Article

Not peer-reviewed version

BiDFNet: A Bidirectional Feature Fusion Network for 3D Object Detection Based on Pseudo-LiDAR

Qiang Zhu and Yaping Wan *

Posted Date: 19 May 2025

doi: 10.20944/preprints202505.1480.v1

Keywords: Pseudo-point cloud; 3d object detection; Substream sparse convolution; Multi-head attention mechanism

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

BiDFNet: A Bidirectional Feature Fusion Network for 3D Object Detection Based on Pseudo-LiDAR

Qiang Zhu and Yaping Wan *

School of Computer Science, University of South China, Hengyang 421001, China
* Correspondence: ypwan@usc.edu.cn

Abstract: This paper presents a Bidirectional Feature Fusion Network (BiDFNet) for 3D object detection, leveraging pseudo-point clouds to achieve bidirectional fusion of point cloud and image features. The proposed model addresses key challenges in multimodal 3D detection by introducing three novel components: (1) the SAF-Conv module, which extends the receptive field through improved submanifold sparse convolution, enhancing feature extraction from pseudo-point clouds while effectively reducing edge noise; (2) the Bidirectional Cross-modal Attention Feature Interaction Module (BiCSAFIM), which employs a multi-head cross-attention mechanism to enable global information interaction between point cloud and image features; and (3) the Attention-based Feature Fusion Module (ADFM), which adaptively fuses dual-stream features to improve robustness. Extensive experiments on the KITTI dataset demonstrate that BiDFNet achieves state-of-the-art performance, with a 3D AP (R40) of 88.79% on the validation set and 85.27% on the test set for the Car category, significantly outperforming existing methods. These results highlight the effectiveness of BiDFNet in complex scenarios, showcasing its potential for real-world applications such as autonomous driving.

Keywords: pseudo-point cloud; 3D object detection; substream sparse convolution; multi-head attention mechanism

1. Introduction

In the advancement of industrial domains such as autonomous driving, robotics, and intelligent traffic monitoring, 3D object detection technology plays a pivotal role. It directly impacts the environmental perception capabilities and decision-making accuracy of emerging transportation tools, including smart vehicles and drones. For instance, in autonomous driving, accurate 3D object detection is crucial for obstacle avoidance, path planning, and ensuring passenger safety. Despite significant progress in recent years, existing methods still face challenges in complex and dynamic environments, such as varying lighting conditions, occlusions, and sparse data from distant objects.

Although significant progress has been achieved in recent years with single-sensor-based 3D object detection including both grid-based LiDAR methods like VoxelNet [1], SECOND [2], PointPillars [3], Voxel R-CNN [4], and VoxelNeXt [5], as well as point-based approaches such as PointRCNN [6], Point-GNN [7], and hybrid methods like PV-RCNN [8], these methods still encounter limitations in complex and dynamic traffic environments. For example, LiDAR provides precise three-dimensional spatial information but is susceptible to extreme weather conditions, such as rain or snow, which can scatter laser signals and reduce data quality. On the other hand, cameras capture rich color and texture details but struggle with accurate depth estimation, especially in low-light conditions or when objects are far away. These limitations highlight the need for multimodal approaches that can combine the strengths of different sensors to improve detection performance.

Consequently, multimodal 3D object detection, which integrates data from diverse sensors such as LiDAR and cameras, has emerged as an effective strategy to enhance detection performance. Early approaches like MV3D [9] explored fusing multiple views, while methods like Frustum PointNets [10] leveraged 2D detections to constrain the 3D search space. More recent works focus on fusing features

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

at various stages, such as combining raw point and image features [11], enhancing point features with image semantics [12], fusing features at the BEV level [13], or employing deep fusion networks [14]. By combining the precise geometric information from LiDAR with the rich semantic details from cameras, multimodal methods can achieve more robust and accurate detection in complex environments. For example, in urban driving scenarios, multimodal approaches can better handle occlusions and varying lighting conditions, leading to improved safety and reliability. Recent studies have demonstrated that multimodal fusion can significantly enhance detection accuracy, particularly for distant or partially occluded objects.

However, existing multimodal methods still face several challenges. For instance, traditional feature fusion techniques fail to fully exploit the complementary information between modalities, resulting in suboptimal detection accuracy. These issues are particularly pronounced in complex scenarios where objects are partially occluded or located at a distance. Furthermore, methods relying on derived representations like pseudo-point clouds (discussed in Section 2) introduce specific challenges related to noise and feature quality. To address these challenges, this paper proposes SAF-Conv module and BiCSAFIM model. The SAF-Conv module uses more precise 3D regions of point clouds for cropping and employs improved submanifold sparse convolution [15] to expand the receptive field for extracting richer information, effectively reducing edge noise. The BiCSAFIM module uses a multi-head cross-attention mechanism [16] to significantly enhance the information interaction between point cloud features and image features from a global perspective, effectively improving the model's utilization of point cloud and image feature information.

To evaluate the performance of the proposed model, the widely recognized KITTI dataset, a benchmark for multimodal 3D object detection, was employed for network performance assessment. Comprehensive experimental results indicate that the model achieves notable performance improvements compared to traditional feature fusion networks. Furthermore, ablation experiments validate the significance and necessity of each module introduced in this study. The primary contributions of this paper are as follows:

- We propose a bidirectional cross-sensor attention feature interaction module (BiCSAFIM), which significantly enhances the interaction between point cloud features and image features.
- A feature extraction module, termed SAF-Conv, is introduced for pseudo-point clouds. This module enlarges the receptive field, thereby improving the extraction of 2D features and concurrently reducing noise to a certain extent.
- Our method surpasses previous approaches, achieving state-of-the-art performance on the KITTI 3D object detection benchmark.

2. Related Work

LiDAR-Based 3D Object Detection

LiDAR provides precise geometric information and is a focal point in single-modal 3D detection. Early methods project point clouds into 2D bird's-eye view (BEV) or range images for 3D detection [3, 17,18]. Recently, mainstream LiDAR-based 3D object detection approaches can be divided into two major categories: point-based methods and grid-based methods. Point-based methods, such as PointNet [19] and PointNet++ [20], directly process raw point clouds using deep learning frameworks. These methods extract features through point-based backbone networks and predict 3D bounding boxes based on downsampled points and features. However, the process of sampling and aggregating features from irregular point clouds is computationally expensive.

