate visual components like forms and textures which are essential for 3D object recognition, monocular vision algorithms rely on pixel intensity. For instance, Mono3D [1] employs 2D detectors to identify candidate regions, refining them using predefined shape characteristics, spatial priors, contextual information, and semantic cues to localize objects. Similar to this, 3DOP [2] presents object identification as an

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

optimisation challenge that involves assessing ground depth and forecasting item sizes, whereas convolutional neural networks provide confidence ratings. Subsequent methods [3] enhance 3DOP's capabilities by generating and ranking 2D candidate regions through depth maps and monocular images. Other approaches address inaccuracies in 3D bounding box estimation caused by pixel depth errors, with methods such as pseudo-LiDAR converting depth maps into LiDAR-like points to enable precise detection using LiDAR-based techniques. This integration bridges depth prediction with robust 3D detection, enhancing the overall detection accuracy. LiDAR-based techniques, on the other hand, fall into three categories: voxel-based, point-based, and multi-view fusion techniques. These techniques mainly use point cloud data for 3D object recognition. Although they provide computational difficulties in real-time applications, voxel-based techniques, like VoxelNet, divide irregular point clouds into organised voxels for 3D bounding box prediction. Innovations like SECOND address these limitations by introducing sparse 3D convolution, reducing processing time and improving performance. Additionally, point-based [4] approaches like CenterPoint assign central points to objects, eliminating reliance on anchor boxes, while Part A2 employs a two-stage framework to refine detection box dimensions. Multi-view fusion techniques, such as PointPillars, transform point features into pseudo-images in BEV space through pooling operations, enabling efficient 2D convolution-based learning suitable for embedded systems with low latency requirements.

Point-based methods are effective in comprehending interactions within a point cloud and extracting local geometric properties. For example, PointNet learns individual point characteristics and global features by pooling using shared multi-layer perceptrons (MLPs) [5], while PointNet++ introduces architectural enhancements to capture local geometric attributes by aggregating adjacent point data. However, traditional sampling methods like Farthest Point Sampling (FPS) can be computationally expensive for large-scale point clouds, leading to the adoption of Random Point Sampling (RPS) to simplify the process. These methods often rely on color data rather than spatial information, necessitating the fusion of point clouds with image data for enriched feature representation, as seen in datasets like SemanticKITTI, which improve network performance through sensor fusion. In multimodal fusion for object detection, LiDAR point clouds are frequently superimposed on camera pictures, generating RGB-D data that 2D convolutional neural networks can handle. LiDAR points are projected onto RGB channels of both BEV and frontal views using techniques such as MV3D, pooling features from each to define 3D regions of interest. Refinements like AVOD improve feature representation by selectively sampling high-confidence regions for projection into feature maps, though convolutional stages may compromise small-object detection. Recent advancements include BEVFusion, which sets a benchmark by concatenating Using LiDAR point clouds and BEV features from surround-view cameras, the average detection accuracy was 71.7%. Its dependence on element-wise concatenation, however, restricts feature interaction and reduces the robustness of detection for distant and obscured objects. In contrast, by resolving these issues, our suggested approach, CL-FusionBEV, exhibits improved accuracy and robustness.

The proposed CL-FusionBEV framework incorporates LiDAR and camera data into the BEV space to improve 3D object identification for self-driving cars in intricate traffic situations. The following are the study's main contributions:

(1) Data Alignment Mechanism: Camera view features are aligned with BEV space via an implicit learning module while transforming LiDAR point clouds into BEV spatial features, enabling effective multimodal fusion.

(2) Self-Attention Mechanism: A self-attention mechanism improves comprehensive feature operations in order to get beyond the drawbacks of multimodal cross-attention, such as inadequate global reasoning and restricted feature interaction across spatial distributions.

(3) Excellent Detection Performance: Comprehensive tests on the nuScenes dataset demonstrate the effectiveness of the suggested framework, which achieved a nuScenes Detection Score (NDS) of 75.5% and a mean Average Precision (mAP) of 73.3%. Notably, it outperforms state-of-the-art methods with detection accuracies of 89% for autos and 90.7% for pedestrians. This research explores the relevant literature on 3D object identification, develops the suggested framework, and uses comparison and ablation experiments to confirm its efficacy. The results highlight CL-FusionBEV's potential as a reliable and accurate 3D object identification system in dynamic and intricate situations.

2. Related Works

Camera-Only Approaches: There has been a notable increase in interest in 3D object identification using camera-based techniques in the field of autonomous vehicle perception. Early techniques in this field primarily relied on 2D object detection outputs, augmented by depth estimation to enable 3D detection functionalities. Bird's Eye View (BEV)-based techniques have drawn a lot of interest lately for 3D object recognition. BEV-based methods, in contrast to previous methods, use multi-view camera inputs to immediately recognise 3D objects, obviating the need for intensive post-processing, especially when object detections overlap. One notable advancement is the Lift-Splat-Shoot (LSS) framework, which presented an end-to-end architecture that can use numerous camera pictures to derive BEV representations. In order to create a coherent representation, the LSS [7] process "lifts" each input picture into separate feature pyramids, which are then "projected" onto a rasterised BEV grid. Building on this framework, later research used depth estimation techniques to convert multi-view camera picture characteristics into the BEV space. A significant innovation in this area is the PETR

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

method, which employs Positional encoding and transformers are used to make BEV detection from camera data easier. Additionally, additional research has investigated the use of deformable attention mechanisms, which allow for the selective querying of pertinent local characteristics in the BEV space. This method guarantees a precise and effective conversion of camera-based data into BEV representations. When combined, these developments offer more efficient [9] and accurate methods for boosting 3D object identification, greatly improving autonomous cars' perception.

