remote sensing

MDPI

Article

CaLiJD: Camera and LiDAR Joint Contender for 3D Object Detection

Jiahang Lyu [1], Yongze Qi [1], Suilian You [1], Jin Meng [1], Xin Meng [1], Sarath Kodagoda [2] and Shifeng Wang [1,3,*]

[1] School of Optoelectronic Engineering, Changchun University of Science and Technology, Changchun 130022, China; 2023200079@mails.cust.edu.cn (J.L.); 2023100366@mails.cust.edu.cn (Y.Q.); 2022100374@mails.cust.edu.cn (S.Y.); 2022200070@mails.cust.edu.cn (J.M.); 2023200069@mails.cust.edu.cn (X.M.)
[2] Faculty of Engineering & Information Technology, University of Technology Sydney, Sydney, NWS 2007, Australia; sarath.kodagoda@uts.edu.au
[3] Zhongshan Institute of Changchun University of Science and Technology, Zhongshan 528400, China
* Correspondence: sf.wang@cust.edu.cn

Abstract: Three-dimensional object detection has been a key area of research in recent years because of its rich spatial information and superior performance in addressing occlusion issues. However, the performance of 3D object detection still lags significantly behind that of 2D object detection, owing to challenges such as difficulties in feature extraction and a lack of texture information. To address this issue, this study proposes a 3D object detection network, CaLiJD (Camera and Lidar Joint Contender for 3D object Detection), guided by two-dimensional detection results. CaLiJD creatively integrates advanced channel attention mechanisms with a novel bounding-box filtering method to improve detection accuracy, especially for small and occluded objects. Bounding boxes are detected by the 2D and 3D networks for the same object in the same scene as an associated pair. The detection results that satisfy the criteria are then fed into the fusion layer for training. In this study, a novel fusion network is proposed. It consists of numerous convolutions arranged in both sequential and parallel forms and includes a Grouped Channel Attention Module for extracting interactions among multi-channel information. Moreover, a novel bounding-box filtering mechanism was introduced, incorporating the normalized distance from the object to the radar as a filtering criterion within the process. Experiments were conducted using the KITTI 3D object detection benchmark. The results showed that a substantial improvement in mean Average Precision (mAP) was achieved by CaLiJD compared with the baseline single-modal 3D detection model, with an enhancement of 7.54%. Moreover, the improvement achieved by our method surpasses that of other classical fusion networks by an additional 0.82%. In particular, CaLiJD achieved mAP values of 73.04% and 59.86%, respectively, thus demonstrating state-of-the-art performance for challenging small-object detection tasks such as those involving cyclists and pedestrians.

Keywords: deep learning; 3D object detection; data fusion; point cloud data

Citation: Lyu, J.; Qi, Y.; You, S.; Meng, J.; Meng, X.; Kodagoda, S.; Wang, S. CaLiJD: Camera and LiDAR Joint Contender for 3D Object Detection. *Remote Sens.* **2024**, *16*, 4593. https://doi.org/10.3390/rs16234593

Academic Editors: Abdul Awal Md Nurunnabi, Meida Chen, Yan Xia and Felicia Norma Rebecca Teferle

Received: 19 October 2024
Revised: 27 November 2024
Accepted: 28 November 2024
Published: 6 December 2024

1. Introduction

As shown in Figure 1, the significant progress in 3D object detection technology in recent years has been driven by advancements in deep learning. Numerous 3D detection models based on various methods have emerged, consistently advancing the rankings of benchmarks in 3D object detection, such as those provided by the KITTI [1] dataset. Researchers have proposed solutions to address the current shortcomings of 3D object detection from various perspectives. Recent advancements in the field of 3D object detection, including methods based on monocular images, stereo vision, LiDAR, and sensor fusion approaches involving LiDAR and cameras, are detailed in references [2–5]. For instance, Zhou et al. [6] performed optimizations to address the issue of inaccurate orientation prediction in the current 3D object detection methods using monocular cameras. Chong

et al. [7] proposed a method that introduced LiDAR signal distillation during the training phase to address the limitation of using only monocular images for 3D object detection, which often suffers from inadequate spatial cues. In the field of 3D detection based on stereo vision, many outstanding works have been produced in succession. In 2019, You et al. proposed PSEUDO-LIDAR++ [8], which substantially improved the pseudo-LiDAR framework by enhancing stereo depth estimation. The performance of the pseudo-LiDAR system is comparable to that of a 64-line LiDAR, although its cost is only 1/70th. In 2020, Sun et al. [9] introduced Disp-RCNN, an instance disparity estimation network. Experiments on the KITTI dataset demonstrate that even without the LiDAR ground truth during training, Disp-RCNN achieves competitive performance. Subsequently, Liu et al. proposed YOLOStereo3D [10], which applies a baseline consisting of only lightweight stereo fusion features and a network output structure similar to monocular vision detection. Then, it introduces a hierarchical feature-fusion network structure to enhance the obtained stereo fusion features.