Grid-based methods, meanwhile, transform point clouds into structured grid formats, such as voxels, pillars, or bird's-eye view (BEV) feature maps. For example, VoxelNet [1] utilizes sparse voxel grids to extract features from points within voxel cells. SECOND [2] introduced 3D sparse convolution for voxels, significantly accelerating the convolution process. PointPillars [3] further simplifies voxels into pillars to enhance processing speed. Despite their efficiency, grid-based methods often struggle with distant or sparse objects due to the limited resolution of LiDAR. Addressing computational

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

efficiency, Ada3D [21] focuses on optimizing inference speed by exploiting spatial redundancy. It employs an adaptive mechanism to dynamically allocate computation, reducing processing in empty or simple regions of the 3D space, thereby improving overall efficiency.

Fusion-Based 3D Object Detection

To address the limitations of single-modal approaches, multimodal 3D object detection methods have been developed to leverage both LiDAR and camera data. These methods can be broadly categorized into frustum-based and multi-view-based approaches. frustum-based methods, represents a sequential, result-level fusion approach. Works like FrustumPointNet [10], FrustumConvNet [22], and CenterFusion [23] use the output of image detectors to lift 2D detections into 3D frustums with explicit depth estimation. While these methods optimize the search space for 3D detection and reduce computational costs, their performance is often limited by inaccurate depth estimation. For example, FrustumPointNet generates 3D proposals by extruding 2D bounding boxes into 3D space, but the accuracy of these proposals heavily depends on the quality of depth estimation, which can be unreliable in complex scenes.

Multi-view-based approaches utilize feature-level fusion to create 3D region proposals from bird's-eye views (BEV) and conduct regression for 3D bounding boxes. A notable early contribution in this field is MV3D [9], developed by Chen. MV3D converts point clouds into BEV representations, extracts relevant features, and generates 3D proposals. These proposals are subsequently projected into range views, where they are combined with high-dimensional image features during the ROI pooling phase, followed by 3D bounding box regression. Building on MV3D, Ku introduced AVOD [24], which integrates image feature maps from VGG-16 with point cloud data during the 3D proposal generation stage. AVOD also incorporates an autoencoder (AE) structure to reduce feature space dimensionality, thereby lowering computational demands. Despite these advancements, such fusion techniques often project 3D point clouds onto 2D planes, leading to substantial information loss. Additionally, the limited alignment between point cloud features and image features hinders the full utilization of semantic information.

Pseudo-Point Cloud Methods

A significant direction within multimodal fusion leverages pseudo-point clouds, which are 3D representations generated primarily from camera data using depth estimation, rather than direct LiDAR sensing. This approach has gained traction due to several potential advantages: (1) Cost-Effectiveness: Enabling capable 3D perception using lower-cost camera-centric sensor suites. (2) Leveraging 2D Representations: Allowing the application of powerful, pre-trained image networks to extract rich semantic features for 3D tasks. (3) Unified Geometric Representation: By converting image-derived depth information into a 3D point cloud format, pseudo-LiDAR brings both camera and (potentially) sparse real LiDAR data into a common geometric domain (point clouds). This unification can simplify the subsequent feature interaction process compared to directly fusing disparate representations like dense 2D feature maps and sparse 3D points, potentially reducing the complexity of explicit cross-dimensional feature alignment mechanisms. (4) Scene Densification: Potentially filling gaps in sparse LiDAR scans when fused, offering a denser scene representation.

Building upon these advantages, recent research has explored various techniques. For instance, MVP [25] uses depth estimation to generate virtual points to enhance the point cloud. PointAugmenting [26] introduces cross-modal data augmentation techniques to generate diverse pseudo-point clouds, enhancing model generalization. SFD [27] employs semantic feature distillation to improve the expression of geometric details and semantic information in pseudo-point clouds. patch refinement proposes a local patch-based optimization method for pseudo-point clouds, reducing the impact of depth estimation errors. DepthGAN [28] leverages generative adversarial networks to produce high-quality pseudo-point clouds, improving the accuracy and robustness of depth estimation. However, existing pseudo-point cloud methods face two major challenges: (1) inaccurate depth estimation often introduces incorrectly placed virtual points around objects, misleading object predictions; and (2) the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

large number of redundant points generated in pseudo-point clouds do not fully align with sparse point cloud data, leading to incomplete utilization of information during fusion.Recent work also explores how to better utilize pseudo-points within detector architectures. For instance, SQD [29] proposes enhancing features around sparse object queries using pseudo-point information, aiming to improve the interaction between sparse queries and the densified feature maps for more accurate detection. These limitations motivate the need for improved feature extraction and fusion techniques, as proposed in this work.

In essence, challenges remain in effectively handling sparse LiDAR data, mitigating noise in pseudo-LiDAR representations, and achieving truly synergistic fusion between modalities. This work introduces BiDFNet to address these issues, proposing dedicated modules for robust pseudo-LiDAR feature extraction (SAF-Conv), deep bidirectional cross-modal interaction (BiCSAFIM), and adaptive fusion (ADFM), detailed next.

3. BiDFNet

This section offers an in-depth exploration of the architectural framework of BiDFNet, a 3D object detection approach leveraging pseudo-point clouds, which combines submanifold sparse convolution with multi-head attention mechanisms. Section 3.1 outlines the overarching design principles of the architecture. Following this, Section 3.2 delves into the design and implementation of the sparse convolutional submanifold network (SAF-Conv) [15]. Section 3.3 details the structure of the bidirectional feature interaction module (BiCSAFIM). Finally, in Section 3.4, the grid feature fusion (ADFM) module is discussed.