LiDAR-based Methods: Point-based, voxel-based, and Bird's Eye View (BEV)-based methods are the three general categories into which LiDAR-based approaches for 3D object recognition may be divided. Point-based methods, like PIXOR, extract useful characteristics from sparse LiDAR data and map them onto a 2D grid to create a 2D BEV representation. Techniques such as PointRCNN and STD directly handle raw point cloud data using PointNet and PointNet++, allowing for the extraction of global features that efficiently represent the environment's general geometric structure. In contrast, voxel-based techniques transform unprocessed point cloud data into a condensed voxel representation. These methods then advantage these voxels are then subjected to convolutional neural networks (CNNs) [11] in order to extract characteristics that are then utilised for 3D object identification. VoxelNet, for instance, creates features in the BEV space by compressing voxel characteristics along the Z-axis. In contrast, BEV-based techniques voxelize the raw point cloud and then use sparse 3D convolutions to extract voxel characteristics. Some methods provide intermediate features, which are then mapped to the BEV space for further processing. For instance, PointPillars divides the point cloud into many vertical pillars, collects information from each pillar, and extracts characteristics specific to BEV-based identification. Collectively, these approaches offer a range of LiDAR data processing methods, from voxelization techniques and BEV representations to direct point cloud analysis, improving the precision and effectiveness of 3D object recognition.

Multi-Modal Fusion for 3D Object identification: Recent years have seen tremendous progress in the field of 3D object identification, mostly due to developments in multi-modal fusion approaches. In this field, well-known techniques like F-PointNet, RoarNet, F-ConvNet, and PointFusion have established new standards. By combining information from point cloud and picture data, F-PointNet improves the accuracy of region proposals. On the other hand, RoarNet [12] leverages rotationally invariant features within RoI-aware frameworks to enhance detection reliability. F-ConvNet capitalizes on 3D convolutional architectures to achieve improved multi-modal data integration, while PointFusion employs advanced point cloud projection methodologies to facilitate more effective data fusion.

Despite these strides, certain challenges persist, particularly in detecting objects within complex environments where occlusion often undermines accuracy. As a result, 3D object identification techniques based on Bird's Eye View (BEV) are receiving more attention. These techniques create a thorough representation of multi-modal information by converting picture data into BEV features and combining them with LiDAR-generated BEV features. These methods might not, however, adequately capture the complex relationships between different modalities.

To address this limitation, methods like BEVFusion have emerged, [13][14] utilizing Transformer-based models for combining complementing data from cameras and LiDAR. While effective, these methods often rely heavily on the seamless cooperation between modalities, potentially reducing robustness when either LiDAR or camera data is unavailable.

Departing from these [15] approaches, this paper introduces the CL-FusionBEV method, which employs a voxelization strategy to achieve robust and accurate multi-modal fusion. In this framework, LiDAR data is converted into a unified grid structure, processed via 3D sparse convolution, and transformed into BEV space features. Simultaneously, camera features from multiple views are propagated Camera-based BEV space features are created by compressing the data along the Z-axis and putting it into the 3D voxel space. In order to improve the fusion process, self-attention mechanisms are used in the BEV space to encourage feature interaction on various scales, and cross-attention mechanisms are employed to aid BEV feature integration. With these advancements, the suggested CL-FusionBEV approach successfully addresses the challenges associated with multi-modal fusion for 3D object identification while exhibiting exceptional robustness and accuracy.

3. Proposed Method

A thorough description of the CL-FusionBEV framework, a 3D object identification method that successfully combines camera and LiDAR data within the Bird's Eye View (BEV) framework, is given in this section. By using the advantages of both senses, the technique improves detection accuracy and spatial awareness. Using VoxelNet and ResNet101 as the foundational models to provide rich feature representations, the procedure starts with the extraction of preliminary shallow features from both point cloud data and pictures. Prior to fusion, this first step guarantees that LiDAR and camera data are efficiently encoded. The LiDAR point cloud is next subjected to voxelization, which standardises its characteristics into an organised grid-like form that makes additional processing easier. To refine these voxelized features, 3D sparse convolution is employed, producing LiDAR BEV spatial features that preserve essential geometric details while optimizing computational efficiency. Simultaneously, important visual information is preserved by extracting and integrating dense multi-view picture characteristics into the 3D voxel space. Compression along the Z-axis is used to align picture features

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

with BEV representation, producing camera BEV features that correspond to the spatial structure of the LiDAR-based BEV features.

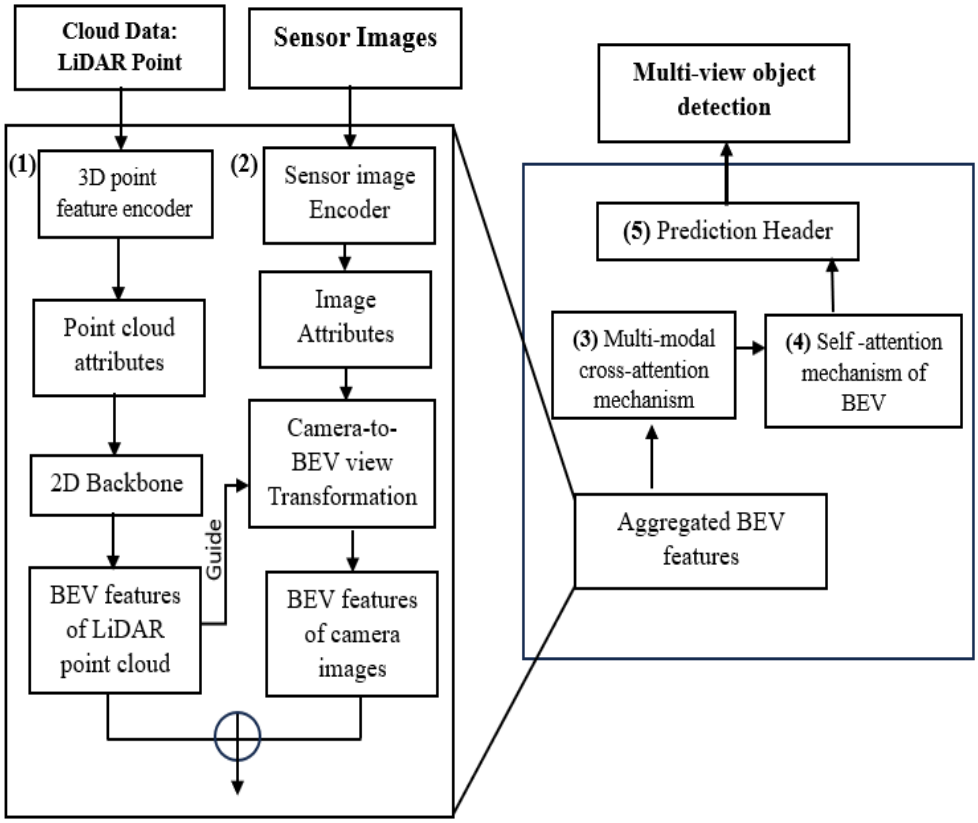

Figure 1. The illustration showcases the CL-FusionBEV architecture with five main components: (1) LiDAR feature extraction and BEV transformation, (2) Camera feature extraction and BEV projection, (3) Multimodal cross-attention (MCM) for LiDAR-camera fusion, (4) BEV Self-Attention Mechanism (BSM) for global feature enhancement, and (5) Prediction head and loss functions for optimized detection.