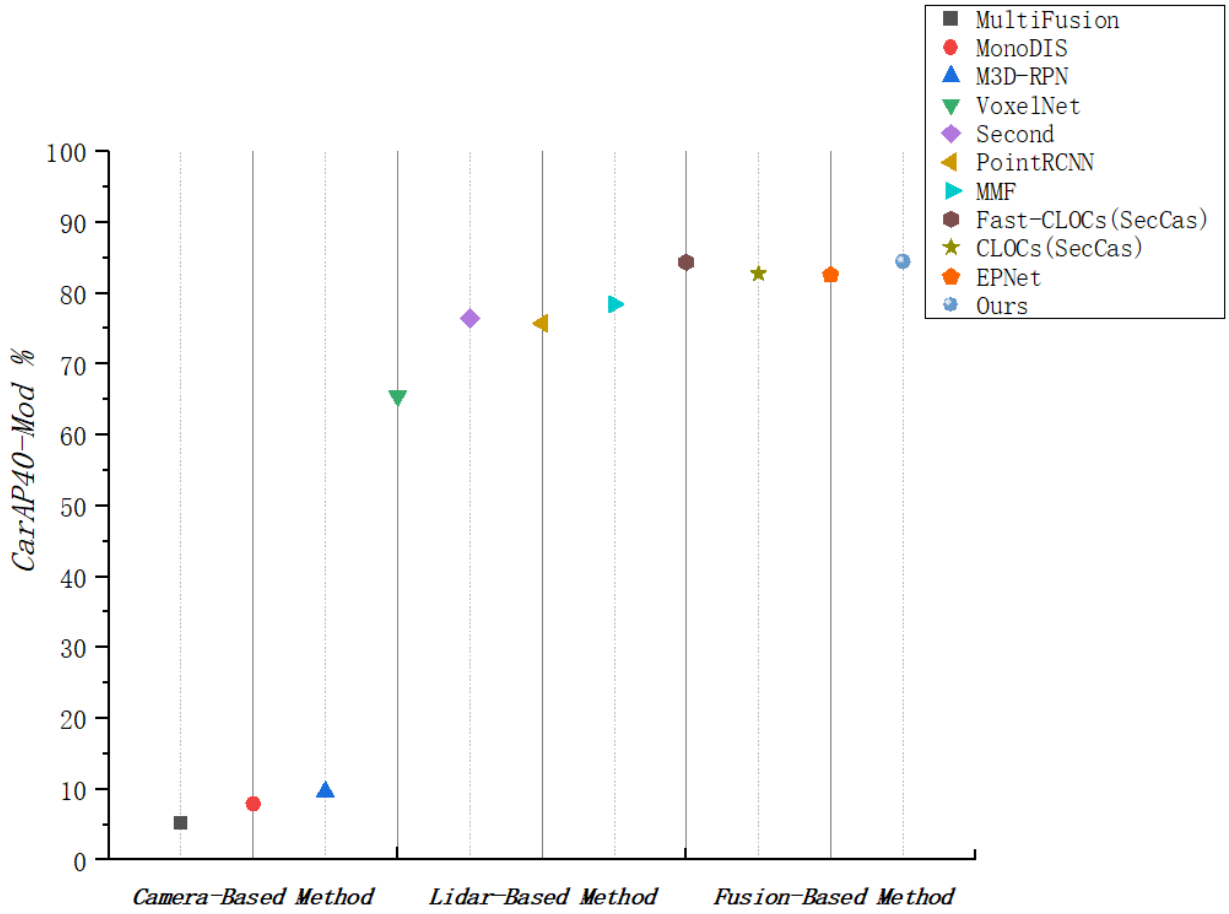

Figure 1. Quantitative Analysis of Camera, LiDAR, Fusion and Proposed Methods. It shows the performance comparison between CaLiJD and other state-of-the-art models on the Kitti dataset, where the horizontal axis represents the different kinds of methods and the vertical axis represents the AP values (Moderate) of the different methods on the kitti dataset.

Existing methods for enhancing depth estimation in 3D detection using stereo vision largely fail to address two issues: the inability to handle dynamic objects and excessive memory overhead. BEVStereos [11] addressed this issue by introducing a dynamic stereo vision approach, which considers the dimensions of the bounding boxes while avoiding the computation of a rotated Intersection over Union (IoU). An improved version, BEVStereo++ [12], was introduced a year later. BEVStereo++ incorporates motion compensation modules and long-sequence frame-fusion techniques, which further enhance

performance and reduce errors. Typically, such approaches focus on the data acquisition phase by employing stereo-vision techniques to obtain three-dimensional information. Owing to the high cost of LiDAR, this remains an unresolved challenge for its applications in the industry. Common approaches in this field include point- and voxel-based methods; however, both are memory-intensive. In 2019, Liu et al. proposed Point-Voxel [13], which addresses this issue by combining two methods using low-frequency voxel and high-frequency point pathways. In 2018, Yan proposed an efficient SpConv implementation based on VoxelNet [14], incorporating data augmentation techniques that significantly improved performance. This paper was entitled SECOND [15]. Subsequently, classic algorithms from 2D detection were drawn upon by [16–18] to construct the entire 3D detection network framework, thereby transforming a single-stage approach into a two-stage approach.

However, as 3D object detection technology continues to advance, researchers have found that inherent limitations are difficult to overcome with single sensors. In object detection involving the fusion of cameras and LiDAR, two primary categories of mainstream methods can be identified: early fusion and late fusion. Early fusion, commonly referred to as data-level fusion, involves the integration of data collected by sensors and relies on the types of sensors that combine similar types of data. In 2017, Chen et al. introduced MV3D [19], which is a multi-view fusion scheme that utilizes different sensors. MV3D introduced innovations in the feature fusion stage by proposing a deep fusion method that enhances detection accuracy through the hierarchical integration of multi-view feature maps. Similarly, the principal innovation of Pointpainting [20] is the fusion of geometric information and point cloud data. The positional relationship between the point cloud and the image is established using the sensor parameters, after which the generated geometric is concatenated with the point cloud data. Considerable success has been achieved with these early fusion methods to some extent. However, the impact of issues such as timestamp synchronization and sensor calibration on the detection results remains non-negligible. Furthermore, information loss arises because this approach does not fully leverage the advantages of both sensors. Consequently, the focus of many researchers has shifted to late-fusion methods for 3D object detection. The multi-sensor fusion technique, BevFusion [21], proposed by Liu et al., integrates data from various sensors into a unified BEV representation. The greatest advantage of this approach is its ability to leverage the strengths of multiple sensors, resulting in improved performance in object detection tasks. Li et al. proposed DeepFusion [22], a modular network architecture designed for the integration of LiDAR, camera, and radar data, to achieve accurate, robust, and long-range 3D object detection. The method employs interchangeable feature extractors, transforming the rich features extracted from each modality into a unified bird's-eye view (BEV) representation, which facilitates convenient fusion within a shared latent space. To address the overfitting issue encountered in the algorithm development for LiDAR and camera fusion detection, Huang et al. [23] constructed a simple prediction pipeline by simulating the data annotation process and utilized a classical fundamental algorithm. The simplest method was then used to train the pipeline to minimize its dependencies and enhance its portability. Moreover, [24–27] proposed novel frameworks from various perspectives, including structure, feature extraction methods, and data selection mechanisms, each achieving outstanding performance. However, the detection accuracy of these methods still requires improvement compared to 2D object detection networks. In particular, all these methods have struggled to effectively balance the detection performance between large and small objects. Therefore, this paper introduces CaLiJD, a novel 3D object detection network based on the detection results of SECOND and Cascade-RCNN [28], which employs a late-fusion approach. The main contributions of this study are summarized as follows:

1. We designed a novel fusion network framework. This fully convolutional network is composed of 1×1 convolutions with different numbers of channels arranged in both serial and parallel configurations. This approach increases the breadth of the fusion network, thereby improving its robustness.
2. We introduce a novel channel attention mechanism known as the Grouped Channel Attention Module (GCAM). This module focuses on the interaction of information between different feature channels and assigns weights to each channel. The traditional 'squeeze' operation in channel attention mechanisms is replaced with different group convolutions, which enhances the correlation between different channels. To utilize GCAM more effectively, it is embedded within a pseudo-feature pyramid module for practical applications.
3. A novel mechanism for selecting bounding boxes is proposed in this study. Based on the normalized distance values, some of the bounding boxes, both near and distant, filtered out in the previous step are reintroduced into the input of the network. Only a single filtering step can lead to missing correctly detected objects. Our method effectively addressed this issue.

2. Related Works

In the context of 3D object detection networks with point cloud and image fusion, it is evident that unimodal 2D object detection and 3D object detection networks are indispensable components. In comparison to 3D object detection, 2D object detection offers several advantages, including high computational efficiency, high robustness, and suitability for real-time applications, as well as compatibility with existing systems. The availability of high-quality 2D image datasets facilitates the training of 2D object detection models. In recent years, a number of high-performance 2D object detection models have been proposed. Fast R-CNN [29] addresses some of the problems, such as low inference speed and low accuracy. In Fast R-CNN, the input image is fed into a convolutional network to generate feature maps and ROI predictions, and then these ROIs are mapped into feature maps for prediction using a pool of ROIs. In contrast to R-CNN, Fast R-CNN does not utilize ROIs as inputs to the CNN layer. Instead, the entire image is processed through the feature maps in order to detect objects. Faster R-CNN [30] employs a comparable approach, although instead of employing a selective search algorithm for the ROI scheme, it utilizes a distinct network that feeds ROIs to the ROI pooling layer and the feature maps. These are then reshaped and employed for prediction. One-stage object detectors, such as YOLO, are more expeditious than two-stage detectors. They are capable of outputting the bounding box directly without the necessity for secondary optimization. In 2023, Ultralytics released YOLOv8 [31], the latest iteration of the YOLO object detection series, which employs advanced neural network architectures (e.g., CSPNet and FPN) to enhance feature extraction. YOLOv8 introduced more efficient training techniques to improve generalization, optimize network and hardware acceleration to improve real-time performance and support multi-task learning. In comparison to the one-stage 2D object detection method, the two-stage object detection method exhibits superior performance, although it can result in increased computational loss. The two-stage network was selected as the 2D backbone network for CaLiJD due to the high data input requirements of the fusion network.