3.1. Overall Architecture

In the BiDFNet model, pseudo-point clouds are first generated through depth completion. The initial point cloud is used to predict 3D RoIs (Regions of Interest). These RoIs are then employed to crop both the original point cloud and the pseudo-point cloud. For the pseudo-point cloud, point cloud information is extracted and voxelized. The voxelized features are fed into SAF-Conv for further feature extraction and a certain degree of denoising. Finally, the features from the initial point cloud and the pseudo-point cloud are input into the dual-stream feature interaction module. The interaction-enhanced features are fused by the ADFM (Attention-based Feature Fusion Module) and then passed to the detection head (Detect Head) to produce the output. The overall architecture of the BiDFNet model is illustrated in Figure 1.

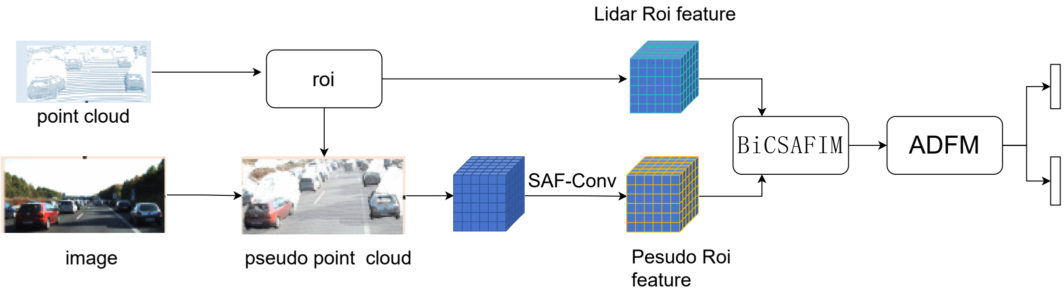

Figure 1. 1. The initial point cloud is used to predict 3D RoIs (Regions of Interest). These RoIs are then employed to crop both the original point cloud and the pseudo-point cloud.For the pseudo-point cloud, point cloud information is extracted and voxelized. 2. The voxelized features are fed into SAF-Conv for further feature extraction and a certain degree of denoising. 3. the features from the initial point cloud and the pseudo-point cloud are input into the dual-stream feature interaction module. 4. The interaction-enhanced features are fused by the ADFM (Attention-based Feature Fusion Module) and then passed to the detection head (Detect Head) to produce the output.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

3.2. *Pseudo-Point Cloud Feature Extraction*

For the original point cloud R, we employ the benchmark Voxel-RCNN [4] method to generate 3D detection regions.Similarly, the region proposals are further divided into G×G×G regular sub-voxels, with the center point serving as the grid point of the corresponding sub-voxel. We observe that a significant amount of non-Gaussian noise is introduced during the generation of pseudo-point clouds due to depth estimation errors. Notably, this noise tends to occur prominently at object edges. While edge detection can roughly remove such noise, it also results in the loss of a substantial number of useful boundary points. To address this, we crop the pseudo-point cloud using 3D proposal regions generated from point clouds rather than pseudo-point clouds. To extract 2D semantic information from the pseudo-point cloud, we utilize CPConvs [27]. Following this, the extracted features are voxelized and then processed for feature extraction . When extracting features from the pseudo-point cloud, we develop an improved submanifold sparse convolution (SAF-Conv), inspired by [30]. This approach expands the receptive field in the pseudo-point cloud flow to the 2D image space, enabling the simultaneous extraction of both 3D and 2D features from the pseudo-point cloud.The structure of this method is illustrated in Figure 2. It is designed specifically to handle the characteristics of pseudo-LiDAR. It expands the receptive field within the sparse convolution framework, allowing the model to incorporate broader spatial context, which is crucial for interpreting potentially noisy points, especially near object boundaries. Furthermore, SAF-Conv enables the simultaneous extraction and adaptive fusion of both 3D geometric features and projected 2D semantic features derived from the pseudo-point cloud, aiming for a more robust representation.

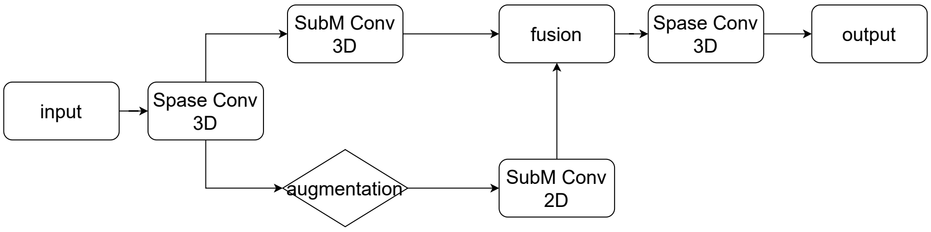

Figure 2. The structure of SAF-Conv. Input voxel features are processed through two parallel streams: (Top) 3D Submanifold Sparse Convolution captures local geometric structure. (Bottom) Features are projected onto the 2D image plane), processed by 2D Submanifold Sparse Convolution to capture semantic context, and then fused back with the 3D features using an attention-based mechanism . This allows adaptive integration of 3D geometry and 2D semantics from pseudo-LiDAR data

Specifically, given N input voxels represented by 3D index vectors $R^{N\times 3}$ and feature vectors $Z \in R^{N\times C^{in}}$ geometric features are first encoded separately in both 3D and 2D image spaces. These two geometric features are then combined to yield $Y \in R^{N\times C^{out}}$ where the number of input and output feature channels are denoted as C^{in} and C^{out}

As depicted in the workflow of Figure 2, SAF-Conv processes features through parallel 2D and 3D spatial encoding paths before fusion. 3D Space Encoding: The geometric features are directly encoded using the 3D submanifold convolution kernel K3D(·).$\hat{Z}_i$are calculated from the non-empty voxels within 3 × 3 × 3 neighborhood based on the corresponding 3D indices .it is expressed as Equation (1) [30].

$$\hat{Z}_i = R\left\{ K^{3D}\left(Z_i, Z_i^{(f_1)}, \ldots, Z_i^{(f_j)}\right)\right\} \tag{1}$$

