symmetry

MDPI

Article

PAFNet: Pillar Attention Fusion Network for Vehicle–Infrastructure Cooperative Target Detection Using LiDAR

Luyang Wang [1,2], Jinhui Lan [1,2,*] and Min Li [1,2]

[1] Department of Instrument Science and Technology, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China; wangluyang@xs.ustb.edu.cn (L.W.); d202110335@xs.ustb.edu.cn (M.L.)
[2] Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
* Correspondence: lanjh@ustb.edu.cn

Abstract: With the development of autonomous driving, consensus is gradually forming around vehicle–infrastructure cooperative (VIC) autonomous driving. The VIC environment-sensing system uses roadside sensors in collaboration with automotive sensors to capture traffic target information symmetrically from both the roadside and the vehicle, thus extending the perception capabilities of autonomous driving vehicles. However, the current target detection accuracy for feature fusion based on roadside LiDAR and automotive LiDAR is relatively low, making it difficult to satisfy the sensing requirements of autonomous vehicles. This paper proposes PAFNet, a VIC pillar attention fusion network for target detection, aimed at improving LiDAR target detection accuracy under feature fusion. The proposed spatial and temporal cooperative fusion preprocessing method ensures the accuracy of the fused features through frame matching and coordinate transformation of the point cloud. In addition, this paper introduces the first anchor-free method for 3D target detection for VIC feature fusion, using a centroid-based approach for target detection. In the feature fusion stage, we propose the grid attention feature fusion method. This method uses the spatial feature attention mechanism to fuse the roadside and vehicle-side features. The experiment on the DAIR-V2X-C dataset shows that PAFNet achieved a 6.92% higher detection accuracy in 3D target detection than FFNet in urban scenes.

Keywords: vehicle–infrastructure cooperative; LiDAR; target detection; feature fusion

Citation: Wang, L.; Lan, J.; Li, M. PAFNet: Pillar Attention Fusion Network for Vehicle–Infrastructure Cooperative Target Detection Using LiDAR. *Symmetry* **2024**, *16*, 401. https://doi.org/10.3390/sym16040401

Academic Editor: Shangce Gao

Received: 7 March 2024
Revised: 27 March 2024
Accepted: 27 March 2024
Published: 29 March 2024

1. Introduction

LiDAR is an active sensor that detects the position of a target by emitting a laser beam and receiving an echo laser [1]. LiDAR has been gradually used in autonomous vehicles [2] because of its advantages of high-range accuracy and because it is not affected by ambient light. With the development of deep learning technology and convolutional neural networks, autonomous vehicles have begun utilizing deep learning for target detection [3,4] of 3D point clouds collected by LiDAR. Unlike 2D images, 3D point clouds are disorderly [5]. PointNet [6] introduced a pioneering method for 3D target detection using a point-based approach. This method uses the T-Net network to learn the spatial transformations of the point cloud so that the point cloud is not affected by the point cloud disorder in the network. Pointnet++ [7] enhances local features by introducing sampling and grouping operations. The PointRCNN [8] network has a high detection accuracy for vehicle targets. PointNeXt [9] and PointMLP [10] use residual structure networks to enhance the overall performance of the model. These methods offer a high detection accuracy. However, point-based methods [11] suffer from slower inferences and require a large number of model parameters. VoxelNet [12] was inspired by 2D images and solves the problem of point

cloud disorder by voxelizing the point cloud. This method converts irregular point clouds into 3D voxels, reducing data dimensions and computation. As a result, the inference speed of 3D point cloud targets is improved. Subsequently, various 3D point cloud target detection methods such as SECOND [13], PointPillars [14], Voxel-FPN [15], PartA2 [16], PillarNet [17], etc. are developed to achieve better detection accuracy in detecting traffic targets such as vehicles, pedestrians, and cyclists. CenterPoint [18,19] is an anchor-free voxel-based detection method for detecting targets in 3D point clouds. It achieves a balance between inference speed and detection accuracy by finding the center point in the heatmap and returning bounding boxes of the target, and has become the most popular method for detecting traffic targets in 3D point clouds. VoxelNeXt [20] and PillarNeXt [21] are two methods proposed on this basis that also use anchor-free detection heads to achieve high detection accuracy.

As the demand for increased sensing range and accuracy in autonomous vehicles grows, vehicle-mounted sensors have limitations due to installation location and cannot achieve long-distance target detection with high accuracy. Vehicle-mounted LiDAR [22] has limitations in detecting traffic information. The point cloud density becomes sparse as the distance between the target and LiDAR increases due to the angular resolution of LiDAR. At long distances, there are only a few points of the vehicle collected by LiDAR, and the sparse point cloud gives a lack of target features and reduces the detection accuracy of the target. At the same time, due to the mutual occlusion between multiple vehicles and environmental objects, the point cloud data collected by LiDAR has limitations. Especially in urban roads with complex traffic environments, the close proximity of vehicles will occlude the LiDAR detection of the rear [23], forming a blind zone behind the vehicles, making distant vehicles undetectable.