The prerequisite for multi-modal fusion detection is to generate a correspondence between the image information and the point cloud information. It is different from a single two-dimensional detection. For LIDAR or cameras, the scanning of the same object at the same moment is a representation of the object at this time; the only difference is the form of the representation. The links that fuse this information are the absolute coordinates. Thus, the position transformation matrix between the LiDAR and the camera can be applied to obtain a coordinate transformation between the coordinate systems of the two sensors. For the object being scanned, it is also possible to denote its coordinates under both sensors as a link for data fusion. Thus, several limitations that exist when a single sensor performs

tasks such as object detection are mitigated. Recent advancements in deep learning for 3D object detection have led to innovations such as Complex-YOLO [32], which employs Euler-RPN to create RGBBEV images from point cloud data. This approach enables YOLO to generate 3D proposals. Another 3D detector proposed in 2021, RAANet [33], exclusively employs LiDAR data to achieve three-dimensional object detection. It utilizes LiDAR point cloud data in the bird's-eye-view (BEV) as input to the Region Proposal Network, which is subsequently employed to generate shared features. These shared features are used as input to an anchor-free network for detecting 3D objects. Although these methods can achieve 3D object detection to a certain extent, they do not effectively reduce the influence of external factors such as lighting or occlusion on detection results obtained from 2D images. Thus, CaLiJD is proposed in this paper as a late-fusion 3D object detection network that integrates results from both 2D and 3D detection. The key benefit of CaLiJD lies in its ability to utilize the information from 2D bounding boxes to optimize 3D detection.

3. Materials and Methods

In this section, we present a comprehensive overview of the overall structure of the CaLiJD. We built on the foundational concept of classical late-fusion networks. Late fusion is applied to replace the non-maximum suppression (NMS) operation by leveraging the geometric consistency of objects across different dimensions. The NMS operation can be effectively prevented from incorrectly filtering valid candidate boxes using this approach. A discriminative network is trained using a late-fusion framework. The network receives as input the output scores of detections across different dimensions, classifications of each detection, and their spatial descriptions.

Figure 2 illustrates the overall structural framework of CaLiJD. The original image and point cloud data are processed within their respective networks to obtain 2D and 3D detection results. The original image and point cloud data are processed within their respective networks to obtain 2D and 3D detection contenders that are not filtered through NMS operations. After obtaining the detection contenders, n 2D detections and m 3D detections are first encoded into a multidimensional tensor of size $m \times n \times 4$. Tensors with an Intersection over Union (IoU) value greater than zero are selected from this encoded tensor. Subsequently, we re-evaluated pairs of contenders with an IoU score of zero based on distance values that exceed 0.9 or are below 0.4. Selected pairs of contenders were fed into the designed fusion network. Initially, 2D and 3D contenders were fused using 1×1 convolutions with different channels across eight layers. Accordingly, the proposed Adaptive Channel Attention Module (ACAM) was incorporated into the pseudo-pyramid structure and linked after the convolutional layers. It was applied to ensure the integrity of both global and local information within the fused features. Ultimately, the processed tensor was mapped to the probability distribution of the object to be detected using max pooling.

3.1. Construction of Input Tensors

The construction of the input tensors primarily involves two components: data encoding and data selection. They can encode the results of two-dimensional and three-dimensional detection into two contender joint representations, specifically a multi-dimensional tensor of size $m \times n \times 4$. Finally, after filtering, these tensors are input into the fusion network for training.

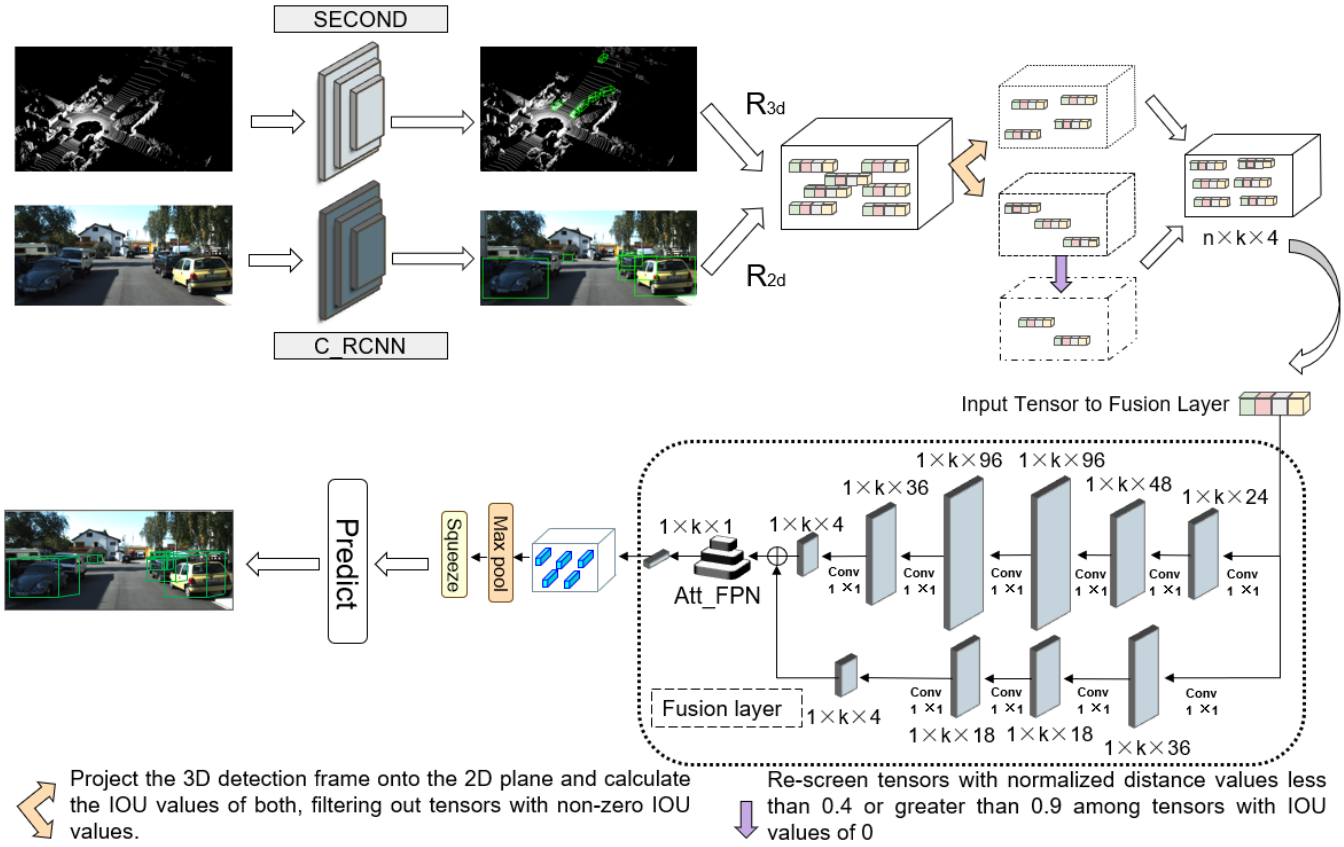

Figure 2. The overall network architecture of CaLiJD. It is mainly divided into a backbone network, a data selection module and a fusion layer. SECOND and C_RCNN represent the 3D and 2D backbone networks of CaLiJD, which are applied to obtain the candidate boxes required for the fusion network. The obtained candidates are filtered and input to the fusion layer for training.