2D Space Encoding:The points are first projected onto the 2D image plane, after which geometric features are encoded using the 2D submanifold convolution kernel K2D(·)The 3D indices are converted into a set of grid points (denoted as G(·)) based on voxelization parameters. Due to the inherent differences between point cloud and image data, data augmentation techniques such as rotation and scaling are commonly applied, as denoted byT(·) [27,30].The grid points are then mapped back to the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

original coordinate system using the data augmentation parameters. Finally, these points are projected onto the 2D image plane, denoted as P(·) using LiDAR-Camera calibration parameters. This process is mathematically expressed as Equation (2) [30].

$$\hat{H} = P\left(T^{-1}(G(H))\right) \tag{2}$$

Here, $\hat{H}$ represents the 2D index vector.For each voxel feature Z_i ,then calculate the noise-aware features from the non-empty voxels within a 3×3 neighborhood based on the corresponding 2D indices.it is expressed as Equation (3) [30].

$$\tilde{Z}_i = R\left\{ K^{2D}\left(Z_i, \tilde{Z}_i^{(f_1)}, \ldots, \tilde{Z}_i^{(f_k)}\right)\right\} \tag{3}$$

$\tilde{Z}_i$ represents the adjacent voxel features generated, and K2D(·) denotes the 2D submanifold convolution kernel. Given that multiple features may exist within a single 2D adjacent voxel, we employ max pooling to retain the most prominent feature for performing 2D convolution, ensuring the emphasis on significant characteristics. After encoding both 3D and 2D features, an attention mechanism is applied to combine these two features, ultimately yielding the ROI (Region of Interest) features of the pseudo-point cloud.it is expressed as Equation (4-6)

$$\alpha = \sigma(\mathrm{MLP}([\tilde{Z}_i; \hat{Z}_i])) \tag{4}$$

$$\overline{Z_i} = \alpha \cdot \tilde{Z}_i + (1 - \alpha) \cdot \hat{Z}_i \tag{5}$$

$$\mathrm{Y} = \left[\overline{Z_1}^{\mathrm{T}}, \ldots, \overline{Z_N}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{6}$$

Multiple SAF-Conv modules are linked in sequence with standard SparseConv3D layers to broaden the 2D receptive field during sparse convolution, resulting in the final output feature vector.

3.3. *Interaction of Bidirectional Features*

Previous methods relied primarily on ROI feature strategies to utilize multimodal data or simply employed channel attention modules to fuse two ROIs. However, these approaches did not fully exploit the complementary information between modalities, resulting in insufficiently effective feature fusion.

To address these limitations, we propose BiCSAFIM (Bidirectional Cross-Modal Selective Attention Feature Interaction Module), as illustrated in Figure 3. This module consists of dual streams, each taking LiDAR features and pseudo-point cloud features generated from RGB images as input, and performs feature interaction and enhancement separately.

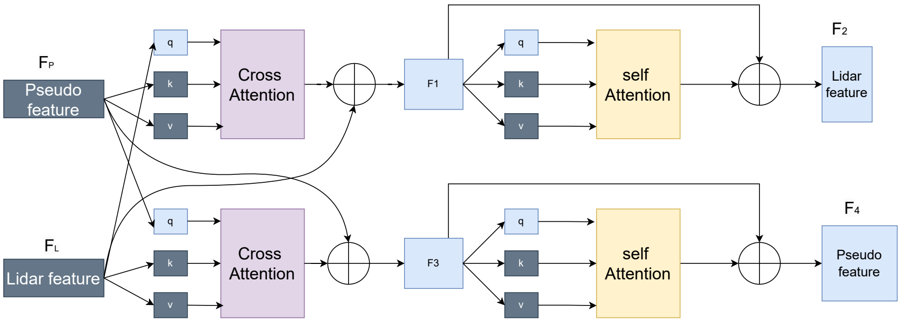

Figure 3. The structure of BiCSAFIM,structured with two branches, both of which take point cloud features and pseudo-point cloud features as inputs. After interaction through multi-head cross-attention, the features are further enhanced via multi-head self-attention.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

Simple fusion methods like concatenation or element-wise operations, or even basic channel attention, often fail to capture the complex, non-linear correlations and spatial misalignments between sparse LiDAR features and denser pseudo-LiDAR/image features. We employ a bidirectional cross-attention mechanism (BiCSAFIM) based on the Transformer architecture [16] because cross-attention explicitly allows features from one modality (acting as queries) to attend to and selectively aggregate information from another modality (acting as keys and values) based on learned relevance. The bidirectional structure, with parallel streams where each modality can query the other, ensures comprehensive information exchange (e.g., pseudo-LiDAR providing denser context where LiDAR is sparse, and real LiDAR providing accurate geometric cues to refine pseudo-LiDAR features). Multi-head attention further enhances this capability by allowing the model to jointly attend to information from different representation subspaces at different positions.

Given the need for cross-modal interaction and the requirement for features from one modality (e.g., pseudo-point ROI features) to perceive the overall features of another modality (e.g., point ROI features) while selectively aggregating and enhancing complementary information, we incorporate both multi-head cross-attention and multi-head self-attention mechanisms [16]. This enables the dual streams to perceive intermodal feature relationships while emphasizing their own feature information.

