

Article

LSCA-RCNN: Large-Kernel Spatial Residual and Cascade Attention Network for Voxel-Based 3D Object Detection

Yuyang Liu ^{1,2,3,*}, Zhanyuan Jiang ¹, Min Mao ^{1,*}, Kun Zhang ¹, Yu Xu ¹, Mingchen Zhu ⁴ and Xianjun Wu ^{3,5}

¹ College of Physics and Electronic Engineering, Xinyang Normal University, Xinyang 464000, China; jiangzhanyuan2002@163.com (Z.J.); zk@xynu.edu.cn (K.Z.); xuyu20084636@126.com (Y.X.)

² School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

³ Xinyang Gumai Optonics Co., Ltd., Xinyang 464000, China

⁴ School of Geomatics Science and Technology, Nanjing Tech University, Nanjing 210009, China; zhu_mc@njtech.edu.cn

⁵ Xinyang Central Semiconductor Technology Co., Ltd., Xinyang 464000, China

* Correspondence: yuyangliu@xynu.edu.cn (Y.L.); maomin@xynu.edu.cn (M.M.)

Abstract

LiDAR-based 3D object detection remains challenging due to sparse and irregular point cloud distributions, which degrade detection accuracy for small and occluded objects. In view of this, this paper proposes a novel two-stage voxel-based 3D detector, namely LSCA-RCNN, to address these issues. First, spatial residual blocks (SRBs) and large-kernel spatial-wise convolutions are integrated into the 3D backbone to suppress feature degradation and to expand the receptive fields for stable multi-scale feature learning. Second, a ConvNeXt-based 2D backbone with spatial attention is constructed to enhance discriminative feature representation of small objects. Third, a cascaded detection head embedded with fine-grained grouped convolutions and cross-stage cross-attention is designed to achieve progressive bounding box refinement and to improve localization precision. Extensive evaluations on the KITTI dataset with the R40 metric show that the proposed method achieves consistent performance improvements over the baseline. In the moderate setting, LSCA-RCNN increases the 3D AP by 2.12%, 7.66%, and 5.43% for cars, pedestrians, and cyclists, respectively, while achieving gains of 1.62%, 5.05%, and 7.05% under the hard setting. These results validate the effectiveness and robustness of the proposed LSCA-RCNN for complex and challenging autonomous driving detection tasks.

Keywords: lidar point cloud; spatial residual blocks (SRBs); spatial-wise convolutions; ConvNeXt-based 2D backbone; 3D object detection; cascade detection head

Academic Editors: Junxing Zheng and Peng Cao

Received: 21 May 2026

Revised: 20 June 2026

Accepted: 24 June 2026

Published: 27 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Recent advances in autonomous driving technologies have made LiDAR a core sensor for environmental perception. 3D object detection serves as a fundamental task in autonomous driving, robotic navigation, and intelligent transportation systems, enabling intelligent platforms to perceive and understand spatial scenes accurately [1]. As a high-precision active sensing modality, LiDAR reliably captures the geometric coordinates, reflection intensity, and structural characteristics of outdoor environments, providing robust perceptual cues for complex driving scenarios [2]. However, LiDAR point clouds are inherently sparse and irregular, with spatially uneven density and significant cross-object

scale variations. These intrinsic properties pose major obstacles for accurate and robust 3D object detection and limit the practical deployment of detection algorithms in real-world complex scenarios.

The existing LiDAR-based 3D object detection methods fall into three dominant categories: point-based, voxel-based, and hybrid point-voxel approaches. Point-based models, including PointNet [3], PointNet++ [4], PointRCNN [5], and 3DSSD [6], process raw point clouds directly and preserve fine-grained geometric information. Nevertheless, these methods incur substantial computational overhead, which constrains their real-time performance on large-scale point cloud data. By contrast, voxel-based methods discretize unstructured point clouds into regular 3D voxel grids and adopt sparse 3D convolutions for efficient spatial feature extraction, with representative architectures, such as SECOND [7], VoxelNet [8], and Voxel R-CNN [9]. Hybrid point-voxel methods exemplified by PV-RCNN and PV-RCNN++ combine voxel-based feature encoding with point-level localization to balance detection accuracy and computational efficiency [10,11]. Even so, their multi-branch architecture introduces inherent feature redundancy and training optimization instability, restricting further performance gains and practical deployment.

Voxel R-CNN is a high-performance two-stage 3D detector that employs voxel features for both proposal generation and bounding box refinement. It avoids dense point sampling and complicated multi-source feature aggregation, yielding high inference efficiency. Owing to structured voxel representation and sparse 3D convolutions, Voxel R-CNN has been widely adopted as a benchmark in both academic research and industrial applications. Despite its merits, this method still suffers from noticeable performance degradation in complex traffic scenarios. First, the backbone network lacks effective residual feature propagation. As network depth increases, geometric details and edge features of small and distant objects gradually decay, resulting in feature degradation. This issue severely impairs the detection of partially occluded pedestrians and distant cyclists, leading to frequent missed detections and low classification confidence. Second, the bird's-eye view (BEV) feature extraction module lacks adaptive feature calibration. Standard convolutions fail to enhance foreground responses or to suppress background noise adaptively, thereby reducing feature discriminability for small and overlapping objects. Third, the single-stage proposal refinement mechanism has inherent limitations. It cannot fully correct coarse localization errors produced by the region proposal network (RPN). The accumulated regression errors further deteriorate localization accuracy in densely distributed and heavily occluded scenes, increasing false-positive detection rates.

To address the above limitations, this work proposes a novel 3D detection framework named LSRA-RCNN. The proposed framework integrates a spatial convolution-enhanced backbone optimized by SRBs, a ConvNeXt-based 2D backbone, and an attention-guided cascaded refinement strategy to achieve accurate and robust 3D object detection [12].

The main contributions of this paper are summarized as follows:

1. We propose a multi-scale residual 3D backbone composed of SRBs and spatial wise convolution modules to alleviate feature degradation by strengthening cross-layer feature propagation and by enhancing multi-scale contextual modeling with enlarged receptive fields.
2. We design a ConvNeXt-based 2D backbone with large scale spatial perception and lightweight spatial attention to enlarge the effective receptive field, to improve small object representation, and to enhance feature learning stability.
3. We propose a cascade component aware detection head with progressive cascade refinement and cross-stage cross-attention to improve fine grained feature perception and to reduce localization errors in single stage RoI prediction.

4. Extensive experiments on the KITTI dataset demonstrate the effectiveness and robustness of the proposed method.

2. Related Work

2.1. Voxel-Based 3D Detection

Voxel-based detection methods transform irregular point clouds into regular voxel grids, enabling convolutional networks to efficiently learn spatial features. Due to their favorable trade-off between accuracy and efficiency, voxel-based frameworks have become mainstream solutions for LiDAR-based 3D perception in autonomous driving. VoxelNet first introduced the voxel feature encoding (VFE) pipeline for end-to-end 3D object detection [8]. However, its dense 3D convolutions incur heavy computation on empty voxels, limiting the inference speed. SECOND addressed this issue by adopting sparse 3D convolutions to skip invalid voxels, significantly improving efficiency [7]. PointPillars further simplified voxel representation into 2D pillar features, achieving real-time detection performance [13]. These methods mainly belong to single-stage detectors, which emphasize speed but often struggle with small and occluded objects.