Data Encoding

m 2D detection contenders and n 3D detection contenders are obtained, denoted as R2d and R3d, respectively. The representations of each element are specified in Formulas (1) and (2).

$$R_i2d = [\, d_ix,\ d_iy,\ d_ih,\ d_iw,\ c_i2d\,] \tag{1}$$

$$R_j3d = \left[\, h_j,\ w_j,\ l_j,\ x_j,\ y_j,\ z_j,\ \theta, c_j3d\,\right] \tag{2}$$

In Formula (1), the first four elements represent the offsets of the 2D bounding boxes, and c_i2d indicates the confidence score for the i-th category of 2D detection. In Formula (2), the first seven elements represent the seven-dimensional vectors of the 3D bounding boxes. Similarly, c_j3d indicates the confidence score for the jth category of 3D detection. Thus, a joint contender based on the geometric consistency between the 2D and 3D data is established. The input tensor, denoted by T, is reconstructed in the specific form presented in Formula (3).

$$T_{ij} = \left[IOU_{i,j},\ c_i2d, c_j3d,\ d_j\,\right] \tag{3}$$

where $IOU_{i,j}$ denotes the Intersection over Union between the i-th 2D bounding box and the j-th 3D bounding box after projection onto the plane. c_i2d and c_j3d represent the confidence scores for the i-th 2D detection and j-th 3D detection, respectively. d_j denotes the normalized distance from the j-th 3D bounding box to the LiDAR in the XY plane, as shown in Formula (5).

$$IOU_{i,j} = \frac{\text{Intersection}}{\text{Union}} \tag{4}$$

$$d_j = \frac{\sqrt{\left(x_j - x_l\right)^2 + \left(y_j - y_l\right)^2}}{82} \quad (5)$$

Formula (5) (x_j,y_j) and (x_l,y_l) represent the coordinates of the j-th object and the LiDAR in the XY plane, respectively.

3.2. Data Selection

An excessive number of joint contenders is unfavorable for the subsequent training of the fusion network. Therefore, it is crucial to establish a mechanism for selecting the data to be processed. This step is divided into two distinct phases in CaLiJD. As shown in Figure 3a, contenders with a non-zero IOUi,j from all the joint contenders are first selected as input data for the fusion network.

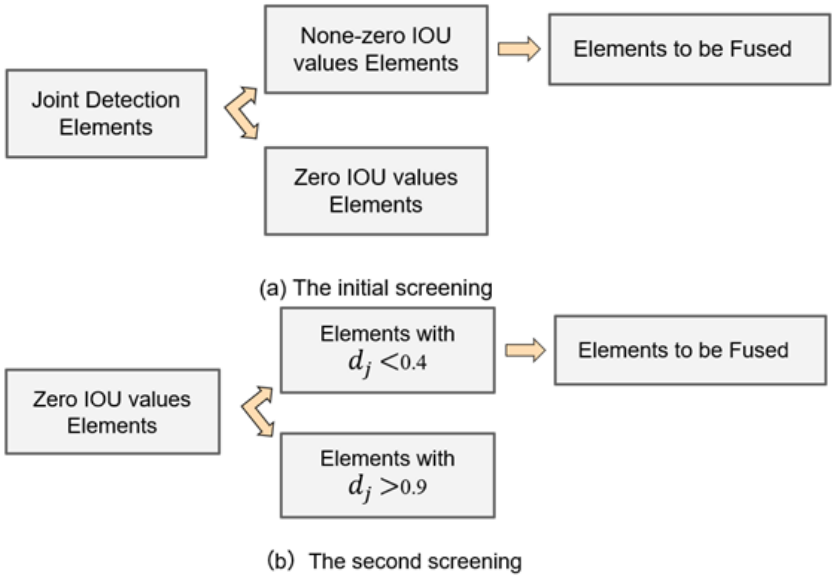

Figure 3. Selection mechanism for CaLiJD. (**a**) demonstrates the data screening method in the traditional late-fusion network, and (**b**) demonstrates the data screening method in CaLiJD.

Contenders with a zero IOUi,j were excluded because of their lack of geometric consistency between the 2D and 3D detection results, which rendered such joint contenders unlikely to have a beneficial effect on subsequent detection. This makes us hesitant to fully discard joint contenders with a zero IOUi,j. Therefore, d_j is applied for the second selection, as illustrated in Figure 3b. Specifically, joint contenders that may miss objects in both 2D and 3D detection are reintegrated into the input of the fusion network. tasks. Occasionally, 3D detection may detect objects that are not detectable in 2D owing to factors such as occlusion and lighting conditions.

3.3. Construction of Fusion Layer

The structure of the fusion network is illustrated in Figure 1, which mainly consists of a feature fusion layer and a pseudo-feature pyramid that includes an adaptive channel attention mechanism.

3.3.1. Fusion Network

The fusion layer is constructed by connecting a 1×1 convolution with different numbers of channels through parallel and sequential arrangements, as shown in Figure 4. The figure specifies the number of input and output channels for each convolutional layer and the corresponding kernel size. An activation function is added after each convolution, except for the last layer.

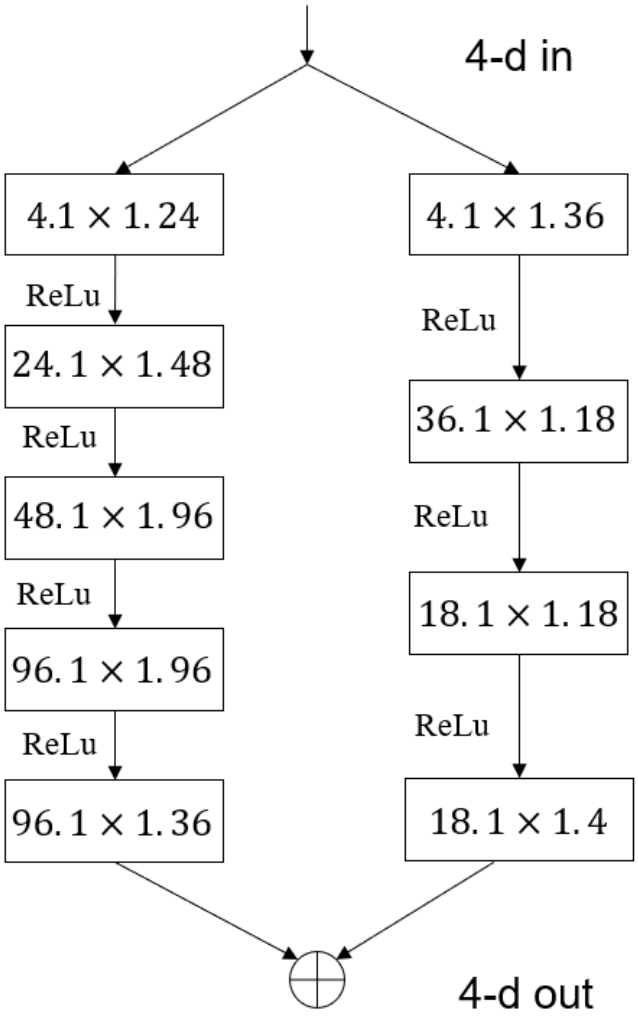

Figure 4. The overall structure of the feature fusion layer.

3.3.2. Grouped Channel Attention Module

The essence of a fusion network is to extract fused features from the constructed joint detection using different convolutional layers. Thus, effectively capturing the interaction information between channels is crucial for enhancing the performance of fusion networks. It is also essential for the network to fuse the 2D and 3D detection results based on geometric consistency. Consequently, a novel attention module was proposed. The structure of the GCAM is shown in Figure 5.