More specifically, initially, for LiDAR ROI features, F_l serve as the query, while pseudo-LiDAR ROI features F_p act as the key and value for intermodal fusion, resulting in feature F1.

$$\mathrm{Attention}(F_l, F_p) = \mathrm{Softmax}\left(\frac{F_l F_p^{\mathrm{T}}}{\sqrt{d_k}}\right) F_p \tag{7}$$

We begin by applying the multihead cross-attention mechanism, where independent learnable linear transformations are applied to the query FL, key FP, and value FP. The i-th head is denoted as in Equation (8), where dimensions are represented as follows.

$$Q_i = F_l W_i^Q, \quad K_i = F_p W_i^K, \quad V_i = F_p W_i^V \tag{8}$$

We then employ m heads to execute multi-head cross-attention, and the result is summed with FL. as shown in Equations (9)–(10).

$$\mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i) \tag{9}$$

$$F_1 = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, ..., \mathrm{head}_h) + F_l \tag{10}$$

where F_1 is the sum of the output from the multi-head attention module and the LiDAR ROI featuresF_L. F_1 is used as queries, keys, and values Subsequently, similar to the multi-head cross-attention mechanism, we utilize m heads to perform multi-head self-attention, aiming to emphasize important information within a single modality, thereby obtaining the intra-modal fusion feature F_2.as shown in Equations (11)–(14)

$$\mathrm{Attention}(F_1, F_1) = \mathrm{Softmax}\left(\frac{F_1 F_1^{\mathrm{T}}}{\sqrt{d_k}}\right) F_1 \tag{11}$$

$$Q_i = F_1 W_i^Q, \quad K_i = F_1 W_i^K, \quad V_i = F_1 W_i^V \tag{12}$$

$$\mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i) \tag{13}$$

$$F_2 = \mathrm{MultiHead}(F_1) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, ..., \mathrm{head}_h) \tag{14}$$

After applying the aforementioned attention mechanisms, we obtain the fused feature F_2 from the point cloud stream branch, which achieves the fusion of pseudo-point cloud features and the enhancement of its own features. The other branch adopts a similar method. After selective fusion of pseudo-point cloud features with point cloud features, it enhances itself to get F_4.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

3.4. Attention-Driven RoI Fusion

After processing the features from the original LiDAR stream (yielding F_2) and the pseudo-LiDAR stream (yielding F_4) through the BiCSAFIM, a final fusion step is needed before prediction. Since the reliability and informativeness of each feature stream (F_2 vs. F_4) might vary spatially across the RoI grid (e.g., pseudo-LiDAR might be more reliable in dense areas far from the ego vehicle, while real LiDAR is better close-up or for fine geometric details), a simple averaging or concatenation might be suboptimal.

To address this, we employ the Attention-Driven Fusion Module (ADFM), inspired by attention mechanisms [16] and selective fusion approaches [31,32]. Since the features F_2 and F_4 represent the same RoI and share a consistent grid structure, ADFM can operate effectively at the grid level. As illustrated in Figure 4, for each corresponding pair of grid features from F_2 and F_4, ADFM dynamically computes spatial attention weights. This allows the network to learn which feature stream to emphasize more at each specific location within the RoI. By adaptively weighting the grid features based on their learned reliability or importance before combining them, ADFM produces a more robust final fused representation for the detection head.

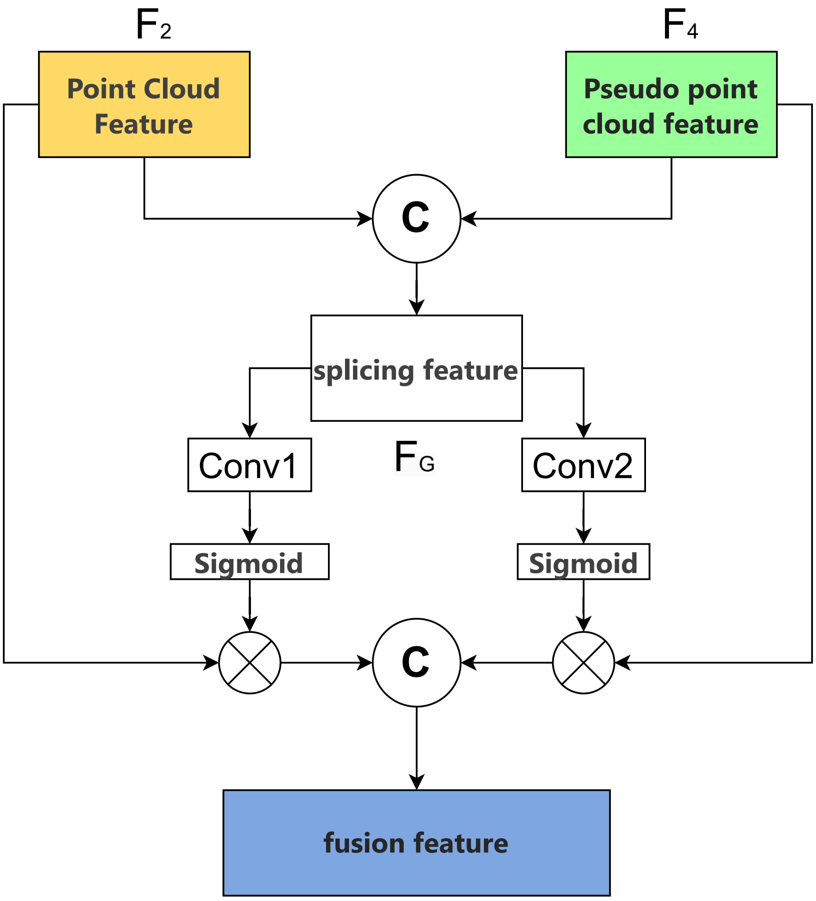

Figure 4. The structure of ADFM.The Attention-Driven Fusion Module (ADFM). Parallel branches generate attention weights (w_1, w_2) for input features F_2(LiDAR) and F_4 (pseudo-LiDAR) using Conv + Signem. Weighted features are concatenated (C) for adaptive fusion before detection.

Specifically, let b represent a single 3D RoI. In the previous section, we obtained the original flow RoI features F_2 and the pseudo flow RoI features F_4. Here, n denotes the total number of grids in the 3D RoI, and C represents the number of feature channels per grid. Given a pair of RoI grid features ($(F_2, F_4$), they are concatenated. Subsequently, two parallel 3×3 convolutional layers are applied to generate higher-level features. These features are then passed through a sigmoid function to produce a

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

pair of weights (w_1, w_2), where both w_1 and w_2 are scalars. Finally, the fused grid feature F is obtained by weighting (F_2, F_4) with (w_1, w_2). The implementation of (F) is as follows

$$F_G = \mathrm{Concat}(\mathrm{F}_2,\ \mathrm{F}_4) \tag{15}$$

$$w_1 = \mathrm{sigmoid}(\mathrm{Conv}(\mathrm{F}_2)), w_2 = \mathrm{sigmoid}(\mathrm{Conv}(\mathrm{F}_4)) \tag{16}$$

$$F = \mathrm{CONCAT}(\mathrm{F}_2 \times w_1,\ \mathrm{F}_4 \times w_2) \tag{17}$$

By fusing the RoI features of the original flow and the pseudo-point cloud flow using the aforementioned method, the final fused RoI features are then utilized to predict class confidence scores and bounding boxes.