Two-stage voxel detectors instead focus on localization accuracy. Voxel R-CNN proposed a fully voxel-based two-stage framework with voxel RoI pooling, achieving a strong balance between accuracy and efficiency [9]. Nevertheless, it still suffers from two limitations: the backbone lacks effective residual connections, causing feature degradation in deep layers, and small convolution kernels provide insufficient receptive fields for capturing long-range spatial dependencies. Although methods such as JPV-Net employ multi-scale feature fusion to improve representation [14], they do not fundamentally resolve deep feature attenuation or limited receptive fields.

To address these limitations, this work reconstructs the 3D backbone of Voxel R-CNN by integrating spatial residual blocks (SRBs) and large-kernel spatial convolutions, thereby enhancing cross-layer feature propagation and enlarging the effective receptive field simultaneously.

2.2. Optimization of 3D Feature Enhancement

Feature representation is crucial for 3D object detection performance. The existing feature enhancement methods are mainly based on residual learning and attention mechanisms. Residual learning improves gradient propagation and preserves low-level spatial details through skip connections [15]. However, most residual designs in LiDAR detection are tailored for hybrid point-voxel frameworks, such as PV-RCNN [10,11], while pure voxel-based two-stage detectors rarely incorporate dedicated residual structures. As a result, Voxel R-CNN suffers from the progressive loss of fine-grained features for pedestrians and cyclists in deep networks.

Attention mechanisms adaptively recalibrate features, suppress background noise, and emphasize foreground targets [16]. Representative works include SENet [17], TANet [18], Bi-Att3DDet [19], and LoGoNet [20], which extend the attention mechanisms to multi-scale and multimodal scenarios. Nevertheless, most existing approaches focus on channel attention or modality fusion, while the spatial attention for BEV feature maps remains insufficiently explored. Since the BEV features encode critical spatial localization information, the lack of spatial calibration limits the discrimination of small objects from background clutter.

To overcome these issues, we optimize feature representation from two aspects. First, SRBs and large-kernel spatial-wise convolutions are embedded into the 3D sparse backbone to alleviate feature degradation and to enlarge receptive fields [21,22]. Second, a

ConvNeXt-based 2D BEV encoder with spatial attention is constructed to enhance small-object representation and to improve the robustness of the BEV feature learning [12].

2.3. Proposal Refinement and Cascade Optimization

Accurate localization in two-stage detectors depends heavily on proposal quality and iterative refinement. Conventional frameworks use only a single-stage regression head to refine the RPN proposals, making it difficult to fully correct localization errors, especially in dense and occluded scenes [23].

Cascade refinement progressively optimizes proposals through multiple regression stages. 3D Cascade R-CNN improves localization by increasing the IoU thresholds across stages [24], while CasA introduces attention-based cross-stage feature aggregation for better refinement [25]. However, the existing cascade methods are mostly designed for hybrid point-voxel architectures and are not well suited to the pure voxel-based structure of Voxel R-CNN. Moreover, they lack fine-grained modeling for small objects, limiting their effectiveness for pedestrians and cyclists.

To address these shortcomings, we design a voxel-oriented cascaded detection head tailored to Voxel R-CNN. The proposed module combines multi-stage progressive regression with cross-stage cross-attention to reduce any accumulated localization errors. In addition, fine-grained grouped convolutions and component-aware mechanisms are introduced to enhance local structural perception for small targets, leading to more accurate localization in complex driving scenarios.

3. Methods

To address the limitations described above, this paper proposes a novel 3D object detection framework, named LSCA-RCNN. The proposed framework integrates a multi-scale residual 3D backbone, a large-scale spatial perception BEV encoder, and a cascade component-aware RoI detection head. These elaborately designed modules collaboratively alleviate the deep feature degradation problem, while substantially enhancing the multi-scale contextual representation capability and object localization accuracy. The overall architecture of the proposed LSCA-RCNN framework is illustrated in Figure 1, which consists of three core functional components.

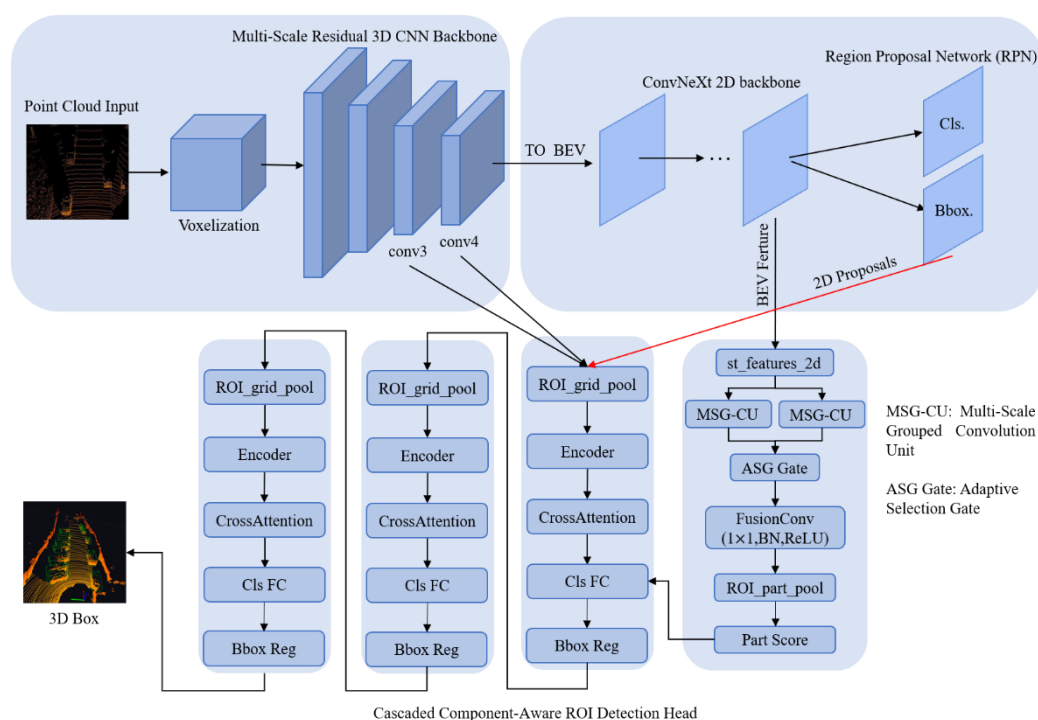


Figure 1. Overall architecture of the LSCA-RCNN. The framework consists of a multi-scale residual 3D CNN backbone with SRBs and large-kernel spatial convolutions for robust feature propagation, a ConvNeXt 2D backbone with spatial attention for enhanced BEV encoding, and a cascaded component-aware RoI head with three-stage progressive refinement and an auxiliary FGCE branch with MSG-CU and ASG for part-aware scoring.

3.1. Multi-Scale Residual 3D CNN Backbone Network

The baseline Voxel R-CNN adopts a 3D sparse convolutional backbone stacked with multi-level SubM blocks [9]. Each block contains two $3 \times 3 \times 3$ SubM sparse convolution layers followed by batch normalization (BN) and ReLU activation. However, this sequential architecture suffers from several limitations. Sparse voxel features tend to degrade during deep feature propagation, resulting in weakened representations for small objects, such as pedestrians and cyclists. Moreover, small convolution kernels provide limited receptive fields and are insufficient for capturing long range spatial dependencies in point clouds. Although large convolution kernels can capture long range context, they significantly increase the computational cost and model complexity. To address these limitations, we design an improved 3D sparse convolutional backbone network, as shown in Figure 2. Dedicated enhancement modules are introduced for both shallow and deep layers. These modules enhance the multi-scale contextual modeling and stabilize the deep feature propagation.