The GCAM aims to capture the interaction information between the channels of the feature map H through different group convolutions and concatenate it along the same dimension. A 1×1 convolution is applied to fully integrate this information, resulting in h′. Subsequently, the 'Extract' operation generates a corresponding weight ww for each channel in h′. Ultimately, the generated weights are multiplied by the feature vectors, resulting in a feature map H that incorporates channel interaction information. The 'Extract' operation is illustrated in Formula (6).

$$w_i = (w_1, w_2, w_3) = Extract(h') = \frac{\exp(h'_i)}{\sum_{j=1}^{n} \exp(h'_i)} \quad (6)$$

where w_i represents the generated weights, with n equal to 3.

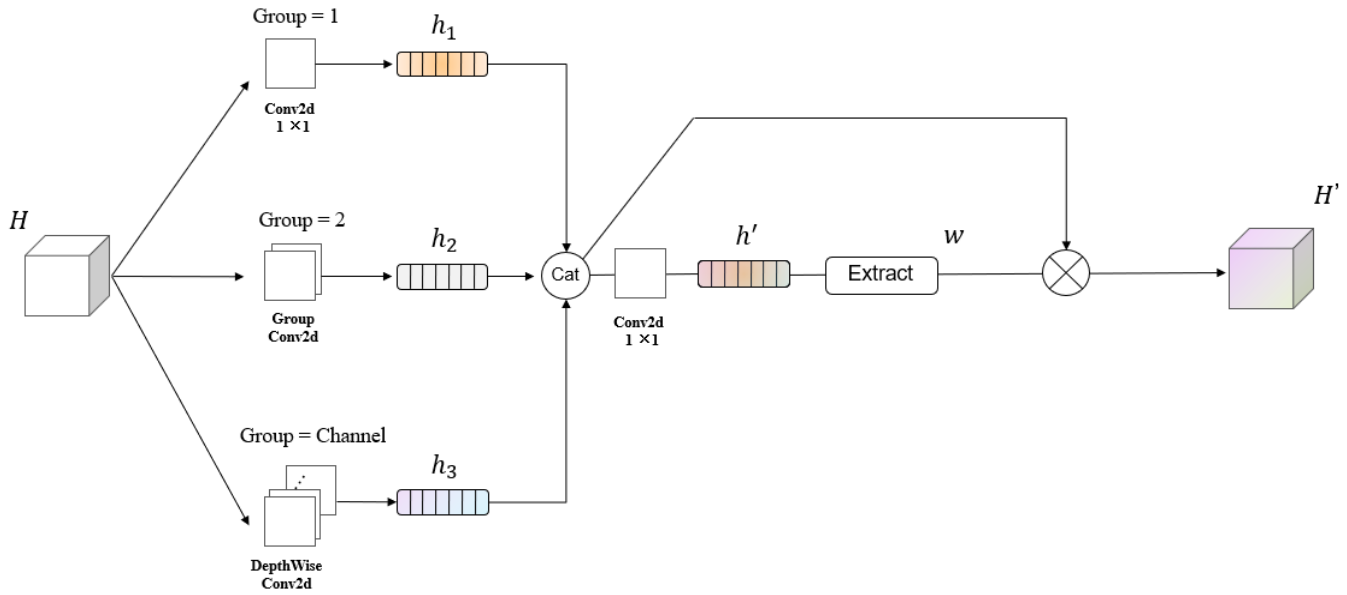

Figure 5. Grouped Channel Attention Module. *H* is the feature map obtained after fusion of features in the fusion layer. *H′* denotes the feature map that has been assigned weights by GCAM. h_1, h_2 and h_3 represent three different features extracted from different grouped convolutions.

To enhance the effectiveness of the GCAM, it is not directly connected to the fusion layer; instead, it is embedded within a pseudo-pyramid structure. As shown in Figure 6, the pseudo-pyramid module differs from the feature pyramid module in that it focuses on the number of channels rather than the size of the input feature map. This enhances the information input to the GCAM at the channel level, making it more efficient.

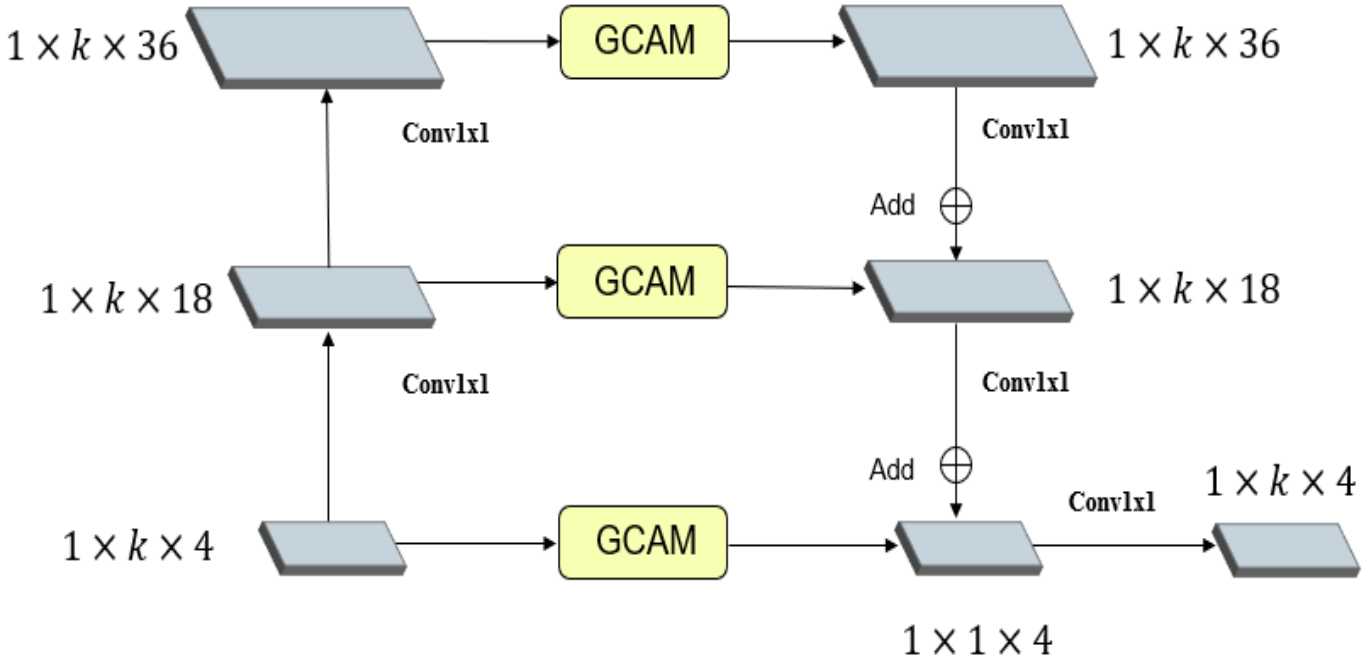

Figure 6. Grouped Channel Attention Module.

As shown in Figure 6, we obtain a feature map with a shape of 1 × k × 4 after the fusion layer. Subsequently, two 1 × 1 convolutions are applied to adjust the feature map shapes to 1 × k × 18 and 1 × k × 36. The three GCAMs process feature maps with three different shapes. Ultimately, the processed feature map is combined with the dimension-reduced feature map to produce an output feature map with the same shape as the input.