A cross-attention approach that computes the correlations between camera and LiDAR characteristics is proposed to facilitate efficient sensor fusion. This mechanism plays a crucial role in enhancing the integration of multi-modal BEV spatial data by tying object depth to feature representation. A BEV self-attention mechanism is also included to improve feature interaction on several levels. The model can effectively analyse objects at different distances and occlusions thanks to this component, which makes multi-scale feature fusion easier. The framework greatly increases 3D object detection's accuracy and resilience by implementing cross-modal and self-attention techniques in diverse scenarios.

Primary Feature Retrieval from LiDAR and Generation of Top-Down (BEV) Representations

A single unprocessed LiDAR point cloud frame usually consists of thousands of three-dimensional data points. Utilizing this data directly for fusion or network processing can impose a significant computational burden, as not all points contribute meaningful semantic information. The high density of raw point clouds can lead to redundant computations and inefficient memory usage, making direct processing impractical for real-time applications. Therefore, compressing the point cloud data into a more efficient feature set is necessary to ensure computational feasibility without compromising critical spatial details. To achieve this, a voxelization method is employed, which groups points into discrete spatial units to reduce complexity while preserving essential geometric structures. Initially, the Multi-Layer Perceptron (MLP) is used to enhance the representation of point cloud data by elevating its features to 512 dimensions. The data has dimensions of $n \times 4$, where n is the number of points and 4 is the x, y, z coordinates and reflectance intensity. The network's capacity to capture intricate spatial relationships within the scene is improved by this change. The input is then aggregated and the number of points is decreased using a max-pooling layer, which efficiently summarises important properties while removing noise.

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

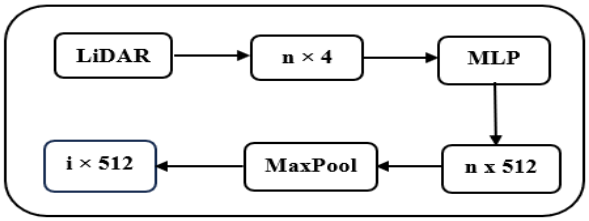

Figure 2. LiDAR initial feature extraction

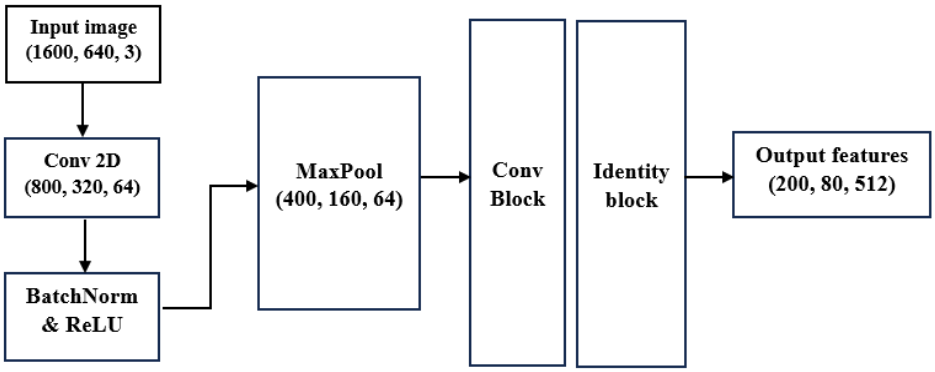

Figure 3. Initial feature extraction from the image

This preserves pertinent information while drastically cutting down on computational overhead, producing a condensed feature vector qi of size 1 × 512 for every point. To improve the 3D object detection process, the resultant point cloud feature vectors are combined to create the input "Query" for the BEV, where "Query" stands for the collection of feature vectors q1, q2,..., qi. These feature vectors will then be further examined or fused across modalities.

Primary feature extraction from the camera and BEV feature establishment:

Bird's Eye View (BEV) space features can currently be created from camera data using two primary methods: BEV-CV or camera to BEV view transformation. Elevate: (1) Splat Shoot (LSS) this technique creates an intermediate BEV representation by forecasting the depth distribution of picture features by estimating their depth using a neural network. (2) BEVFormer : this method use a spatiotemporal transformer to learn a single BEV representation by capturing spatial and temporal information using pre-formulated grid-like BEV queries. Both methods learn the translation from the camera view to BEV space through implicit supervision. In addition to using implicit supervision to generate camera BEV space features, our CL-FusionBEV fusion method builds on the state-of-the-art perception methodology BEVFusion [16]. This method converts input photos with multiple views. In addition, depiction finishing the image BEV space feature development.

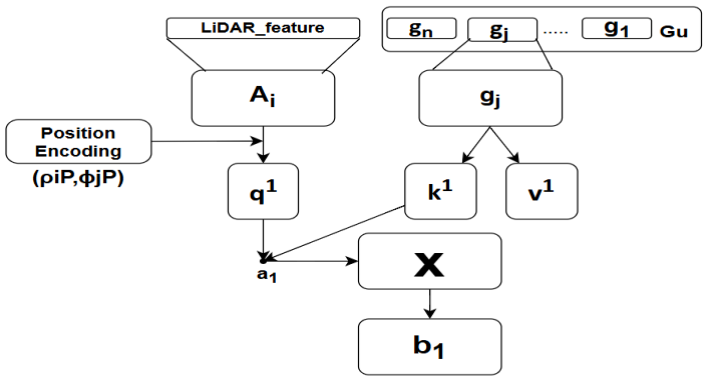

Figure 4. The multi-modal cross-attention mechanism's general architecture

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

Building a multi-modal cross-attention method for feature fusion (MCM) of LiDAR-Camera BEV:

The overall design of the multi-modal cross-attention system we have created is depicted in Fig. 5. We begin by linearly transforming the input feature matrices using a set of learnable weights, w1 and w2, taking into account the input BEV LiDAR features Ai and camera BEV feature gj. After that, position encoding is applied to the BEV LIDAR features (ρiP, ØPj), and the encoded features are assigned the designation q1. Similarly, the linearly translated camera attributes gj are assigned to k1 and v1, respectively. After that, qi is used to query the correlation of each image feature sequence kj, resulting in a correlation matrix ai where the normalisation factor

$$a_i = \frac{q_i k_i}{\sqrt{D}} \in R^{H \times W \times H \times W} \quad (1)$$

where the numerical range is constrained using the normalisation factor √D. The location of the maximum correlation matrix ai can be used to determine the relationship between LIDAR and picture features. The SoftMax operation is then used in the article for normalisation and differentiation:

$$m_i = \text{softmax}(a_i) \in R^{H \times W \times H \times W} \quad (2)$$

The final attention feature is also generated by computing a weighted sum of the normalised weights mi and the original linear transformation vi of camera features gj. When computation yields the final attention feature bi, the query update process for the BEV features is finished.