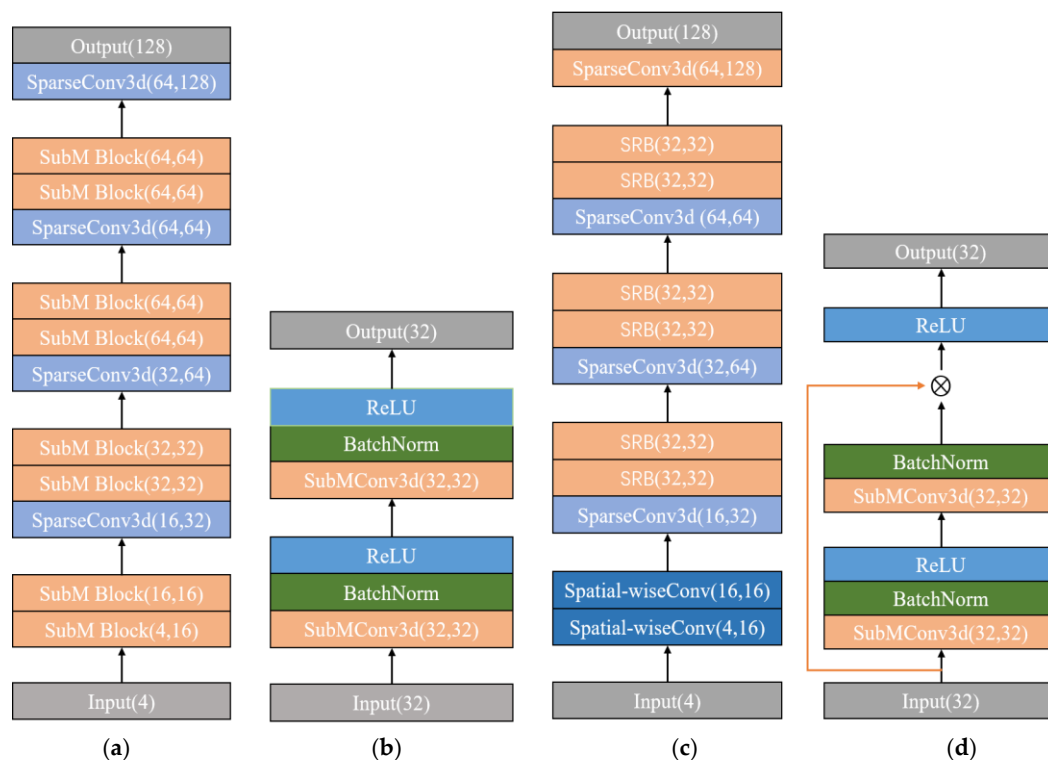


Figure 2. Comparison of backbone structures, where (a,b) illustrate the modules of Voxel R-CNN and (c,d) show the structures of the proposed method: (a) 3D backbone structure of Voxel R-CNN; (b) SubM block; (c) Multi-scale residual network structure; (d) Spatial Residual Block (SRB).

3.1.1. Spatial Residual Block

To alleviate feature degradation in the baseline backbone, the spatial residual block (SRB) is introduced to replace the original vanilla SubM blocks [21]. As illustrated in Figure 2d, the SRB employs two consecutive submanifold sparse convolution layers as the main feature extraction branch. Meanwhile, an identity residual connection is

incorporated to fuse shallow geometric information with high-level semantic representations across the network layers. The aggregated features are activated by a ReLU function, facilitating stable gradient propagation during training. By strengthening cross-layer information flow, the SRB effectively preserves geometric details and mitigates feature degradation in deep sparse convolutional networks. As a result, it enhances the representation capability for small objects and improves the overall robustness of the detector.

3.1.2. Spatial Dimensional Convolutional Modules

To strengthen the multi-scale spatial context modeling in the shallow feature encoding, a dual-branch parallel architecture is introduced in the shallow stages of the backbone network, as illustrated in Figure 3a. The two branches adopt $5 \times 5 \times 5$ large-kernel convolutions and $3 \times 3 \times 3$ SubMConv3d, respectively. A weight transformation mechanism is further introduced to enable parameter sharing across the multi-scale convolution kernels and to facilitate cross-branch feature fusion. The spatial-wise convolution module leverages large-kernel operations to expand the global receptive field and to capture long-range spatial dependencies within the sparse point cloud data [21,22]. Meanwhile, the embedded parameter sharing strategy eliminates redundant network parameters, reduces the optimization difficulty of large-kernel convolutions, and mitigates the risk of overfitting.

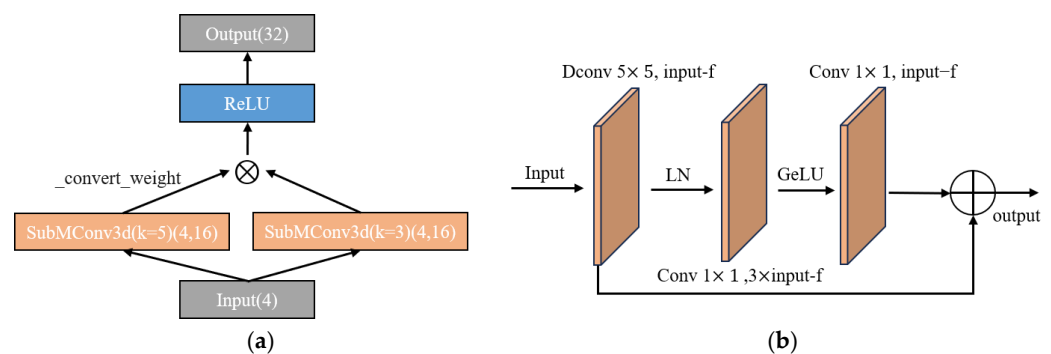


Figure 3. Illustration of key network modules: (a) Spatial-wise convolution module; (b) ConvNeXt module.

3.2. ConvNeXt 2D Backbone Network

The baseline model utilizes stacked 3×3 standard convolutions for the BEV feature encoding, which inherently suffers from constrained receptive fields, redundant computational overhead, unstable mini-batch training, and insufficient modeling capability for small-scale contextual features. To overcome these drawbacks, this paper proposes a ConvNeXt-based 2D backbone network, as depicted in Figure 3b. Derived from the classical ResNet architecture, ConvNeXt adopts 7×7 depthwise separable convolutions for hierarchical feature extraction [12,26,27]. Different from conventional standard convolutions, this design efficiently expands the network receptive field with limited computational and parameter costs, enabling superior capture of fine-grained features and contextual information for small objects.

Equipped with LayerNorm normalization, GELU activation functions, and learnable gamma scaling parameters, the proposed 2D backbone network mitigates gradient vanishing issues and stabilizes mini-batch training. Furthermore, a spatial attention module is embedded at the end of each network stage [28]. This module generates 2D spatial attention maps by aggregating the channel-wise average and maximum pooling features, which adaptively enhances foreground target responses and suppresses irrelevant background noise on the BEV feature plane. Additionally, a multi-scale feature fusion strategy

is integrated into the framework. Hierarchical down-sampling operations are adopted for cross-stage feature compression and semantic encoding, while deconvolution-based up-sampling is employed to reconstruct the multi-scale BEV feature representations. The finally aggregated discriminative features provide reliable feature priors for subsequent RoI refinement, object classification, and bounding box regression tasks.