4. Results

4.1. Implement

The KITTI dataset was applied to evaluate the performance of CaLiJD. The dataset consists of both LiDAR point cloud data and camera image data, containing 7481 training samples and 7518 testing samples. Ground truth labels are only applicable to the training samples. To evaluate the testing samples, it is necessary to submit the detection results to the KITTI server. In the experiment, the original training samples were divided into 3712 training samples and 3769 validation samples. In this section, our method is compared with the current state-of-the-art single-modal 3D object detection and multi-modal fusion methods. The training and validation processes were performed on an NVIDIA RTX A6000 GPU, supported by an Intel Xeon Gold 6226R CPU with 10 cores. The entire training process consisted of 10 epochs, with each epoch divided into 3700 steps. An Adam optimizer with an initial learning rate of 0.003 was employed during the training process, and the learning rate decayed by a factor of 0.8.

4.2. Comparative Experiments with Single-Modal 3D Object Detection Networks

To evaluate the performance of joint detection, we compared CaLiJD with other state-of-the-art single-modal 3D object detection networks on the car split of the KITTI validation set.

$$mAP = \frac{1}{C} \times \sum_{c \in C} (AP_c) \tag{7}$$

The calculation method for the Mean Average Precision (mAP) values of each category detected at varying difficulty levels is illustrated in Formula (6). As shown in Table 1, compared to the current state-of-the-art single-modal 3D object detection networks, CaLiJD achieved state-of-the-art performance in 3D detection for the easy and moderate split, with Average Precision (AP) values of 93.05% and 84.49%, respectively. SECOND was significantly surpassed as a baseline, with the mAP increasing by 7.54%. Our mAP reached 93.01% for BEV detection, further validating the effectiveness of fusion detection in CaLiJD.

Table 1. Comparative experiments with single-modal 3D object detection networks for the car split on the KITTI validation set (40 recall). Numbers in bold in the table indicate the best results.

Methods	3D AP (%)				Bird's Eye View AP (%)			
	Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP
VoxelNet [14]	81.79	65.46	62.85	70.09	89.60	84.81	78.57	84.33
PointPillars [34]	86.99	77.12	74.98	79.70	89.77	87.02	83.71	86.83
Point-GNN [35]	87.89	78.34	77.38	81.21	89.82	88.31	87.16	88.43
EFNet [36]	86.71	77.25	75.73	79.90	89.66	87.27	85.61	87.51
SECOND [15]	87.43	76.48	69.10	77.67	95.61	89.54	86.96	90.71
3DSSD [37]	88.36	79.57	74.55	80.83	N/A	N/A	N/A	N/A
PV-RCNN [38]	92.10	84.36	**82.48**	**86.44**	N/A	N/A	N/A	N/A
PointRCNN [39]	92.54	82.16	77.88	84.19	95.58	88.78	86.34	90.23
Pointformer [40]	90.05	79.65	78.89	82.86	95.68	90.77	88.46	91.64
CaLiJD (Ours)	**93.05**	**84.49**	78.09	85.21	**96.68**	**93.69**	**89.67**	**93.35**

Figure 7 shows that CaLiJD is more effective than SECOND in reducing false detections. Our method remains more effective even when the objects are located at a considerable distance from LiDAR. This also indirectly demonstrates that our late-fusion network outperforms traditional NMS operations.

Figure 7. Visualization of the KITTI dataset. (**a**,**d**,**g**,**j**) show 2D images from four different scenes. (**b**,**e**,**h**,**k**) present the visualizations of 3D detection using the SECOND for these four scenes. (**b**,**e**,**h**,**k**) present visualizations of 3D detection using SECOND for these four scenes. (**c**,**f**,**i**,**l**) show the visualization results for the CaLiJD. The green 3D bounding boxes represent the detection results, whereas the areas within the yellow circles indicate the erroneous detections.

4.3. Comparative Experiments with Advanced Fusion Networks for 3D Object Detection

Table 2 clearly shows that CaLiJD exhibits highly competitive performance superior to that of most other fusion networks, even fusion detection networks. Notably, compared with CLOCs, another classical late-fusion approach, CaLiJD achieved improvements of 0.70% and 1.76% in AP values on the Easy and Mod splits, respectively. The mAP reached 85.21%—an increase of 0.82%.

Table 2. Comparative experiments with image and point cloud fusion networks for the car split on the KITTI validation set (40 recall). Numbers in bold in the table indicate the best results.

	Methods	3D AP (%)				Bird's Eye View AP (%)			
		Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP
Early Fusion and Deep Fusion	MV3D [19]	71.29	62.68	56.60	63.52	86.55	78.10	76.67	80.44
	MVXNet [41]	85.50	73.30	67.40	75.40	89.50	84.90	79.00	84.47
	F-ConvNet [42]	89.02	78.80	77.09	81.64	90.23	88.79	86.84	88.62
	3D-CVF [43]	88.84	79.72	72.80	80.45	N/A	N/A	N/A	N/A
	F-PointNets [44]	83.76	70.92	63.65	72.78	88.16	84.92	76.44	83.17
	ContFusion [45]	86.32	73.25	67.81	75.79	95.44	87.34	82.43	88.40
	PI-RCNN [46]	88.27	78.53	77.75	81.52	N/A	N/A	N/A	N/A
	MAFF-Net [47]	88.88	79.37	74.68	80.98	93.23	89.31	86.61	89.72
	EPNet [48]	92.28	82.59	**80.14**	85.00	95.51	91.47	**91.16**	92.80
Late Fusion	CLOCs(SC) [49]	92.35	82.73	78.10	84.39	96.34	92.57	89.36	92.76
	CaLiJD (Ours)	**93.05**	**84.49**	78.09	**85.21**	**96.68**	**93.69**	89.67	**93.35**

We performed quantitative experiments on the car category of the KITTI validation set at recall rates of 11 and 40, respectively. The results are shown in Figure 8, where the blue, red, and yellow bars represent SECOND, CLOCs, and CaLiJD, respectively. Regardless of the 3D and BEV detection benchmarks, CaLiJD outperformed the other two networks in terms of mAP values.