The performance of autonomous vehicles is extended by the advent of cooperative autonomous driving. LiDAR sensors installed on roadside traffic infrastructure [24] supplement the surrounding environmental information collected by autonomous vehicles on the road, enabling the perception of the surrounding traffic environment. Due to cost constraints, LiDAR units mounted on vehicles typically have fewer channels, often below 64 lines, and a detection range of generally less than 150 m. A vehicle-mounted LiDAR with 32 lines can capture 600,000 points per second at a frame rate of 20 Hz. In contrast, roadside LiDAR, as a type of road infrastructure, has channel counts of 128 or even higher. Having a higher channel number allows roadside LiDAR to achieve higher resolution, enabling it to detect objects beyond 150 m. A roadside LiDAR with 128 lines can capture 2,400,000 points per second at a frame rate of 20 Hz. The data collected by roadside LiDAR and vehicle-side LiDAR can be symmetrically fused with each other, enhancing the perception capabilities of autonomous vehicles. The fusion of point cloud data from both the vehicle and road with spatial diversity can greatly enhance the target detection accuracy for autonomous vehicles [25]. When it comes to vehicle–infrastructure cooperative (VIC) [26] traffic sensing, there are three schemes of fusion target detection methods for roadside LiDAR and automotive LiDAR data, depending on the fusion data: early fusion [27], feature fusion [28], and late fusion [29].

Early fusion is a method of fusing raw LiDAR point cloud data. This involves transmitting the point cloud data collected by roadside LiDAR directly to the vehicle and fusing it with each frame of the point cloud collected by the automotive LiDAR for target detection. The benefit of early fusion is that it retains all point cloud information, resulting in the highest target detection accuracy. However, transmitting and sharing large amounts of raw point cloud data streams quickly and accurately is a significant challenge [30] due to limited bandwidth. This is especially difficult when multiple autonomous vehicles need to share raw point cloud information with the same roadside LiDAR, where the communication bandwidth cannot meet the demand of the vehicles.

Late fusion is a method of fusing target detection results [31] of point clouds collected by roadside LiDAR and automotive LiDAR. It is used to detect the target of the point cloud collected by roadside LiDAR, and transmit the results to the vehicle side [32,33]. These

results are then fused with the detection results of the automotive LiDAR. Yu proposes the Time Compensation Late Fusion (TCLF) [26] method for VIC 3D object detection. This method transmits target detection results between roadside and vehicle side, minimizing the amount of data and solving the problem of data transfer between the vehicle and infrastructure. However, late fusion is limited by the detection capability of a single sensor. The target detection process uses only individual LiDAR data, which cannot take advantage of cooperative sensing, resulting in generally low target detection accuracy.

Feature fusion, also called middle fusion, is the fusion of features after feature encoding of each LiDAR point cloud, and target detection is performed based on fused features [34]. To ensure efficient transmission between the roadside and vehicles, middle fusion can effectively utilize information from both sides to achieve better detection accuracy. The balance between communication bandwidth and target detection accuracy is realized, which is suitable for vehicle-to-road target detection fusion. The point cloud data collected by automotive LiDAR and roadside LiDAR are first sampled and feature encoded by voxel or pillar grid through the backbone network to form a feature map. This process achieves feature extraction at both the roadside and the vehicle side. The feature maps from both the roadside and vehicle side are combined to create a fusion feature map, which is then used for target detection. Bai proposed the PillarGrid [34] model for VIC target detection, which uses grid-wise feature fusion (GFF) to fuse features and anchor-based detection heads for target detection. Experiments on the Carla simulation dataset demonstrate that this method provides better detection results than baseline PointPillars on both the roadside and vehicle sides. FFNet [35] uses PointPillars as a feature extraction network to extract roadside and vehicle-side Bird's Eye View (BEV) features, generating a feature stream with feature prediction capability using a flow-based method, and using an anchor-based detection head for target detection. It performs target detection by feature compression, transmission, and feature decompression to reduce the cost of data asynchronous transmission in cooperative target detection. In summary, there are currently few feature fusion methods for VIC target detection. Table 1 shows the details of 3D target detection methods. All current methods use pillars to extract point cloud features from vehicle sides and roadsides. Feature fusion can be achieved by either concatenating simple features or selecting the maximum value of a feature. Anchor-based detection heads are the most commonly used target detection methods.

Table 1. Details of 3D target detection methods using LiDAR.