3.3. Cascaded Component Aware RoI Detection Head

The conventional single-stage RoI prediction paradigm adopted in baseline detectors often suffers from localization errors and insufficient bounding-box refinement, particularly for small and sparsely distributed objects. To address these limitations, a cascaded component-aware RoI detection head is proposed, as illustrated in Figure 4a. Inspired by cascaded attention networks, the proposed design extends the cascaded optimization strategy to 3D LiDAR-based object detection by constructing a multi-stage attention-guided refinement framework within the RoI head [29].

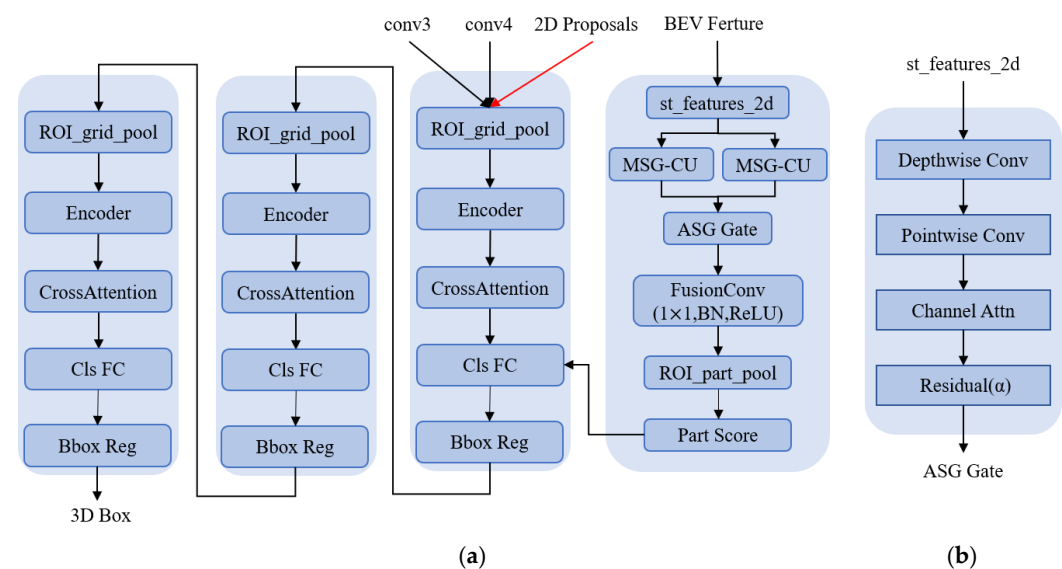


Figure 4. Structure of the cascaded component aware ROI detection head and multi-scale grouped convolution unit: (a) Cascaded component aware RoI detection head, consisting of the cascaded detection head and the fine-grained grouped convolutional component aware module; (b) Multi-scale grouped convolution unit (MSG-CU).

The cascaded structure progressively refines object proposals and facilitates more effective feature learning across stages. In addition, the component-aware feature extraction mechanism enhances the perception of fine-grained structural information, leading to more discriminative object representations. Through iterative optimization, the proposed detection head improves the localization quality and feature refinement, thereby achieving a more robust detection performance for challenging targets, such as pedestrians and cyclists in complex traffic environments.

3.3.1. Fine-Grained Grouped Convolutional Component-Aware Module

To strengthen the structural feature modeling of small targets, this work embeds a fine-grained part-aware branch into the cascaded detection head. Different from conventional standard convolution-based detection branches, this paper proposes a fine-grained group convolution enhancement (FGCE) module built upon depthwise separable convolutions. This module is specifically tailored for small objects (e.g., pedestrians and cyclists) that suffer from limited BEV projection areas, sparse point cloud distributions, and insufficient structural feature representation. Since standard convolutions fail to capture the

subtle structural correlations of such small targets effectively, a multi-scale grouped convolution unit (MSG-CU) is further designed to refine fine-grained feature learning, with its detailed structure depicted in Figure 4b.

Given the input feature $F \in R^{B \times C \times H \times W}$, MSG-CU first splits the input channel dimension into $G = 16$ disjoint groups: $F = [F^{(1)}; F^{(2)}; \dots; F^{(G)}]$, where each group feature satisfies $F^{(g)} \in R^{B \times \frac{C}{G} \times H \times W}$. Each group is independently processed via depthwise separable convolution to extract group-specific spatial features, formulated as follows:

$$F_{dw}^{(g)} = W_{dw}^{(g)} * F^{(g)}, g = 1, 2, \dots, G \quad (1)$$

where $W_{dw}^{(g)} \in R^{k \times k \times \frac{C}{G}}$ denotes the learnable depthwise convolution kernel, and the kernel size $k \in (3, 5)$ for multi-scale feature modeling. Subsequently, pointwise convolution is adopted to achieve adaptive cross-group feature fusion and to integrate multi-group complementary information, which can be expressed as follows:

$$F_{pw} = W_{pw} \cdot \text{Concat}[F_{dw}^{(1)}, \dots, F_{dw}^{(G)}], W_{pw} \in R^{C_{out} \times C} \quad (2)$$

To enhance the channel-wise feature sensitivity and to suppress redundant channel responses, a channel attention mechanism is introduced to recalibrate the fused features, as follows:

$$\omega_c = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W F_{pw}(i, j))) \quad (3)$$

where $W_1, W_2 \in R^{C_{out} \times \frac{C_{out}}{4}}$ denotes the linear projection layers. The recalibrated feature is obtained as $F_{ca} = F_{pw} \odot \omega_c$. Furthermore, a learnable residual scaling factor $\alpha = \tanh(\alpha_l)$ is incorporated, which stabilizes network training at the initial stage and adaptively regulates the degree of feature enhancement during optimization.

To address the inherent scale discrepancy between pedestrians and cyclists in the BEV feature maps, the FGCE module adopts a dual-branch multi-scale design. The 3×3 convolution branch captures fine-grained local structural features for pedestrian perception, while the 5×5 convolution branch leverages a larger receptive field to model the global structural characteristics of cyclists. An adaptive spatial gating mechanism $G = \sigma(W_g * F_{in})$ is employed to fuse dual-branch features, which dynamically adjusts the contribution of each branch according to spatial position variations. Finally, the component-aware perceptual features are integrated with geometric classification outputs to generate the final prediction results, providing explicit structural prior knowledge to facilitate the high-precision detection of small and easily confused traffic objects.

3.3.2. Cross-Stage Cross-Attention

To facilitate efficient global feature propagation and to eliminate redundant computation across cascaded optimization stages, a shared feature connection layer is embedded at the input of the cascaded detection head. On this basis, a cross-attention mechanism is integrated to enable sufficient feature interaction and complementary learning between adjacent stages [30]. Let $\hat{F}^j$ and F^j denote the feature maps derived from the previous and current cascaded stages, respectively. The cross-attention calculation is formulated as follows:

$$\hat{F}_i^j = \text{softmax}\left(\frac{Q_i^j (K_i^j)^T}{\sqrt{c^j}}\right) V_i^j \quad (4)$$

where Q^j , K^j , and V^j represent the query, key, and value embeddings linearly projected from stage-specific features, respectively. This mechanism enables the network to

adaptively model long-range dependencies across different cascaded stages, thereby achieving hierarchical feature aggregation and significantly improving feature representation robustness. The refined multi-stage interactive features are ultimately fed into the classification and regression branches to produce accurate 3D object detection results. The overall pipeline of the proposed cross-stage cross-attention module is illustrated in Figure 4a.