4. Experiments

In this section, we present a comprehensive evaluation of the BiDFNet model on the KITTI dataset [33], a widely recognized benchmark for 3D object detection. The evaluation is structured as follows: Section 4.1 provides an overview of the KITTI dataset and its evaluation metrics. Section 4.2 details the experimental setup, including the network architecture, loss function, data augmentation techniques, and training procedures. Section 4.3 discusses the design and implementation of the proposed model. Section 4.4 presents ablation studies to analyze the contributions of individual modules. Finally, Section 4.5 summarizes the experimental results and highlights the key findings.

4.1. KITTI Dataset

The KITTI dataset [33] is utilized for our experiments. It provides real-world driving data including LiDAR point clouds and camera images, along with annotations for 3D object detection. We conduct our experiments on this dataset as it remains a widely adopted and challenging benchmark for 3D object detection in autonomous driving research. Its real-world driving scenarios and standardized evaluation protocol facilitate fair comparison with numerous state-of-the-art methods [4,12,27,29,30, 34–39]. While larger datasets like nuScenes or the Waymo Open Dataset offer greater diversity in environments and sensor configurations, KITTI is still highly relevant for evaluating core aspects of multimodal fusion and performance on varying object distances and occlusion levels, which are central to our proposed method. For evaluation, we follow the standard practice of splitting the official training set of 7,481 samples into a training set (3,712 samples) and a validation set (3,769 samples). Performance is reported using the official Average Precision (AP) metric with IoU thresholds of 0.7 for the 'Car' category, calculated across Easy, Moderate, and Hard difficulty levels.

4.2. Experimental Details

The development environment used for the experiments includes: (1) Ubuntu 18.04 operating system; (2) GPU: RTX 4090; (3) open-source framework: PyTorch 1.6.9. The BiDFNet model restricts the point cloud range to [0, 70.4] m on the X-axis, [-40, 40] m on the Y-axis, and [-3, 1] m on the Z-axis. The input voxel size is set to (0.05 m, 0.05 m, 0.1 m), with an IoU threshold of 0.7 and a recall rate of 40. The code for this work is implemented based on an open-source 3D object detection framework, Voxel-RCNN.

Loss and Data Augmentation

In the BiDFNet model, we use the same training loss as the benchmark method [4]. Local and global data enhancement methods such as ground-truth sampling, rotation, translation and flip are followed in the model [27,30].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

Training and Inference Details

The BiDFNet model undergoes training from the ground up on the KITTI dataset, utilizing an adaptive optimization algorithm across 80 epochs. Training is performed on a single NVIDIA 4090 GPU, with learning rates dynamically adjusted to 0.01, 0.01, and 0.005 at various phases of the training process. To ensure an equitable comparison with existing models, all experimental settings—excluding the learning rate—are aligned with the baseline framework outlined in Voxel-RCNN [4].

4.3. Main Results

We evaluate BiDFNet on the widely-used KITTI dataset [33] following the standard evaluation protocol, reporting 3D Average Precision (AP) with an IoU threshold of 0.7 for the 'Car' category across Easy, Moderate, and Hard difficulty levels, using both R40 and R11 recall sampling points. For comparison against existing methods in Tables 1 and 2, results for competing approaches are primarily sourced from their original publications or the official KITTI leaderboard, unless otherwise noted, ensuring a fair comparison under identical evaluation settings

Table 1. Comparison of experimental results under the KITTI validation set.

Method	Reference	Car AP (R40)			Car AP (R11)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
Voxel-RCNN [4]	AAAI 2021	92.38	85.26	82.26	89.41	84.52	78.93
Pyramid-PV [34]	ICCV 2021	92.26	85.23	82.94	89.37	84.38	78.84
D-Conformer-TSD [35]	ICASSP 2023	-	-	-	89.69	85.36	79.53
3D-CVF [37]	ECCV 2020	89.67	79.88	78.47	-	-	-
EPNet [12]	ECCV 2020	92.28	82.59	80.14	-	-	-
SFD [27]	CVPR 2022	93.21	87.85	84.56	88.96	86.32	84.79
VirConv-L [30]	CVPR 2023	93.36	88.71	85.83	-	-	-
Ours	-	93.47	88.79	86.41	89.71	86.81	85.70

Results on KITTI validation set. From the experimental results of car detection accuracy comparison in Table 1, it can be observed that the BiDFNet model, by integrating the SAF-Conv, BiCSAFIM, and ADFM modules, significantly enhances its feature representation capability in complex environments. Specifically, compared to the baseline model Voxel-RCNN, BiDFNet shows improvements across Easy, Moderate, and Hard detection levels: the 3D AP (40) increased by 1.09%, 3.53%, and 4.15%, respectively, while the 3D AP (11) increased by 0.30%, 2.29%, and 6.77%, respectively. Furthermore, compared to some mainstream 3D vehicle detection algorithms, the proposed method achieves the best results across all three detection difficulty levels.

Table 2. Comparison of experimental results on the KITTI test set.