Modal information from the camera and LiDAR is seamlessly integrated during this process. Any combination of sensors can be used with our suggested modality-agnostic feature fusion algorithm. It allows for variable customisation of the number of cross-attention mechanism enquiries and the output feature dimensions, depending on the size of the picture features and the specific requirements of the work.

BEV self-attention mechanism (BSM):

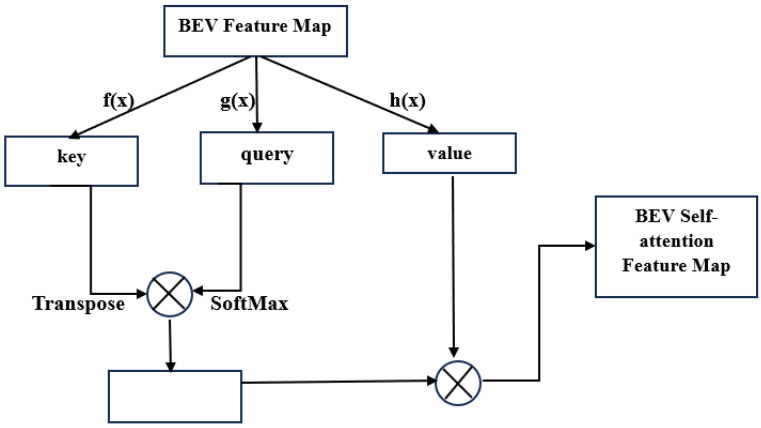

Figure 5. BEV self-attention mechanism flow chart

Features with diverse spatial distributions cannot interact because the context component of the multimodal cross-attention mechanism does not properly use global reasoning. At this point, the characteristics lack comprehensive global rationale and are more likely to convey local information. We to tackle this obstacle, a Bird's Eye View (BEV) self-attention mechanism for extensive global feature operations was designed. By identifying the contextual locations of the fused features across the entire BEV architecture, this approach makes simpler to create aggregated information on the shapes of pertinent objects. As the instrument for global feature interaction, we have opted for a self-attention mechanism that relies on soft attention. Fig. 6 displays the flowchart for the BEV self-attention mechanism. Three linear transformations are first applied to the original BEV features in order to produce query, key, and value feature sequences. Weights are then determined by calculating the similarity with the query using the transposed key. These weights are then normalized using the SoftMax procedure. A weighted summation is used for integrating the normalized weights with the appropriate values in order to generate the final self-attention feature map.Concentrate on self-mechanism can be expressed mathematically,

$$\text{Attention}(q,\ k,\ v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d}}\right)v \quad (3)$$

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

where q stands for the query for the locations and point cloud items that are being processed at the moment. The value of each element in the series is denoted by v, while the key matrix corresponding to each camera BEV feature element in the sequence is indicated as k. In order to keep the gradient from vanishing or blowing up, the attention score is calculated by scaling the dot product using the key's dimension, represented by d. In addition, k^T is the key's transposition, allowing q and k to execute the dot product operation.

Prediction header: As our prediction module, we have selected PointPillars' object detection head [10], which feeds a Single Shot Multi Box with the fused Features of the Bird's Eye View (BEV) for 3D object detection use the Detector (SSD) [18]. The SSD object detecting head uses convolutional neural networks and is a one-step process. It improves the real-time performance of object detection by concurrently doing object classification and regression within the network architecture. The SSD framework efficiently advances the network's training process by using a variety of loss functions to improve object position and categorization prediction. This method is ideal for real-time 3D object detection requirements applications because it enables more accurate and efficient optimization.

Loss functions: Three different loss functions are used for assessment in the object classification loss, 3D bounding box regression loss, and azimuth classification loss of the CL-FusionBEV fusion approach presented in this study. for 3D object boxes. The total of these separate loss components is the network's composite loss. We have chosen to use the Focal Loss function for in order to manage challenging data and balance the positive and negative sample distribution. By concentrating on difficult-to-classify The Focal Loss specifically improves the network's learning ability by solving the issue of class imbalance and giving samples from a wide variety of sample data.

$$L_{cls} = -\alpha_a (1 - p^a)^\gamma \ln(p^a) \tag{4}$$

where αa is the weight of prediction box a, which balances the significance of various prediction boxes, and L cls is the classification loss. p a stands for the likelihood of the prediction box a. A hyperparameter called γ is utilized to modify the loss function's form. γ is 2.0 while α is 0.25.

$$\begin{aligned} &\Delta x = \frac{x^{gt} - x^a}{d^a}, \ \Delta y = \frac{y^{gt} - y^a}{d^a}, \ \Delta z = \frac{z^{gt} - z^a}{d^a} \\ &\Delta w = \ln \frac{w^{gt}}{w^a}, \ \Delta l = \ln \frac{l^{gt}}{l^a}, \ \Delta h = \ln \frac{h^{gt}}{h^a} \\ &\Delta\theta = \theta^{gt} - \theta^a \end{aligned} \tag{5}$$

where l, w, h stands for the 3D bounding box's dimensions, θ for rotation, and (x, y, z) for the location of the 3D bounding box's center point.
angle. The 3D bounding box's ground truth values are indicated by $*^{gt}$, and the predicted values are indicated by $*a$. The regression residuals are computed using the formula d a } √(wa)2 + (l a)2. This study uses the Smooth L1 loss function to calculate the geometric loss, which yields the 3D bounding box regression loss, in order to avoid excessive loss when mistakes are large.