3.4. Loss Function

Unlike the original Voxel R-CNN, which employs a single-stage RoI refinement head, the proposed framework adopts a three-stage cascade refinement strategy. Accordingly, the overall training objective is formulated as the sum of the RPN loss and the losses from all cascade stages: $L = L_{RPN} + \sum_{k=1}^K L_{roi}^{(k)}$, where L_{RPN} is the RPN loss, $L_{roi}^{(k)}$ denotes the loss of the (k)-th cascade stage, and ($K = 3$) is the total number of cascade stages. All cascade stages are equally weighted during training. The detailed formulations are defined as follows:

$$L_{RPN} = \frac{1}{N_p} \left[\sum_i L_{cls}(a_i, \hat{a}_i) + \lambda_1 I(IoU_i > u) \sum_i L_{reg}(\delta_i, \hat{\delta}_i) \right] \quad (5)$$

$$L_{roi}^{(k)} = \frac{1}{N_b} \left[\sum_j L_{cls}(a_j^{(k)}, \hat{a}_j^{(k)}) + \lambda_2 I(IoU_j^{(k)} > u^{(k)}) \sum_j L_{reg}(\delta_j^{(k)}, \hat{\delta}_j^{(k)}) \right] \quad (6)$$

For L_{RPN} , N_p denotes the number of anchors; a_i and δ_i represent the predicted classification scores and bounding box offsets, while $\hat{a}_i$ and $\hat{\delta}_i$ denote the ground-truth labels and targets; $I(\cdot)$ is the indicator function with $u = 0.5$. $L_{cls}(\cdot)$; and $L_{reg}(\cdot)$ denotes the focal loss and smooth L_1 loss, with a balancing weight $\lambda_1 = 1.0$. For the K -th stage $L_{roi}^{(k)}$, N_b denotes the number of RoIs; $a_j^{(k)}$ and $\delta_j^{(k)}$ are the predicted outputs, while $\hat{a}_j^{(k)}$ and $\hat{\delta}_j^{(k)}$ are the assigned labels and targets.

4. Experiment

4.1. Dataset and Evaluation Metrics

The KITTI dataset is one of the most authoritative and widely adopted benchmarks for LiDAR-based 3D object detection in autonomous driving scenarios, owing to its diverse traffic scenes, high-precision annotations, and reliable data quality [31]. This study strictly follows the official data partition protocol of Voxel R-CNN for a fair comparison, dividing the entire trainval set into 3712 training samples and 3769 validation samples. Consistent with the official KITTI evaluation criteria, all test samples are classified into three difficulty levels (easy, moderate, and hard) according to the degree of object occlusion and shooting distance. In this work, 3D average precision (AP) with 40 recall points (R40) is adopted as the core evaluation metric [32]. Standard IoU thresholds are employed for quantitative evaluation, including 0.7 for car detection and 0.5 for pedestrian and cyclist detection tasks.

4.2. Experiment Details

All experiments are conducted on a server equipped with two NVIDIA GeForce RTX 3090 GPUs (24 GB VRAM) under the Ubuntu 20.04 LTS operating system. The programming language used is Python 3.6.9. The software environment is strictly unified for reproducibility: PyTorch 1.8.0, CUDA 11.1, cuDNN 8.0.5. The adopted third-party libraries

include sparse 3D convolution spconv 2.1.25 and NumPy 1.21.6 and Open3D 0.11.0 Scikit-learn 1.3.2

We adhere to the standard preprocessing workflow of Voxel R-CNN for the coordinate transformation and value normalization of raw LiDAR point clouds. To remove invalid distant points, we constrain all point coordinates within the range $x \in [0, 70.4]$, $y \in [-40, 40]$, $z \in [-3, 1]$. The voxelization hyperparameters are predefined: the voxel dimension is (0.05 m, 0.05 m, 0.05 m), the maximum number of points contained in a single voxel is 5, and the maximum voxel quantity is set to 16,000 for the training set and 40,000 for the test set [9]. A series of conventional augmentation strategies for LiDAR 3D detection are utilized during training only, namely *gt_sampling*, random world flip, random world rotation, random world scaling, and random local pyramid augmentation. These augmentations are disabled for validation and inference.

The network is trained over 80 epochs with a per-GPU batch size of 4, using the AdamW optimizer for parameter updating. For a rigorous and fair experimental comparison, all configurations, including hyperparameters, weight initialization rules, and random seed policies, strictly conform to the Voxel R-CNN baseline, and no extra hyperparameter adjustment is implemented in our experiments.

4.3. Comparison with Other Algorithms

To fully evaluate the performance of the proposed LSCA-RCNN, we conduct extensive comparative experiments on the KITTI validation set. The proposed framework is compared against several state-of-the-art 3D object detection approaches, with the quantitative results listed in Tables 1 and 2. Additional comparisons with representative multimodal detectors, including Bi-Att3DDet, LoGoNet, SQD, SFD, and VirConv-L, are presented in Table 3 [19,20,33–35].

As shown in Table 1, LSCA-RCNN achieves 3D AP scores of 86.64%, 67.33%, and 76.37% for cars, pedestrians, and cyclists, respectively, under the moderate difficulty setting. Compared with the baseline Voxel R-CNN, the proposed method improves the 3D AP by 2.12%, 7.66%, and 5.43% for the three categories, demonstrating consistent accuracy gains while maintaining balanced performance across different object classes. Notably, LSCA-RCNN outperforms several representative two-stage LiDAR-based detectors, including PV-RCNN++, CasA, and CIA-3D, in pedestrian and cyclist detection, indicating its effectiveness in handling small and sparsely distributed objects.

To further assess the robustness of the reported results without incurring the high computational cost of multiple independent training runs (approximately 20 GPU-hours per run), a bootstrap resampling analysis is conducted on the predictions of a single trained model. Specifically, 3769 test samples are randomly sampled with replacement to generate $N = 100$ bootstrap datasets. The KITTI AP metric is recomputed for each dataset, and the final performance is reported as the mean $\pm$ standard deviation over all bootstrap iterations. As shown in the last row of Table 1, the proposed method achieves low performance variance across the different categories and difficulty levels, demonstrating the statistical reliability and robustness of the reported results.

Table 1. Comparison of Different Algorithms on the Validation Set. The bold results indicate the best performance. Note: Results marked with (*) are reproduced by our team under identical training, validation, and test configurations to ensure fair comparison; unmarked results are directly cited from the original published papers.

Model	Stage	Car.3D (APR40)			Ped.3D (APR40)			Cyc.3D (APR40)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
SECOND * [7]	1-stage	88.61	78.62	77.22	56.55	52.98	47.73	80.58	67.15	63.10
PointPillars * [13]	1-stage	86.46	77.28	74.65	57.75	52.29	47.90	80.05	62.68	69.70
VFL3D [36]	1-stage	91.89	85.11	82.64	59.90	53.35	49.11	84.05	65.86	61.97
PointRCNN * [5]	2-stage	91.78	80.72	78.22	60.01	53.38	45.78	90.60	71.80	67.45
Part A2 * [37]	2-stage	89.36	80.23	78.88	65.70	61.32	55.40	85.60	69.24	65.63
PV-RCNN++ * [11]	2-stage	90.85	83.66	81.56	63.36	57.55	53.11	87.93	69.77	65.97
CIANet [38]	2-stage	89.68	84.89	79.32	67.95	62.54	56.91	87.12	73.78	70.69
RCAVoxel-RCNN [39]	2-stage	92.81	85.78	83.51	71.46	64.74	58.72	92.92	76.61	71.62
CasA * [29]	2-stage	90.01	86.30	82.83	73.89	66.55	59.89	92.61	73.04	68.44
Voxel R-CNN * [9]	2-stage	92.24	84.52	82.54	65.51	59.67	53.73	90.72	70.94	66.88
Ours	2-stage	93.27	86.64	84.16	74.57	67.33	60.78	91.71	76.37	71.93
Bootstrap mean $\pm$ standard deviation		93.30 $\pm$ 0.18	86.39 $\pm$ 0.75	83.34 $\pm$ 1.08	74.91 $\pm$ 1.58	67.05 $\pm$ 1.44	60.80 $\pm$ 1.30	92.22 $\pm$ 1.66	76.48 $\pm$ 2.16	72.21 $\pm$ 2.35