Method	Reference	Car AP (R40)		
		Easy	Mod.	Hard
Voxel-RCNN [4]	AAAI 2021	90.90	81.62	77.06
Pyramid-PV [34]	ICCV 2021	88.39	82.08	77.49
CasA [38]	TGRS 2022	91.58	83.06	80.08
D-Conformer-TSD [35]	ICASSP 2023	89.13	82.18	79.75
PV-RCNN++ [36]	IJCV 2023	90.53	81.60	77.07
3D-CVF [37]	ECCV 2020	89.20	80.85	73.11
Graph-Vol [39]	ECCV 2022	91.89	83.27	77.78
SFD [27]	CVPR 2022	91.73	84.76	77.92
VirConv-L [30]	CVPR 2023	91.41	85.05	80.56
Ada3D [21]	ICCV2023	87.46	79.41	75.63
SQD [29]	ACM MM 2024	91.58	81.82	79.07
Ours	-	**91.79**	**85.27**	**80.72**

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

The results on the KITTI test set demonstrate that our BiDFNet model achieved significant improvements in 3D AP (R40) across the easy, moderate, and hard vehicle categories, with increases of 0.89%, 3.65%, and 3.66%, respectively, compared to the baseline model Voxel RCNN [4]. Moreover, BiDFNet has consistently outperformed numerous state-of-the-art methods in recent years. For a comprehensive comparison, the detailed results are summarized in Table 2. While improvements on the 'Easy' subset might appear incremental due to near-saturation by top methods, the significant gains observed for the 'Moderate' and particularly the 'Hard' difficulty levels are noteworthy. These Hard cases often involve distant, small, or heavily occluded objects, where robust feature extraction and effective fusion are most critical. The consistent improvements in these challenging scenarios strongly suggest the effectiveness of our proposed modules: SAF-Conv likely contributes by enhancing pseudo-point cloud features and mitigating noise, BiCSAFIM improves cross-modal feature interaction, and ADFM enables more robust final fusion. This ability to better handle difficult cases highlights the practical value of BiDFNet for real-world autonomous driving applications where robustness is paramount.

To validate the generalization capability of our approach, we extended BiDFNet to simultaneously detect pedestrians, and cyclists using a unified model. As shown in Table 3, our method demonstrates consistent improvements across all object categories compared to the multi-class Voxel-RCNN baseline. These results confirm that BiDFNet's bidirectional fusion mechanism effectively preserves its advantages when adapted to multi-class detection scenarios, achieving significant performance gains while maintaining architectural simplicity.

Table 3. 3D Detection results (3D AP (R40)) of multi-class(KITTI validation set).

Class	Method	3D AP (R40)		
		Easy	Mod.	Hard
Pedestrian	Baseline	70.55	62.92	57.35
Pedestrian	Ours	73.47	66.82	61.01
Cyclist	Baseline	89.86	71.13	66.67
Cyclist	Ours	90.13	73.87	69.15

4.4. Ablation Study

To assess the efficacy of the proposed approach, a comprehensive set of experiments was carried out. As depicted in Table 4, experiment (a) serves as a tailored baseline model derived from the Voxel R-CNN framework, utilizing solely the original point cloud data as input. Conversely, experiments (b) through (i) integrate pseudo-point clouds to enable multimodal enhancement, facilitating a balanced comparison with the single-modal approach of experiment (a). These experiments aimed to evaluate the synergistic effects and overall performance of the SAF-Conv, BiSCAFIM, and ADFM modules.

Table 4. Performance Impact of Various Components on Car 3D Detection Using the KITTI Validation Set, Measured by Average Precision at 40 Recall Positions

Experiment	pseudo	SAF-Conv	BiCSAFIM	ADFM	Car AP (R40)			times(ms)
					Easy	Mod	Hard	
(a)					92.38	85.26	82.26	40
(b)	✓				92.41	85.63	83.97	62
(c)	✓	✓			92.81	86.87	85.09	70
(d)	✓		✓		92.85	86.95	85.13	72
(e)	✓			✓	92.49	85.91	83.04	64
(f)	✓	✓	✓		93.20	88.34	86.10	80
(g)	✓		✓	✓	92.97	87.63	85.41	76
(h)	✓	✓		✓	93.02	87.54	85.51	74
(i)	✓	✓	✓	✓	93.47	88.79	86.41	84

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025

doi:10.20944/preprints202505.1480.v1

The integration of SAF-Conv and BiCSAFIM into the baseline model led to a notable enhancement in detection accuracy, as evidenced by the results in Table 4. Specifically, SAF-Conv contributed to improvements in 3D AP (R40) by 0.40%, 1.24%, and 1.12% for the easy, moderate, and hard difficulty levels, respectively. Similarly, BiCSAFIM achieved gains of 0.44%, 1.32%, and 1.16% across the same categories. When both modules were combined, the accuracy saw further enhancement, with the hard class accuracy reaching 86.10%. The addition of the ADFM fusion attention module further elevated the hard class detection accuracy to 86.41%, representing a 4.15% overall improvement compared to the baseline. These results underscore the significant contributions of each module to the model's performance.

Finally, we analyze the computational overhead introduced by our modules using the inference times reported in Table 4 (measured in ms on an RTX 4090). The Voxel-RCNN baseline (Experiment (a)) runs in 40 ms. Integrating the pseudo-LiDAR stream and our proposed components increases the runtime incrementally: SAF-Conv adds approximately 8 ms, BiCSAFIM adds about 10 ms, and ADFM adds 4 ms. Our full BiDFNet model (Experiment (i)) results in a total inference time of 84 ms. While this is roughly 2.1 times slower than the baseline (corresponding to 11.9 FPS), this increased computational cost directly correlates with significant performance gains, particularly the substantial 4.15% AP (R40) improvement on the difficult 'Hard' category. This demonstrates a clear trade-off, where the enhanced robustness and accuracy in challenging scenarios justify the additional computation, though further optimization could be explored for applications demanding higher frame rates.

The experimental results reveal that the SAF-Conv module significantly improves the feature representation of pseudo-point clouds, whereas the BiCSAFIM and ADFM modules enhance the integration and effective use of data across modalities. These advancements lead to a marked increase in detection accuracy, especially in challenging scenarios, where substantial improvements are evident for the hard category.