$$L_{reg} = \sum_{b \in (x, y, z, l, w, h, \theta)} \text{Smooth} L1(\Delta b) \tag{6}$$

where Δb is a vector representing the difference between the ground truth and the expected values of the 3D bounding box, of the form (Δx, Δy, Δz, Δl, Δw, Δh, Δθ). A 3D bounding box orientation regression loss function is used in this investigation. represented as Lreg_θ, which is defined as follows because orientation regression loss can have a substantial impact on model training precision:

$$L_{reg_\theta} = \text{Smooth} L1\left(\sin\left(\theta^{gt} - \theta^a\right)\right) \tag{7}$$

where θ^{gt} is the object's actual orientation and θ^a is its expected orientation. The original regression loss gets closer to zero when $\theta a = \theta^{gt} \pm \pi$, reducing the impact of opposing object orientations and improving model training efficacy. We use the the orientation regression loss in 3D bounding boxes by treating oppositely oriented predicted boxes as identical and using the cross-entropy loss function, denoted as Ldir, to train the orientation categories. The 3D object prediction head is the source. More accurate orientation category predictions are the goal of this method.

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

$$L_{dir} = -\theta_{dir}^{gt} 1_b(\theta_{dir}^a) - \left(1 - \theta_{dir}^{gt}\right) 1_b(1 - \theta_{dir}^a) \qquad (8)$$

where dir is the expected azimuth category and θ gt is the ground truth azimuth angle. When the condition enclosed in parenthesis is true, the indicator function 1b returns 1; otherwise, it returns 0 The weighted combination of the four previously mentioned loss functions creates the CL-FusionBEV method's ultimate total loss function:

$$L_{all} = \lambda_{cls} L_{cls} + \lambda_{reg}\left(L_{reg} + L_{reg_\theta}\right) + \lambda_{dir} L_{dir} \qquad (9)$$

4. Results and Discussion

Through thorough testing experiments, we hope to validate our framework's performance in this part. Additionally, we will carry out ablation studies to validate the efficiency, adaptability, and resilience of the model's constituent part**s.**

For our tests and analyses, we have chosen the nuScenes dataset [19], which is intended especially for scene perception in autonomous driving. Both Singapore and Boston's complex road environments are captured in this dataset. More than 400,000 frames of data and 1000 scenes make up this collection. To gather multimodal data, such as radar data, LiDAR point clouds, and 360-degree panoramic photos, nuScenes employs six sensors. The dataset offers thorough annotations for important components like traffic signs, automobiles, pedestrians, and 123 Complex & Intelligent Systems (2024) 10:7681–7696 7689. NuScenes provides a wide selection of scenarios and data for various environmental perception tasks, such as 3D object detection, tracking, and high-definition (HD) map production, in addition to thorough annotations.

The nuScenes Detection Score:

We employ the mean Average Precision (mAP) and nuScenes Detection Score (NDS) for evaluation.measures supplied by the nuScenes dataset [17] [19], to verify the efficacy of our suggested approach. Bird's-eye view center distances D=0.5,1,2,4 meters and the class set CC are used to compute the mean accuracy across matching thresholds, or mAP. Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE) are among the object attribute detection results that NDS combines with mAP.

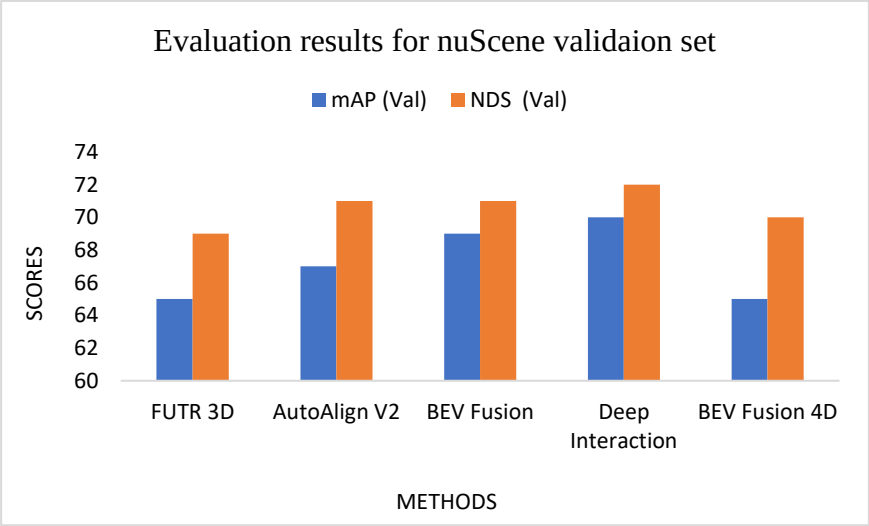

Figure 6. Evaluation results for nuScene validation set with comparison to the existing methods

Mean Average Precision

Where Average Precision mean is denoted by mAP. T Pc is the collection comprises five average category c error metrics, whereby A PC, d stands for the Average Precision for a specific category c and difficulty level d.As a result, NDS evaluates more than just detection performance; it also assesses the quality of detection by taking into account characteristics, velocity, size, orientation, and bounding box position. This comprehensive evaluation makes it possible to evaluate 3D object detection findings in-depth. Additionally, we provide thorough comparisons of the results for each of the ten detection categories. high-level semantic characteristics found in pictures. We have selected VoxelNet the LiDAR branch's backbone network for point cloud feature learning [6]. VoxelNet first splits the point cloud in order to extract spatial information.use intensive 3D convolution after entering a voxel grid.VoxelNet can deal directly with raw point clouds, avoiding data loss. In contrast to other point cloud processing networks. We set the image resolution to 1600 x 640 pixels

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

during the testing phase. We fixed the voxel size for the LiDAR point cloud to (0.075 m, 0.075 m, 0.2 m) in accordance with earlier works [6, 10, 37, 39]. An Ubuntu 18.04 server with an Intel Core i7-10,700 CPU and a GeForce RTX is used for our training and inference procedures. Python 3.7 is used to conduct the experiments, while the To build the models, the PyTorch deep learning framework is utilized. We enhance the network with a learning rate of 2e-4 and a weight decay of 1e-2.specifications using the AdamW optimizer.