Table 2 further presents the 3D and BEV detection results for the car category. LSCA-RCNN achieves 3D AP scores of 93.27%, 86.64%, and 84.16% under the easy, moderate, and hard settings, respectively. Under the moderate setting, it surpasses Voxel R-CNN by 2.12% in 3D AP and 1.02% in the BEV AP, demonstrating its strong capability in spatial feature representation and object localization. Although some methods, such as SA-Voxel R-CNN, and TED, achieve slightly higher 3D AP under the hard setting, LSCA-RCNN still maintains a competitive 3D AP of 84.16% while achieving a BEV AP of 89.90%. Moreover, it delivers a consistently strong performance across all evaluation settings, indicating a favorable balance between localization accuracy and overall detection robustness.

Table 2. Comparison of Different Algorithms for the Car Category on the KITTI Validation Set in terms of 3D and BEV Performance. The bold results indicate the best performance. Note: Results marked with (*) are reproduced by our team under identical training, validation, and test configurations to ensure fair comparison; unmarked results are directly cited from the original published papers. The symbol “-” indicates that the corresponding data were not reported in the original publication.

Model	Reference	Modality	Car.3D (APR40)			Car.BEV (APR40)		
			Easy	Mod.	Hard	Easy	Mod.	Hard
PV-RCNN++ * [11]	IJCV 2023	LiDAR	90.85	83.66	81.56	92.90	89.64	88.79
CasA * [29]	IEEE TGRS 2022	LiDAR	90.01	86.30	82.83	93.93	90.22	87.74
TED [40]	AAAI 2023	LiDAR	92.59	85.81	85.26	95.05	91.43	91.22
SA-Voxel-RCNN [41]	WEVJ 2025	LiDAR	93.94	86.62	83.54	95.43	91.29	90.03
SPS-RCNN [42]	Sensors 2025	LiDAR	92.79	85.49	83.19	-	-	-
Voxel R-CNN * [9]	AAAI 2021	LiDAR	92.24	84.52	82.54	95.34	91.17	88.68
AVOD [43]	IROS 2018	LiDAR + RGB	84.41	74.44	68.65	86.30	85.44	77.73
MV3D [44]	CVPR 2017	LiDAR + RGB	71.29	62.68	56.56	86.02	76.90	68.49
EPNet [45]	ECCV 2020	LiDAR + RGB	92.28	82.59	80.14	94.22	88.47	83.69
Ours	-	LiDAR	93.27	86.64	84.16	95.96	92.19	89.90

By jointly analyzing the results in Tables 1 and 2, we further investigated the effectiveness of LSCA-RCNN across different object categories and the factors influencing its

detection performance. Compared with Voxel R-CNN, the proposed method achieves consistent improvements for cars, pedestrians, and cyclists. However, the magnitude of the improvement varies across the categories.

Cars usually contain dense point clouds and well-preserved geometric structures, allowing most existing detectors to localize them accurately. As a result, most of the existing algorithms could achieve a higher performance for cars. By contrast, pedestrians and cyclists are considerably more challenging due to their small physical size, sparse point cloud distributions, frequent occlusions, and limited BEV projection areas. These characteristics often lead to incomplete feature representations and a reduced detection accuracy.

The larger improvements observed for pedestrians and cyclists demonstrate the effectiveness of the proposed architecture. By combining multi-scale feature extraction, enhanced BEV feature encoding, and cascaded RoI refinement, LSCA-RCNN can capture more discriminative fine-grained features and better preserve the structural information for small objects. Consequently, the proposed method achieves superior robustness and generalization capability when detecting challenging road users in complex autonomous driving scenarios.

Table 3. Comparison with Multimodal Methods on the Validation Set. Best results are highlighted in bold. The bold results indicate the best performance.

Model	Reference	Car.3D (APR40)			Ped.3D (APR40)			Cyc.3D (APR40)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
LoGoNet [20]	CVPR 2023	92.04	85.04	84.31	70.20	63.72	59.46	91.74	75.35	72.42
Bi-Att3DDet [19]	Sensors 2025	95.25	88.32	85.91	74.22	67.57	61.14	95.46	73.60	68.90
SQD [33]	ACM MM 2024	95.18	84.82	86.07	-	-	-	-	-	-
SFD [34]	CVPR 2022	95.47	88.56	85.74	-	-	-	-	-	-
VirConv-L [35]	CVPR 2023	93.36	88.71	85.83	-	-	-	-	-	-
Ours	-	93.27	86.64	84.16	74.57	67.33	60.78	91.71	76.37	71.93

As shown in Table 3, LSCA-RCNN is compared with several state-of-the-art multimodal detection methods. Although these approaches exploit the complementary information from both LiDAR and images, LSCA-RCNN, as a purely voxel-based single-modal framework, still achieves a competitive performance on multiple evaluation metrics. For pedestrian detection under the moderate setting, LSCA-RCNN achieves a 3D AP of 67.33%, which is comparable to Bi-Att3DDet (67.57%) and surpasses LoGoNet (63.72%). For cyclist detection, LSCA-RCNN attains 76.37% for 3D AP under the moderate setting, outperforming LoGoNet (75.35%) and remaining competitive with Bi-Att3DDet.

These results demonstrate the effectiveness of the proposed feature representation and component-aware detection strategy for challenging small-object detection tasks. Although multimodal methods generally benefit from richer complementary information, LSCA-RCNN achieves a comparable performance while relying solely on LiDAR point clouds. This indicates that the proposed voxel feature enhancement and cascaded detection framework can effectively exploit geometric information from sparse point clouds, providing a favorable balance between detection accuracy, model simplicity, and computational efficiency.

Figure 5 presents a qualitative visualization comparison between the proposed LSCA-RCNN and the baseline Voxel R-CNN, with the samples arranged as ground truth, Voxel R-CNN, and LSCA-RCNN from left to right. The visualization results intuitively demonstrate the superior performance of LSCA-RCNN in complex traffic scenarios, especially for distant and occluded small objects. In the first two rows of visualization cases, the baseline Voxel R-CNN fails to detect distant vehicles, mid-road pedestrians, and bus

stop-occluded objects, whereas LSCA-RCNN successfully identifies these challenging targets. Meanwhile, LSCA-RCNN effectively reduces false-positive detections. In the last two rows, Voxel R-CNN misclassifies distant tree shadows and roadside vegetation as cars and cyclists, while the proposed method suppresses such invalid interference and achieves accurate detection.



Figure 5. Visualization of detection results on the KITTI dataset. Blue boxes represent pedestrians, yellow boxes represent cyclists, and green boxes represent vehicles. The red box denotes the comparison region.

Although a small number of false negatives still exist in the detection results, LSCA-RCNN achieves a comprehensive performance improvement over the baseline model. The qualitative results further verify that the proposed modules can significantly enhance the model's perception ability for sparse, small, and occluded objects in complex scenes, fully demonstrating the effectiveness and superiority of the proposed LSCA-RCNN framework.