5. Conclusions

This paper introduces BiDFNet, a novel 3D object detection framework integrating pseudo-point clouds and bidirectional feature fusion between point cloud and image data streams. Our approach, leveraging the proposed SAF-Conv, BiCSAFIM, and ADFM modules, ensures comprehensive information exchange across modalities while enhancing feature robustness, particularly for pseudo-LiDAR data. Experimental results on the KITTI dataset demonstrate that BiDFNet achieves state-of-the-art performance in the highly competitive Car category, significantly surpassing existing methods, especially in challenging 'Hard' scenarios. Furthermore, we validated the generalizability of our architecture by extending it to multi-class detection, showing consistent performance improvements for 'Pedestrian' and 'Cyclist' categories over the baseline (Table 3).

We acknowledge several limitations in the current study which outline important directions for future work. Firstly, while we demonstrated multi-class capability, the primary focus remained on the 'Car' category, and further dedicated evaluation and optimization are needed for pedestrian and cyclist detection to fully assess performance across all crucial road users . Secondly, and critically for autonomous driving applications, this work lacks a dedicated robustness analysis under adverse weather conditions (e.g., rain, snow) or scenarios simulating sensor degradation, which is essential for evaluating real-world reliability.Thirdly, while inference times were reported (Section 4.4), a comprehensive computational complexity analysis, including hardware-independent metrics like detailed efficiency comparisons, was not performed, limiting a full assessment of the model's computational demands.Future research will prioritize addressing these limitations. Specifically, we aim to conduct rigorous robustness testing, perform detailed computational benchmarking, and extend the optimization efforts to pedestrian and cyclist detection. Despite these acknowledged limitations, this study offers an efficient and innovative solution built upon attention-driven bidirectional fusion, establishing the effectiveness of the proposed components and providing a strong foundation for future advancements in robust and accurate multimodal 3D object detection.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Author Contributions: Conceptualization, Q.Z. and Y.W.; methodology, Q.Z.; software, Q.Z.; validation, Q.Z., and Y.W.; formal analysis, Q.Z.; investigation, Q.Z.; resources, Y.W.; data curation, Q.Z.; writing—original draft preparation, Q.Z.; writing—review and editing, Q.Z., Y.W.; visualization, Q.Z.; supervision, Y.W.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Project 2024JJ7428 of the Hunan Provincial Natural Science Foundation of China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in the KITTI dataset at http://www.cvlibs.net/datasets/kitti/.

Conflicts of Interest: the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
2. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*. https://doi.org/10.3390/s18103337.
3. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection From Point Clouds. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
4. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; Li, H. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35, pp. 1201–1209.
5. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21674–21683.
6. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
7. Shi, W.; Rajkumar, R. Point-gnn: Graph neural network for 3d object detection in a point cloud. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1711–1719.
8. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10529–10538.
9. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.
10. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum pointnets for 3d object detection from rgb-d data. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 918–927.
11. Xu, D.; Anguelov, D.; Jain, A. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 244–253.
12. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3d object detection. In Proceedings of the Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XV 16. Springer, 2020, pp. 35–52.
13. Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; Tang, Z. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems* **2022**, *35*, 10421–10434.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025

doi:10.20944/preprints202505.1480.v1

14. Li, Y.; Yu, A.W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q.V.; et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17182–17191.
15. Graham, B.; Van der Maaten, L. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307* **2017**.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
17. Beltrán, J.; Guindel, C.; Moreno, F.M.; Cruzado, D.; Garcia, F.; De La Escalera, A. Birdnet: a 3d object detection framework from lidar information. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 3517–3523.
18. Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; Du, D. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790* **2021**.
19. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.
20. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* **2017**, *30*.
21. Zhao, T.; Ning, X.; Hong, K.; Qiu, Z.; Lu, P.; Zhao, Y.; Zhang, L.; Zhou, L.; Dai, G.; Yang, H.; et al. Ada3d: Exploiting the spatial redundancy with adaptive inference for efficient 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17728–17738.
22. Wang, Z.; Jia, K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019, pp. 1742–1749.
23. Nabati, R.; Qi, H. Centerfusion: Center-based radar and camera fusion for 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 1527–1536.
24. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3d proposal generation and object detection from view aggregation. In Proceedings of the 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2018, pp. 1–8.
25. Yin, T.; Zhou, X.; Krähenbühl, P. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems* **2021**, *34*, 16494–16507.
26. Wang, C.; Ma, C.; Zhu, M.; Yang, X. Pointaugmenting: Cross-modal augmentation for 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11794–11803.
27. Wu, X.; Peng, L.; Yang, H.; Xie, L.; Huang, C.; Deng, C.; Liu, H.; Cai, D. Sparse fuse dense: Towards high quality 3d detection with depth completion. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5418–5427.
28. Li, Y.; Wang, Y.; Lu, Z.; Xiao, J. DepthGAN: GAN-based depth generation of indoor scenes from semantic layouts. *arXiv preprint arXiv:2203.11453* **2022**.
29. Mo, Y.; Wu, Y.; Zhao, J.; Hou, Z.; Huang, W.; Hu, Y.; Wang, J.; Yan, J. Sparse Query Dense: Enhancing 3D Object Detection with Pseudo points. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 409–418.
30. Wu, H.; Wen, C.; Shi, S.; Li, X.; Wang, C. Virtual sparse convolution for multimodal 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 21653–21662.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
32. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.
33. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *The international journal of robotics research* **2013**, *32*, 1231–1237.
34. Mao, J.; Niu, M.; Bai, H.; Liang, X.; Xu, H.; Xu, C. Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2723–2732.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 May 2025 doi:10.20944/preprints202505.1480.v1

35. Zhao, X.; Su, L.; Zhang, X.; Yang, D.; Sun, M.; Wang, S.; Zhai, P.; Zhang, L. D-conformer: Deformable sparse transformer augmented convolution for voxel-based 3d object detection. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
36. Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; Li, H. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. *International Journal of Computer Vision* **2023**, *131*, 531–551.
37. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In Proceedings of the Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16. Springer, 2020, pp. 720–736.
38. Wu, H.; Deng, J.; Wen, C.; Li, X.; Wang, C.; Li, J. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–11.
39. Yang, H.; Liu, Z.; Wu, X.; Wang, W.; Qian, W.; He, X.; Cai, D. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 662–679.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.