We compared our suggested approach with state-of-the-art multimodal fusion 3D object detection techniques, assessing performance on nuScenes [19] test and validation sets. As shown in Table 1, our approach outperforms state-of-the-art AutoAlignV2 [21], BEVFusion [16], BEVFusion [20], and DeepInteraction with regard to the nuScenes validation set's detection performance. Specifically, our method shows a noteworthy 2.4% increase in mAP and 1.8% increase in NDS detection accuracy in comparison to DeepInteraction, a multimodal fusion technique. Furthermore, with improvements of 1.4% and 1.5%, respectively, our mAP and NDS detection scores outperform the state-of-the-art BEV fusion approach BEVFusion4D. In contrast to BEVFusion, a highly sophisticated BEV fusion technique, with even larger gains of 2.7% in mAP and 2.3% in NDS, our CL-FusionBEV approach represents a substantial breakthrough in multimodal BEV. 3D object detection using fusion. The test's validation outcomes set are shown in Table 2, where we contrast our approach with earlier methods that rely only on LiDAR. Our methodology achieves detection scores of 75.5% in NDS and 73.3% in mAP on the test set, which is a significant improvement over these earlier methods. According to Fig 6, our approach outperforms earlier cutting-edge techniques in the majority of detecting categories. Our suggested the MCM module, the network might offer BEV fusion capabilities that are responsible for this overall performance improvement. The spatial priors of the combined camera and LiDAR features in the BEV space are highlighted in both modules. Furthermore, our approach improves the detection capabilities for commonplace things including bicycles, people, Lorries, buses, and vehicles.

Examining Fig. 8 shows that adding the BEV-CV module improves mAP and NDS values in every category, demonstrating the potency of our suggested BEV-CV fusion method. The accuracy of 3D object detection has significantly increased, especially for automobiles, people, and bicycles, with improvements of 1.0%, 1.8%, and 0.3%, respectively. Our approach results in a 1.1% increase in NDS detection scores of 0.7% and a rise in mAP relative to the baseline. The point cloud's sparsity, which limits the amount of meaningful information for far items as well as causes the baseline network to perform less well in 3D object detection, is most likely the cause of these gains. Additional semantic information from photos is needed for these far-off items. By adding the MCM module, the network may provide BEV fusion features that are more in line with actual distributions, improving the detection accuracy of our approach even further. In particular, the detection accuracy of bicycles, vehicles, and people increased by 1.1%, 0.2%, and 0.9%, respectively.

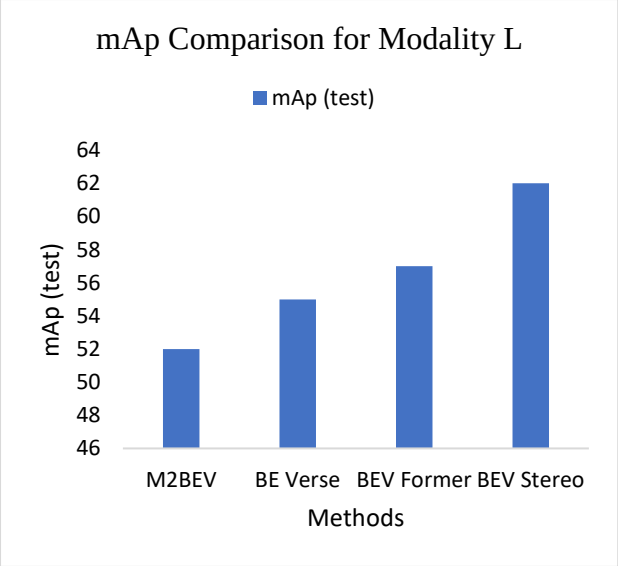

Figure 7. mAp Comparison graph for modality L

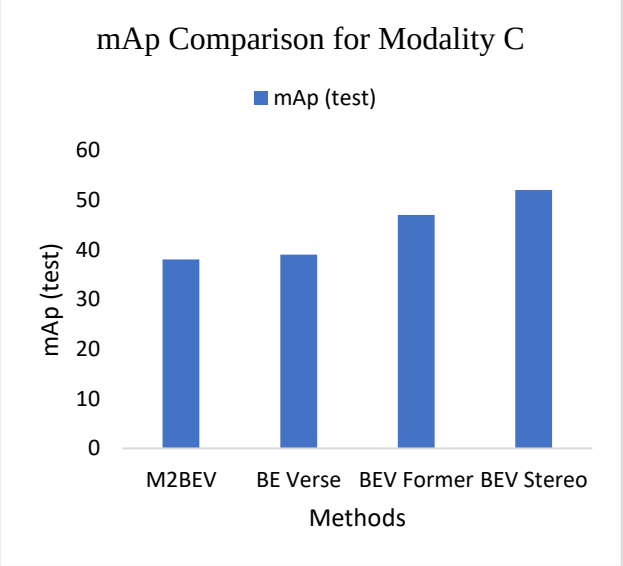

Figure 8. mAp Comparison graph for modality C

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

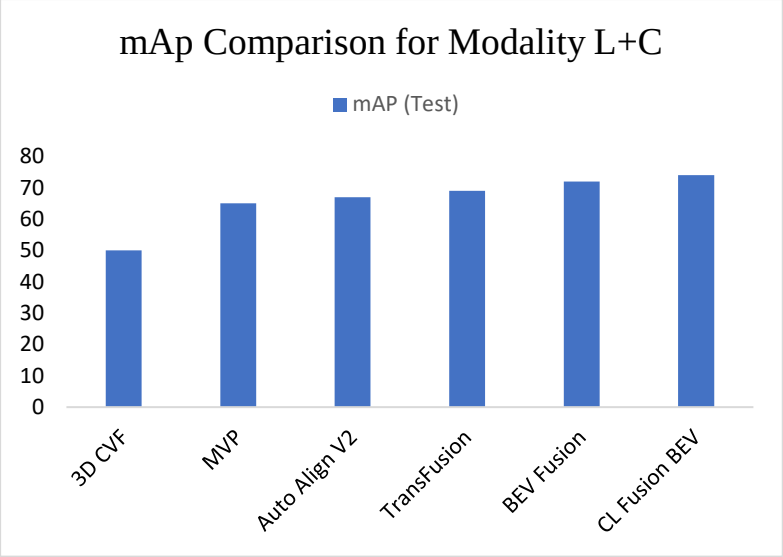

Figure 9. mAp Comparison graph for modality L+C

Nevertheless, there is only a slight improvement in detecting accuracy for tiny items like bicycles and pedestrians. This might be Due to the lack of global context reasoning in the BEV fusion features generated by the MCM, features from various spatial distributions interact with one another. Consequently, the accuracy of small objects is improved less noticeably by BEV fusion features, which only supply localized information rather than full global reasoning. BEV fusion features can now infer their introduction of the BSM module, which gives these features a global operation, allowed for placements throughout the entire BEV. Arrangement. This helps to further improve our method's detection effectiveness by making it easier to generate aggregated information on the shapes of pertinent items. Bicycles saw improvements of 2.4%, automobiles by 0.6%, and pedestrians by 0.4%. Our strategy increased mAP by 1.1% and NDS detection scores by 1.2% when compared to the baseline. The efficacy of every module in our network architecture is convincingly demonstrated by this set of ablation trials.