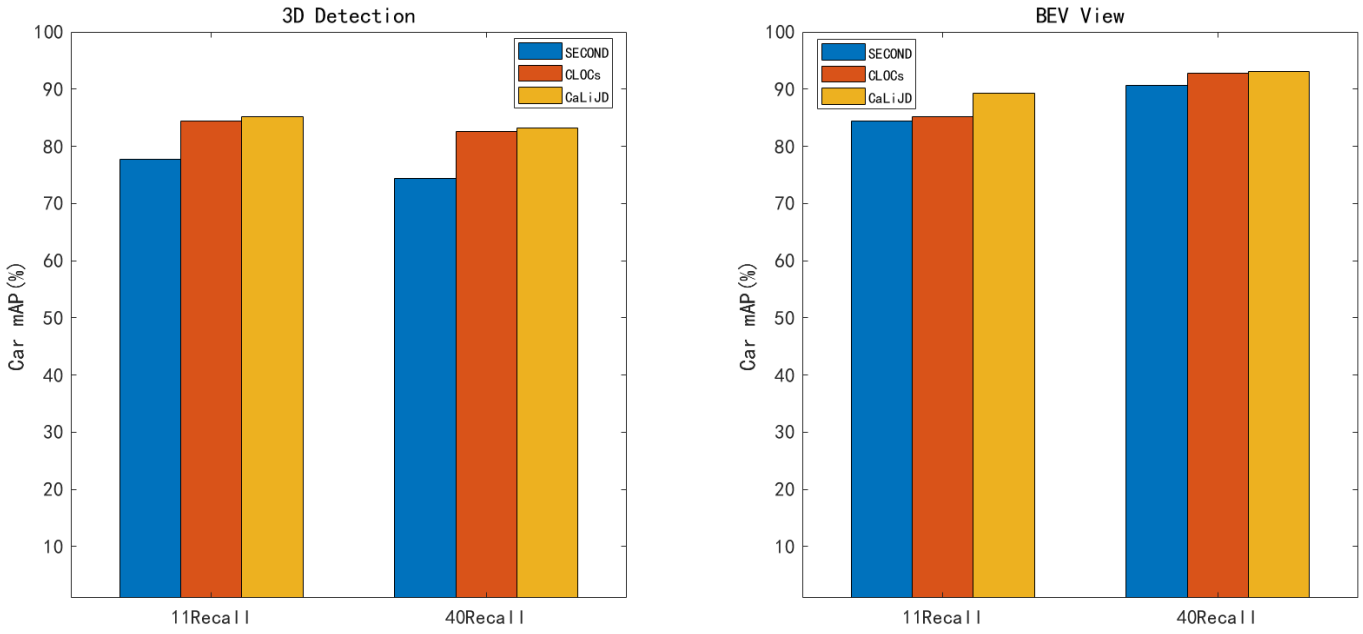

Figure 8. Comparison of mAP values for detection results on the car split. The figure presents a bar chart that visually compares the AP values of three algorithms, including CaLiJD, on the KITTI dataset's car split under recall levels of 11 and 40. Yellow represents CaLiJD, red represents CLOCs, and blue represents SECOND. (**a**) shows the results for 3D object detection, while (**b**) displays the results for BEV detection.

4.4. Detection Experiments on Small Objects

Compared to the detection performance on large objects such as cars, CaLiJD demonstrates superior performance in detecting small objects. Detection experiments were performed on small objects to demonstrate this. This is shown in Table 3.

Table 3. Quantitative analysis of 3D detection for the cyclist and pedestrian splits on the KITTI validation set (40 recall). Numbers in bold in the table indicate the best results.

Methods	Modality	Cyclist				Pedestrian			
		Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP
SECOND [15]	L	78.50	56.74	52.83	62.69	58.01	51.88	47.05	52.31
PointPillars [34]	L	82.31	59.33	55.25	65.63	58.53	51.42	45.20	51.72
STD [50]	L	81.36	67.23	59.35	69.31	60.02	48.72	44.55	51.09
PointRCNN [21]	L	82.56	67.24	60.28	70.06	54.77	46.13	42.84	47.91
IPOD [51]	L and I	78.19	59.40	51.38	62.99	60.88	49.79	45.43	52.03
F-PointNet [13]	L and I	77.26	61.37	53.78	64.14	57.13	49.57	45.48	50.73
AVOD-FPN [52]	L and I	69.39	57.12	51.09	59.20	58.49	50.32	46.98	51.93
F-ConvNet [42]	L and I	84.16	68.88	60.05	71.03	57.04	48.96	44.33	50.11
Painted (PR) [20]	L and I	83.91	**71.54**	**62.97**	72.81	58.70	49.93	46.29	51.64
CLOCs (SM) [49]	L and I	85.47	59.47	55.00	66.65	62.54	56.76	52.26	57.19
CaLiJD (Ours)	L and I	**88.58**	68.07	62.46	**73.04**	**66.22**	**60.05**	**53.31**	**59.86**

From Tables 3 and 4, it is evident that CaLiJD outperforms current state-of-the-art networks in detecting small objects (cyclists, pedestrians), regardless of whether single-modal or camera and LiDAR fusion methods are used. Notably, the mAP value for the pedestrian split improved significantly by 2.67%, whereas the enhancement for the Cyclist split was 0.23%. Relatively, the mAP values for the two categories improved by 4.02% and 1.44%, respectively, in BEV detection.

Table 4. Quantitative analysis of BEV detection for the cyclist and pedestrian splits on the KITTI validation set (40 recall). Numbers in bold in the table indicate the best results.

Methods	Modality	Cyclist				Pedestrian			
		Easy	Mod	Hard	mAP	Easy	Mod	Hard	mAP
SECOND [15]	L	81.91	59.33	55.53	65.59	61.97	56.77	51.27	56.67
PointPillars [34]	L	84.65	67.39	57.28	69.77	66.97	59.45	53.42	59.95
CLOCs (SM) [49]	L and I	88.96	63.40	59.81	70.72	69.35	63.47	58.93	63.92
CaLiJD (Ours)	L and I	**91.31**	**68.81**	**64.11**	**74.74**	**71.48**	**65.21**	**59.39**	**65.36**

4.5. Ablation Study

An ablation study was conducted on the KITTI validation set for the car and cyclist splits to further demonstrate the contribution of each component in CaLiJD to the final results. As shown in Table 5, an ablation study was designed to investigate the contributions of each component in the CaLiJD, including the fusion layer, data selection, the GCAM, and the PFPN, to the overall performance of the network. SECOND served as the baseline to which each component was added for the evaluation. The results demonstrate that the fusion network, trained on both 2D and 3D detection results, provides the greatest contribution to performance improvement, whether for the car or cyclist split. Another highly effective improvement was the incorporation of the GCAM into the fusion layer. It is worth noting that the PFPN module is only effective when applied in conjunction with the GCAM.

Table 5. The contribution of each component in our CaLiJD fusion pipeline. (√ indicates that the network contains this module).

Fusion Layer	Data Selection	GCAM	PFPN	Car AP (%)	Cyclist AP (%)
				76.48	56.74
√				83.56	65.32
√	√			83.70	65.96
√	√	√		84.28	67.43
√	√		√	83.65	65.89
√	√	√	√	84.49	68.07

5. Discussion

CaLiJD, proposed in this study, exhibits superior performance. However, it still has certain shortcomings that are currently difficult to resolve. While a novel paradigm for late-fusion networks has been introduced, multi-class detection is still unable to be accomplished within a single training and inference cycle. We have yet to develop a balanced loss function for multi-object detection. Future iterations of CaLiJD could incorporate a multi-task learning framework to handle multi-class detection more effectively, potentially using a mixed objective function that balances the requirements of various detection tasks. Additionally, as with other classic late-fusion networks, changing a more advanced 2D or 3D baseline could also improve the performance of the network. These aspects can be addressed in the future.

6. Conclusions

CaLiJD is proposed in this study. This is a late-fusion 3D object detection network based on 2D and 3D detection results. A secondary selection mechanism based on normalized distance is proposed within the late-fusion framework. A novel fusion network is designed and trained. Moreover, the proposed PFPN module embedded with a GCAM is incorporated into the fusion layer. It is employed to extract channel-level interactive information from the feature vectors. Experiments on 3D detection and BEV detection were conducted using the KITTI benchmark dataset. The results demonstrate that CaLiJD enhances the performance to some extent, regardless of whether it is applied to single-modal detection or traditional fusion detection networks. Compared to the single-modal baseline, our method achieved a 7.54% improvement in mAP on the car split. Additionally, for the detection of small objects (cyclists and pedestrians), CaLiJD demonstrates enhancements of 0.23% and 2.67% in mAP compared to existing state-of-the-art fusion detection networks.

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

Funding: This research is funded by the International Cooperation Foundation of Jilin Province (20210402074GH), the Public Welfare Science and Technology Foundation of Zhongshan City (2023SYF05), and Innovative Research Team Funding (CXTD2023002).

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the author upon reasonable request.