4.4. Ablation Experiment

Effectiveness of the proposed 3D backbone: To evaluate the effectiveness of different 3D backbone designs, we compared the original Voxel R-CNN with two backbone variants on the KITTI validation set, and the AP40 results for pedestrian and cyclist detection are reported in Table 4. The backbone incorporating only the spatial-dimensional convolution module (SDCM) achieves the highest cyclist detection accuracy across all difficulty levels, whereas its pedestrian detection performance is inferior to that of the spatial residual block (SRB)-based backbone. The SRB effectively alleviates feature degradation during deep feature propagation, resulting in more stable feature learning and superior pedestrian detection, although its gains for cyclists are relatively limited. To exploit the complementary strengths of both modules, we integrate an SDCM and an SRB into a unified backbone. The resulting hybrid architecture achieves the best pedestrian detection performance across all difficulty levels and consistently improves cyclist detection. Although it does not attain the highest score in every category, it delivers the strongest overall performance. These results demonstrate that the proposed backbone effectively enlarges the receptive field of shallow layers while mitigating deep feature degradation, thereby enhancing feature representation and improving detection accuracy.

Table 4. Ablation Results of Different 3D Backbone Designs for Pedestrian and Cyclist Detection on the KITTI Validation Set. The bold results indicate the best performance.

Model	Ped.3D (APR40)			Cyc.3D (APR40)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Voxel R-CNN	65.51	59.67	53.73	90.72	70.94	66.88
Voxel R-CNN + SDCM	71.92	64.74	58.18	91.95	74.41	69.96
Voxel R-CNN + SRB	72.27	65.60	59.34	89.13	72.57	68.34
Voxel R-CNN + SDCM + SRB	72.77	65.90	60.88	91.62	73.88	69.49

Effectiveness of the proposed detection head: To evaluate the effectiveness of the proposed detection head, we compared the original Voxel R-CNN with different detection head configurations. As shown in Table 5, the cross-stage cross-attention (CSCA) mechanism improves the detection performance by progressively refining object proposals through multi-stage optimization. However, the CSCA primarily focuses on bounding box refinement and provides limited enhancement for the sparse and incomplete structural features of pedestrians and cyclists. To address this issue, the fine-grained grouped convolutional component-aware (FGCA) module is introduced to strengthen fine-grained feature perception during RoI refinement. The results show that the combination of the CSCA and FGCA achieves the best cyclist detection performance across all difficulty levels and attains the highest AP40 for pedestrians under the hard setting. These improvements demonstrate that the proposed detection head effectively enhances component-level feature representation and fine-grained feature learning, leading to more accurate localization and recognition of challenging objects.

Table 5. Ablation Results of Different Detection Head Designs for Pedestrian and Cyclist Detection on the KITTI Validation Set. The bold results indicate the best performance.

Model	Ped.3D (APR40)			Cyc.3D (APR40)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Voxel R-CNN	65.51	59.67	53.73	90.72	70.94	66.88
Voxel R-CNN + CSCA	71.99	64.92	59.18	90.91	74.02	69.56
Voxel R-CNN + CSCA+ FGCA	71.07	64.45	59.45	91.38	75.15	70.57

Effects of different components: To validate the effectiveness of the proposed modules, comprehensive ablation studies were conducted on the KITTI validation set, with the results reported in Table 6. All experiments are built upon Voxel R-CNN and evaluated using AP40. Compared with the baseline, the proposed MSR3DCNN (Mutli-scale Residual 3D CNN) backbone significantly improves detection accuracy. As shown in Group A, the moderate AP40 increases by 1.54%, 6.23%, and 2.94% for cars, pedestrians, and cyclists, respectively, demonstrating the effectiveness of enlarged receptive fields and enhanced feature propagation. To further enhance the BEV feature extraction, a ConvNeXt-based 2D backbone is introduced. As shown in Group E, combining MSR3DCNN with ConvNeXt further improves pedestrian detection performance. However, the gains for cars and cyclists remain limited, indicating that the enhanced BEV features alone cannot be fully exploited without effective RoI refinement. Therefore, the cascaded component-aware RoI detection head (CasA_FGCE) is incorporated to improve the proposal refinement and structural feature learning. As shown in Groups C and D, CasA_FGCE consistently outperforms the baseline. Finally, the full model achieves the best overall performance, demonstrating the complementary advantages and strong synergy of the three proposed modules.

Table 6. Ablation Results on the KITTI Validation Set Using Different Components. The bold results indicate the best performance.

Method	Module			Car 3D (AP ₄₀)			Pedestrian 3D (AP ₄₀)			Cyclist 3D (AP ₄₀)		
	MSR3 DCNN	Conv Next	CasA_FGCE	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Voxel R-CNN				92.24	84.52	82.54	65.51	59.67	53.73	90.72	70.94	66.88
A	√			92.80	86.06	83.64	72.77	65.90	60.88	91.62	73.88	69.49
B		√		93.25	86.44	82.02	72.11	66.33	59.68	91.30	74.23	69.65
C			√	93.02	86.19	83.78	71.07	64.45	59.45	91.38	75.15	70.57
B	√		√	93.07	86.43	83.94	72.34	65.64	59.45	91.23	75.50	71.06
C	√	√		92.97	86.24	83.86	73.13	66.24	59.79	91.74	74.22	69.80
Ours	√	√	√	93.27	86.64	84.16	74.57	67.33	60.78	91.71	76.37	71.93
Improvement				+1.03	+2.12	+1.62	+9.06	+7.66	+7.05	+0.99	+5.43	+5.05

Computational complexity analysis: To assess the computational efficiency of the proposed modules, Table 7 reports the parameter count and inference latency of different model configurations. The baseline MSR3DCNN backbone requires only 6.03 M parameters and 47.20 ms inference time. Adding the ConvNeXt-based BEV backbone introduces a slight increase in the computational cost, while the CasA_FGCE detection head contributes a larger overhead due to cascaded RoI refinement and feature enhancement. Nevertheless, both modules yield notable accuracy improvements. The full model achieves the best overall performance with 11.51 M parameters and 70.20 ms latency, attaining mean AP₄₀ scores across the easy, moderate, and hard settings of 88.02%, 67.56%, and 80.00% for cars, pedestrians, and cyclists, respectively. These results indicate that the proposed framework provides a favorable trade-off between detection accuracy and computational efficiency.

Table 7. Computational Complexity and Inference Latency Analysis of Different Model Configurations on the KITTI Validation Set.√

MSR3DCNN	ConvNeXt	CasA_FGCE	Car.3D	Ped.3D	Cyc.3D	#Params (M)	Runtime (ms)
√			87.50	66.52	78.33	6.03	47.20
	√		87.24	66.04	78.39	5.76	45.10
		√	87.66	64.99	79.03	11.01	64.30
√	√		87.69	66.39	78.58	6.41	50.90
√		√	87.81	65.81	79.26	11.43	69.60
√	√	√	88.02	67.56	80.00	11.51	70.20

5. Conclusions

This work presents LSCA-RCNN, a novel multi-stage voxel-based 3D detection framework for autonomous driving. To mitigate feature degradation and insufficient multi-scale perception in traditional voxel backbones, we embed SRBs and large-kernel spatial convolutions into 3D sparse convolutions, retaining the geometric details and enlarging the receptive fields. Combined with a ConvNeXt 2D backbone equipped with depthwise separable convolutions and spatial attention, the model strengthens the BEV feature representation and improves the detection of small and hard objects. Moreover, the cascaded RoI detection head with grouped convolution and cross-stage cross-attention refines the bounding boxes progressively, reducing localization errors for sparse and occluded targets. Evaluated on the KITTI dataset, LSCA-RCNN outperforms Voxel R-CNN in various scenarios and object classes. The experimental results confirm the efficacy and strong generalization of our method, which shows great robustness for sparse point

clouds, distant objects, and occlusions for practical outdoor 3D detection. Nevertheless, the proposed LSCA-RCNN still has certain limitations, including degraded detection performance in extreme weather scenarios and slightly insufficient real-time inference speed due to increased computational overhead. In future work, we will optimize the network lightweight structure and introduce adaptive feature adjustment mechanisms to improve the model's real-time performance and environmental generalization ability for practical autonomous driving deployment.