5. Computational and Thematic Evaluation

1. Verification of multimodal cross-attention mechanism and bev self-attention mechanism effectiveness.

Figure 6 provides a detailed visual representation of the multimodal cross-attention and BEV self-attention mechanisms' efficacy. The attention feature map produced by the multimodal cross-attentio. The attention feature map produced by the multimodal cross-attention is displayed in Figure 6a. Which is unclear, indicating that the camera and LiDAR BEV elements may not be integrated as well as they may be. On the other hand, the BEV self-attention mechanism generates a more perceptive attention map in Fig. 6b. Stronger feature correlations are made possible by this approach, which encourages efficient interaction between the fused BEV features. Figure 6: Attention feature map comparison. As seen in Fig. 6, we created comparable attention feature maps and chose forecasts with high confidence scores to enhance visualization. Higher values on the attention maps suggest a stronger level of significance. While b shows the maps created by the BEV self-attention mechanism, a show the attention feature map created by the multimodal cross-mechanism of focus. Interestingly, the map in b exhibits more clearly discriminative features, indicating a higher level of ability to identify pertinent places. In contrast, the BEV self-attention mechanism in b more successfully highlights important aspects, maintaining and improving relevance, whereas the attention map from the multimodal cross-attention in a seems less distinct. This comparison highlights our method's notable advantage in feature attention. Figure 6: Attention feature map comparison. As seen in Fig. 6, we created comparable attention feature maps and chose forecasts with high confidence scores to enhance visualization. Higher values on the attention maps suggest a stronger level of significance. While b shows the map created by the BEV self-attention mechanism, a shows the attention feature map created by the multimodal cross-attention mechanism. Interestingly, the map in b exhibits more clearly discriminative features, indicating a higher level of ability to identify pertinent places. In contrast, the BEV self-attention mechanism in b more successfully highlights important aspects, maintaining and improving relevance, whereas the attention map from the multimodal cross-attention in a seems less distinct. This comparison highlights our Attention. Figure 7 shows the nuScenes test set's comparative experimental detection findings.

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

2. Comparison of 3D object detection results from the BEV perspective.

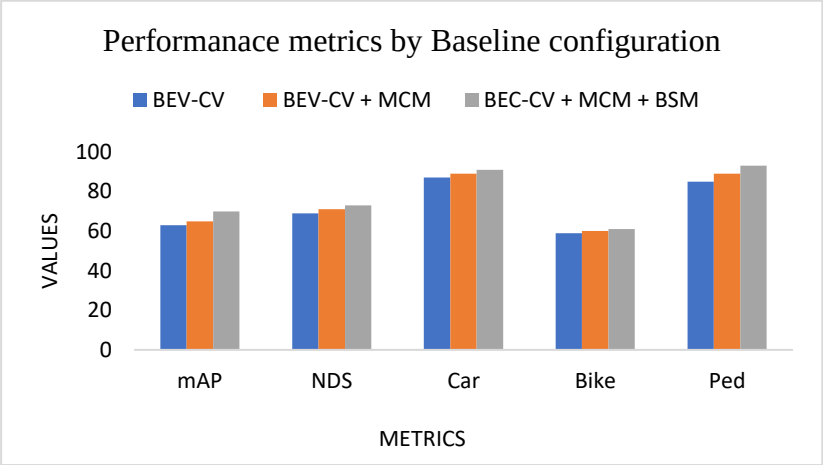

Figure 10. Performance metrics by baseline configuration comparison graph

We deliberately chose BEVFusion as a sample 3D object identification approach and performed comparative experiments with other state-of-the-art methods to further validate the efficacy of our proposed CLFusionBEV method. A qualitative comparison of our CLFusionBEV and BEVFusion is shown in Figure 7. As can be seen, BEVFusion has several restrictions when it comes to identifying far-off and obscured objects, most likely because of its shortcomings in feature fusion and context awareness. In contrast, our CL-FusionBEV approach effectively enhances the feature ability to express and the model's ability to reason globally by incorporating a multi-modal cross-attention function with a BEV self-attention mechanism, showing significant advantages in these challenging scenarios.

We use red circles to highlight a number of common overlooked objects in Fig. 7. These items, which BEVFusion's detection findings did not correctly identify, are obviously noted in the results of CL-FusionBEV. This demonstrates that CL-FusionBEV provides more dependable identification results due to its increased resilience when dealing with occlusion and long-distance objects. By means of these compared tests with cutting-edge technologies, we demonstrate the CL-FusionBEV method's potential and benefits for real-world applications in addition to confirming its efficacy.

3. 3D recognition of objects results are visualized from a BEV standpoint

We provide visualizations demonstrating the 3D detection outcomes using CL-FusionBEV in several settings, as illustrated in Fig. 8, to confirm further the effectiveness of the suggested network in this paper. The nuScenes test set's qualitative detection results are Six cameras in the nuScenes dataset are used to take the six images in the figure that represent the front-left, front-center, front-right, rear-left, rear-center, and rear-right perspectives. The object detection findings obtained using the LiDAR clouds of points from the BEV perspective are displayed in the image on the far right. Vehicles (filled circle), motorbikes (filled circle), trucks (filled circle), bicycles (filled circle), people (filled circle), and buses (filled circle) are all represented by a variety of colorful bounding boxes.