Acknowledgments: The authors wish to thank the anonymous reviewers and the associated editor for their valuable suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
2. Arnold, E.; Al-Jarrah, Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A survey on 3d object detection methods for autonomous driving applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
3. Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. *IEEE Tran. Intell. Veh.* **2023**, *8*, 3781–3798. [CrossRef]
4. Qian, R.; Lai, X.; Li, X. 3D Object Detection for Autonomous Driving: A Survey. *Pattern Recognit.* **2022**, *130*, 108796. [CrossRef]
5. Wang, Y.; Mao, Q.; Zhu, H.; Deng, J.; Zhang, Y.; Ji, J.; Li, H.; Zhang, Y. Multi-modal 3d object detection in autonomous driving: A survey. *Int. J. Comput. Vis.* **2023**, *131*, 2122–2152. [CrossRef]
6. Zhou, Y.; He, Y.; Zhu, H.; Wang, C.; Li, H.; Jiang, Q. Monocular 3d object detection: An extrinsic parameter free approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 7556–7566.
7. Chong, Z.; Ma, X.; Zhang, H.; Yue, Y.; Li, H.; Wang, Z.; Ouyang, W. Monodistill: Learning spatial features for monocular 3d object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022; pp. 1–17.
8. You, Y.; Wang, Y.; Garg, D.; Pleiss, G.; Hariharan, B.; Campbell, M. Pseudolidar++: Accurate depth for 3d object detection in autonomous driving. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 30 April–4 May 2020; pp. 1–18.
9. Sun, J.; Chen, L.; Xie, Y.; Zhang, S.; Jiang, Q.; Zhou, X.; Bao, H. Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 10548–10557.
10. Liu, Y.; Wang, L.; Liu, M. Yolostereo3d: A step back to 2d for efffcient stereo 3d detection. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13018–13024.
11. Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; Li, Z. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In Proceedings of the Annual AAAI Conference on Association for the Advancement of Artificial Intelligence (AAAI), Hawaii, HI, USA, 7–11 February 2023; pp. 1486–1494.
12. Li, Y.; Yang, J.; Sun, J.; Bao, H.; Ge, Z.; Xiao, L. Bevstereo++: Accurate depth estimation in multi-view 3d object detection via dynamic temporal stereo. *arXiv* **2023**, arXiv:2304.04185.
13. Liu, Z.; Tang, H.; Lin, Y. Point-voxel cnn for efffcient 3d deep learning. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–12 December 2019; pp. 965–975.
14. 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 (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499.
15. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embed-ded convolutional detection. *Sense* **2018**, *18*, 3337–3354. [CrossRef] [PubMed]
16. Shi, S.; Li, H.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X. PV-RCNN++:Point-Voxel Feature Set Abstraction With Local Vector Representation for 3D Object Detection. *arXiv* **2021**, arXiv:2102.00463. [CrossRef]
17. Chen, Y.; Liu, S.; Shen, X.; Jia, J. Fast point r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9775–9784.
18. Li, Z.; Wang, F.; Wang, N. Lidar r-cnn: An efffcient and universal 3d object detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 7546–7555.
19. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
20. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Se-quential fusion for 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 4604–4612.
21. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2774–2781.
22. Li, Y.; Yu, A.; Meng, T.; Ben, C.; Ngiam, J. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 26 December 2022; pp. 560–567.
23. Huang, J.; Ye, Y.; Liang, Z.; Shan, Y.; Du, D. Detecting As Labeling: Rethinking LiDAR-camera Fusion in 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 439–455.
24. Xu, K.; Yang, Z.; Xu, Y.; Feng, L. A Novel Interactive Fusion Method with Images and Point Clouds for 3D Object Detection. *Appl. Sci.* **2019**, *9*, 1065–1074. [CrossRef]
25. Guo, Y.; Hu, H. Multi-Layer Fusion 3D Object Detection via Lidar Point Cloud and Camera Image. *Appl. Sci.* **2024**, *14*, 1348–1364. [CrossRef]
26. Chen, Y.; Lin, Q.; Sun, J.; Feng, Y.; Liu, S. Cascaded Cross-Modality Fusion Network for 3D Object Detection. *Sensors* **2020**, *20*, 7243–7257. [CrossRef] [PubMed]
27. Yang, Q.; Liu, F.; Qu, J.; Jing, H.; Kuang, B.; Chai, W. Multi-sensor fusion of sparse point clouds based on neuralnet works. In Proceedings of the International Conference on Robotics, Intelligent Control and Artificial Intelligence(RICAI), Guilin, China, 3–5 December 2021; pp. 1742–1750.

28. Cai, Z. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef] [PubMed]
29. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
31. Varghese, R.; S, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceeding of the International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6.
32. Simon, M.; Milz, S.; Amende, K. Complex-YOLO: Real-time 3D Object Detection on Point Clouds. *arXiv* **2018**, arXiv:1803.06199.
33. Lu, Y.; Hao, X.; Li, Y.; Chai, W.; Sun, S.; Velipasalar, S. Range-Aware Attention Network for LiDAR-Based 3D Object Detection With Auxiliary Point Density Level Estimation. In *IEEE Transactions on Vehicular Technology*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–14.
34. Lang, A.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beibom, O. PointPillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–21 June 2019; pp. 12697–12705.
35. Shi, W.; Rajkumar, R. Point-GNN: Graph neural network for 3D object detection in a point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 1711–1719.
36. Meng, X.; Zhou, Y.; Du, K.; Ma, J. EFNet: Enhancing feature information for 3D object detection in LiDAR point clouds. *J. Opt. Soc. A.* **2024**, *4*, 739–748. [CrossRef] [PubMed]
37. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3DSSD: Point-Based 3D Single Stage Object Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 11037–11045.
38. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 10529–10538.
39. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–21 June 2019; pp. 770–779.
40. Pan, X.; Xia, Z.; Song, S.; Li, E.; Huang, G. 3D Object Detection with Pointformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 7235–7245.
41. Vishwanath, A.; Zhou, Y.; Oncel, T. MVX-Net: Multimodal VoxelNet for 3D Object Detection. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7276–7282.
42. Wang, Z.; Jia, K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1742–1749.
43. Jin, H.; Kim, Y.; Kim, J.; Choi, W. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, Scotland, 29 September–4 October 2020; pp. 439–455.
44. Qi, C.; Liu, W.; Wu, C.; Su, H.; Leonidas, J. Frustum PointNets for 3D Object Detection from RGB-D Data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 918–927.
45. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep continuous fusion for multi-sensor 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 641–656.
46. Xie, L.; Xiang, C.; Yu, Z. PI-RCNN: An Efficient Multi-sensor 3D Object Detector with Point-based Attentive Cont-conv Fusion Module. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
47. Zhang, Z.; Shen, Y.; Li, H.; Zhao, X.; Yang, M.; Tan, W.; Pu, S.; Mao, H. Maff-net: Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion. In Proceedings of the IEEE International Conference on Intelligent Transportation(ICITS) Systems, Xiamen, China, 24–27 June 2022; pp. 369–376.
48. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 3735–3752.
49. Pang, S.; Morris, D.; Radha, H. Clocs: Camera-lidar object candidates fusion for 3d object detection. In Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020; pp. 10386–10393.
50. 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(ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1951–1960.
51. Yang, Z.; Sun, Y.; Shu, L.; Shen, X.; Jia, J. IPOD: Intensive Point-based Object Detector for Point Cloud. *arXiv* **2018**, arXiv:1812.05276.
52. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S. Joint 3D proposal generation and object detection from view aggregation. In Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8.