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

Funding: This research was funded by the Support Project for Young Key Teachers of Xinyang Normal University (Grant Number 2025GGJS-03)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available in KITTI official website: <https://www.cvlibs.net/datasets/kitti/> (accessed on 15 August 2025).

Conflicts of Interest: Author Xianjun Wu was employed by the companies Xinyang Gumai Optronics Co., Ltd. and Xinyang Central Semiconductor Technology Co., Ltd., the author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LSCA-RCNN	Large-Kernel Spatial Residual and Cascade Attention in Voxel R-CNN
SRBs	Spatial Residual Blocks
3D	Three-Dimensional
2D	Two-Dimensional
SE	Squeeze-and-Excitation
RoI	Region of Interest
BEV	Bird's-Eye View
RPN	Region Proposal Network
VFE	Voxel Feature Encoding
BN	Batch Normalization
FGCE	Fine-Grained Group Convolution Enhancement
MSG-CU	Multi-scale Grouped Convolution Unit
MSR3DCNN	Multi-scale Residual 3D CNN
CasA_FGCE	Cascaded Component-Aware RoI Detection Head
SDCM	Spatial-Dimensional Convolutional Modules
FGCA	Fine-Grained Grouped Convolutional Component-Aware
CSCA	Cross-Stage Cross-Attention

References

1. Qian, R.; Garg, D.; Wang, Y.; You, Y.; Belongie, S.; Hariharan, B.; Campbell, M.; Weinberger, K.Q.; Chao, W.-L. End-to-end pseudo-lidar for image-based 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5881–5890.
2. Wang, Z.; Huang, Z.; Fu, J.; Wang, N.; Liu, S. Object as query: Lifting any 2D object detector to 3D detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 3791–3800.

3. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
4. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, pp. 5015–5114.
5. Shi, S.; Wang, X.; Li, H. Pointcnn: 3D object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 770–779.
6. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3DSSD: Point-based 3D single stage object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11040–11048.
7. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. <https://doi.org/10.3390/s18103337>.
8. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499.
9. 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 AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 1201–1209. <https://doi.org/10.1609/aaai.v35i2.16207>.
10. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10529–10538.
11. 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. *Int. J. Comput. Vis.* **2023**, *131*, 531–551. <https://doi.org/10.1007/s11263-022-01710-9>.
12. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
13. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12697–12705.
14. Song, N.; Jiang, T.; Yao, J. JPV-Net: Joint point-voxel representations for accurate 3D object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; pp. 2271–2279. <https://doi.org/10.1609/aaai.v36i2.20125>.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, pp. 6000–6010.
17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
18. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3D object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11677–11684. <https://doi.org/10.1609/aaai.v34i07.6837>.
19. Gao, X.; Zhao, Y.; Wang, Y.; Shang, J.; Zhang, C.; Wu, G. Bi-att3Ddet: Attention-based bi-directional fusion for multi-modal 3D object detection. *Sensors* **2025**, *25*, 658. <https://doi.org/10.3390/s25030658>.
20. Li, X.; Ma, T.; Hou, Y.; Shi, B.; Yang, Y.; Liu, Y.; Wu, X.; Chen, Q.; Li, Y.; Qiao, Y. Logonet: Towards accurate 3D object detection with local-to-global cross-modal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 17524–17534.
21. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. Largekernel3D: Scaling up kernels in 3D sparse cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13488–13498.
22. Zhang, G.; Chen, J.; Gao, G.; Li, J.; Liu, S.; Hu, X. Safdnet: A simple and effective network for fully sparse 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 14477–14486.
23. Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. Std: Sparse-to-dense 3D object detector for point cloud. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1951–1960.
24. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.

25. Sheng, H.; Cai, S.; Liu, Y.; Deng, B.; Huang, J.; Hua, X.-S.; Zhao, M.-J. Improving 3D object detection with channel-wise transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2743–2752.
26. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029. <https://doi.org/10.48550/arXiv.1603.08029>.
27. Fran, C. Deep learning with depth wise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
28. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6688–6697.
29. 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 Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. <https://doi.org/10.1109/TGRS.2022.3203163>.
30. Chen, C.-F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention Multi scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 357–366.
31. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. <https://doi.org/10.1177/0278364913491297>.
32. Simonelli, A.; Buló, S.R.; Porzi, L.; López-Antequera, M.; Kotschieder, P. Disentangling monocular 3D object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1991–1999.
33. 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 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 409–418. <https://doi.org/10.1145/3664647.3681420>.
34. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5418–5427.
35. Wu, H.; Wen, C.; Shi, S.; Li, X.; Wang, C. Virtual sparse convolution for multimodal 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 21653–21662.
36. Li, B.; Chen, J.; Li, X.; Xu, R.; Li, Q.; Cao, Y.; Wu, J.; Qu, L.; Li, Y.; Diniz, P.S. VFL3D: A single-stage fine-grained lightweight point cloud 3D object detection algorithm based on voxels. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 12034–12048. <https://doi.org/10.1109/TITS.2024.3373227>.
37. Shi, S.; Wang, Z.; Shi, J.; Wang, X.; Li, H. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2647–2664. <https://doi.org/10.1109/TPAMI.2020.2977026>.
38. Zhou, Y.; Onder, O.F.; Dou, Q.; Tsougenis, E.; Chen, H.; Heng, P.-A. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In Proceedings of the International Conference on Information Processing in Medical Imaging, Hong Kong, China, 2–6 June 2019; pp. 682–693. https://doi.org/10.1007/978-3-030-20351-1_53.
39. Lu, Y.; Zhang, Y.; Fan, X.; Cai, D.; Gong, R. Fusing Residual and Cascade Attention Mechanisms in Voxel-RCNN for 3D Object Detection. *Sensors* **2025**, *25*, 5497. <https://doi.org/10.3390/s25175497>.
40. Wu, H.; Wen, C.; Li, W.; Li, X.; Yang, R.; Wang, C. Transformation-equivariant 3D object detection for autonomous driving. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 2795–2802. <https://doi.org/10.1609/aaai.v37i3.25380>.
41. Wang, Z.; Huang, X. A Self-Attention-Enhanced 3D Object Detection Algorithm Based on a Voxel Backbone Network. *World Electr. Veh. J.* **2025**, *16*, 416. <https://doi.org/10.3390/wevj16080416>.
42. Xu, H.; Yang, L.; Zhao, S.; Tao, S.; Tian, X.; Liu, K. SPS-RCNN: Semantic-Guided Proposal Sampling for 3D Object Detection from LiDAR Point Clouds. *Sensors* **2025**, *25*, 1064. <https://doi.org/10.3390/s25041064>.
43. 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), Madrid, Spain, 1–5 October 2018; pp. 1–8. <https://doi.org/10.1109/IROS.2018.8594049>.

44. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
45. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3D object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK 23–28 August 2020; pp. 35–52. https://doi.org/10.1007/978-3-030-58555-6_3.

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.