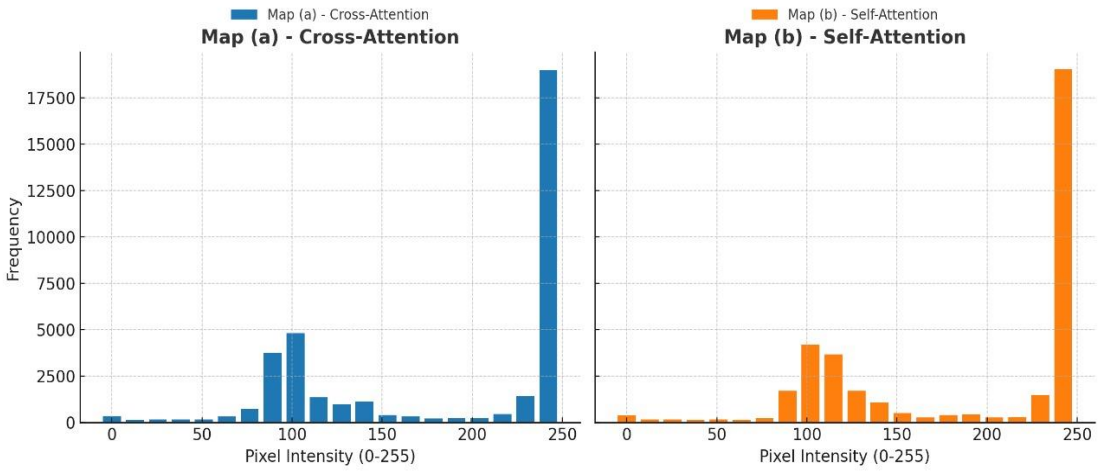

Figure 11. Performance metrics by baseline configuration comparison graph

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

CL-FusionBEV's smooth integration of the BEV self-attention mechanism and multimodal crossattention mechanism allows for efficient use of depth-interacted BEV fusion features, improving the detection of occluded objects. Even when there are partial or full Overall, our approach continues to exhibit commendable performance, demonstrating its adaptability and efficiency in challenging environments. Additionally, Figs. 8c, d, and f show that our approach continues to provide reliable detection performance in high-traffic situations, such as city streets and intersections with a significant volume of vehicles. It also has the ability to recognize the relevance of our suggested approach for 3D object detection in intricate traffic scenarios is highlighted by the inclusion of small things like bikers and pedestrians. These outcomes demonstrate our method's persuasive dependability while handling difficulties in complicated scenarios, in addition to highlighting the notable gains in detection performance it has attained.

6. Conclusion

The advanced 3D object recognition method CL-FusionBEV, which we provide in this work, efficiently integrates camera and LiDAR data using BEV) perspective. Our main goal is to improve object detection's accuracy and resilience for autonomous automobiles, especially in situations with fluctuating traffic. We accomplish this by improving the way data from many sensor modalities is integrated. The first step in our method is to construct BEV spatial features. By deftly incorporating rich multi-view picture information within the 3D voxel space of characteristics and compressing them along the Z-axis, we generate camera-derived BEV spatial features. Voxelization is then used to convert raw LiDAR-captured point cloud information into a homogeneous gridded representation.

3D sparse convolution techniques are then used to improve this representation, producing accurate spatial aspects of LiDAR BEV. We provide a new cross-attention method at the heart of BEV feature fusion that is specifically intended for blending camera and LiDAR BEV features. This mechanism is enhanced by a BEV self-attention mechanism, which improves feature communication and integration in several dimensions. Through extensive experiments on the nuScenes dataset, we demonstrate the effectiveness of our technique by showing that CL-FusionBEV performs remarkably well across a range of assessment parameters. Although these findings are encouraging, we recognize that for wider use, our model needs more optimization. This entails lowering processing requirements and evaluating its applicability in a larger range of environmental circumstances. In order to improve CLFusionBEV's functionality and performance, we are dedicated to addressing these issues in subsequent studies.

Reference

[1] C. Yan and E. Salman, "Mono3D: Open Source Cell Library for Monolithic 3-D Integrated Circuits," 2017.
[2] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection," 2017.
[3] C. Pham and J. W. Jeon, "Robust Object Proposals Re-ranking for Object Detection in Autonomous Driving Using Convolutional Neural Networks," 2017.
[4] B. Xu and Z. Chen, "Multi-Level Fusion Based 3D Object Detection from Monocular Images," 2018.
[5] H. Dou, Y. Liu, S. Chen, and H. Bilal, "A Hybrid CEEMD-GMM Scheme for Enhancing the Detection of Traffic Flow on Highways," 2023.
[6] Y. Zhou and O. Tuzel, "VoxelNet: End-to-End Learning for Point Cloud-Based 3D Object Detection," 2018.
[7] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely Embedded Convolutional Detection," 2018.
[8] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-Based 3D Object Detection and Tracking," 2021.
[9] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From Points to Parts: 3D Object Detection from Point Cloud with Part-Aware and Part-Aggregation Network," 2020.
[10] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast Encoders for Object Detection from Point Clouds," 2019.
[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," 2017.
[12] C. Zhang, X. Pan, and H. Li, "A Hybrid MLP-CNN Classifier for Very Fine Resolution Remotely Sensed Image Classification," 2018.
[13] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," 2017.
[14] H. Bilal, W. Yao, Y. Guo, Y. Wu, and J. Guo, "Experimental Validation of Fuzzy PID Control of Flexible Joint System in Presence of Uncertainties," 2017.
[15] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," 2019.
[16] Z. Liu, H. Tang, A. Amini, X. Liu, X. Yu, S. Han, and D. Rus, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," 2023.

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025

[17] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," 2017.
[18] Q. Wu, X. Li, K. Wang, and H. Bilal, "Regional Feature Fusion for On-Road Detection of Objects Using Camera and 3D-LiDAR in High-Speed Autonomous Vehicles," 2023.
[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A Multimodal Dataset for Autonomous Driving," 2020.
[20] H. Yan, X. Yu, Y. Zhang, S. Zhang, X. Zhao, and L. Zhang, "Single Image Depth Estimation with Normal Guided Scale Invariant Deep Convolutional Fields," 2017.
[21] H. Bilal, B. Yin, M. S. Aslam, and H. Wu, "A Practical Study of Active Disturbance Rejection Control for Rotary Flexible Joint Robot Manipulator," 2023.

DOI: https://doi.org/10.54216/FPA.190202
Received: January 03, 2025 Revised: February 13, 2025 Accepted: March 05, 2025