	Method	Anchor	Dataset	LiDAR Type
Point-based	PointNet [6]/Pointnet++ [7]	√	Vehicle-side Dataset	Mechanical Scanning [36]
	PointRCNN [8]	√		
	PointNeXt [9]	Anchor-free		
Voxel-based	VoxelNet [12]/VoxelNeXt [20]	√		
	Voxel-FPN [15]	√		
	PartA2 [16]	√		
	SECOND [13]	√		
	CenterPoint [18,19]	Anchor-free		
Pillar-based	PointPillars [14]/PillarNet [17]	√		
	PillarNeXt [21]	Anchor-free		
	PillarGrid [34]	√	VIC, Carla Simulation [37]	Simulation mechanical scanning [38]
	FFNet [35]	√	VIC	V: Mechanical Scanning, I: Solid-State [36]

In order to improve the accuracy of feature fusion detection methods, we propose Pillar Attention Fusion Network (PAFNet) for VIC target detection. The contributions of this work are as follows:

1. A novel anchor-based VIC feature fusion target detection network is proposed in this paper. The proposed network combines the advantages of a center point detection scheme, improves the accuracy of target detection through feature fusion, and provides a new solution for VIC autonomous driving target detection.
2. A method for preprocessing point cloud features with spatial–temporal coordination is proposed. The accuracy of feature fusion can be improved by optimizing the matching of roadside point cloud frames and vehicle-side point cloud frames. It is also enhanced by unifying the vehicle-side and roadside point cloud coordinate systems to the world coordinate system.
3. A method for fusing VIC features based on spatial attention, called Grid Attention Feature Fusion (GAFF), is proposed. The information contained in roadside feature maps and vehicle-side feature maps is maximally preserved through feature extraction and feature fusion using a spatial attention model.

The article is organized as follows: Section 2 describes the VIC 3D target detection dataset and presents the PAFNet proposed in this paper. Section 3 shows the visualization of the experimental results and the target detection results compared with other methods. Section 4 discusses PAFNet and the experimental results. Section 5 gives conclusions and future research directions.

2. Materials and Methods

2.1. Dataset

We prepared a VIC autonomous driving dataset for the analysis and experiments on the proposed PAFNet network. The DAIR-V2X-C dataset [26] is the world's first large-scale, multi-modal, multi-view 3D target detection dataset for VIC autonomous driving research. It includes 38,845 frames of image data and 38,845 frames of point cloud data. The DAIR-V2X-C dataset is based on the Beijing High-Level Autonomous Driving Demonstration Area, which uses cameras and roadside LiDAR at intersections in complex traffic scenarios. Figure 1 shows two representative samples from the dataset. Autonomous vehicles in the area are equipped with cameras and LiDAR. When an autonomous vehicle drives past roadside equipment, the roadside and vehicle-side sensors record data for that period of time. The data is synchronized using vehicle and roadside GPS timing. The roadside LiDAR has 300 lines and a sampling frame rate of 10 Hz. The vehicle-side LiDAR is a Hesai Pandar 40-line LiDAR with a sampling frame rate of 10 Hz. The dataset contains information on the type, 3D bounding box, occlusion, and truncation of 10 types of targets, including vehicles, pedestrians, and cyclists. Calibration data and coordinate conversion parameters for roadside LiDAR and automotive LiDAR are also provided. Table 2 shows the parameters of the DAIR-V2X-C dataset. This dataset enables the implementation of the VIC 3D target detection task.

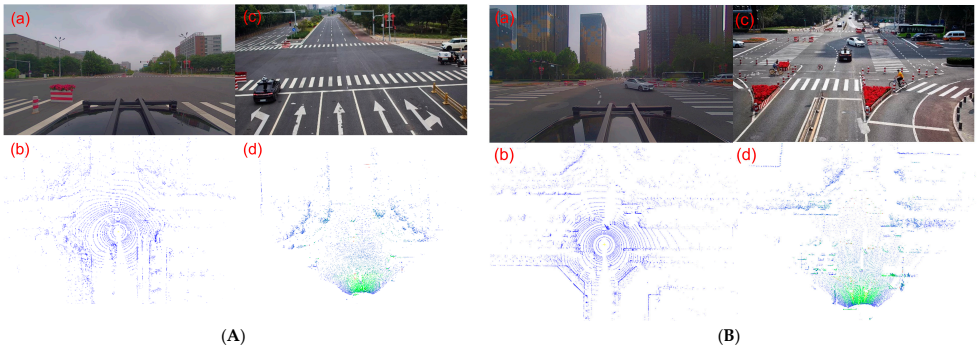

Figure 1. Representative samples of DAIR-V2X-C dataset. (**A**) Visualization of scenario 1 at an intersection. (**B**) Visualization of scenario 2 at an intersection. (**a**) Image collected by the vehicle-side camera. (**b**) Point cloud collected by the vehicle-side LiDAR. (**c**) Image collected by the roadside camera. (**d**) Point cloud collected by the roadside LiDAR.

Table 2. The characteristics of the DAIR-V2X-C dataset.

Information of DAIR-V2X-C [26]			Parameters
Roadside (RS) equipment	LiDAR	Laser lines	300 lines
		Capture frequency	10 Hz
		Horizontal/Vertical FOV	100°/40°
		Detection distance accuracy [36]	≤3 cm
	Camera	Resolution	1920 × 1080
		Capture frequency	25 Hz
Vehicle-side (VS) equipment	LiDAR	Laser lines	40 lines
		Capture frequency	10 Hz
		Horizontal/Vertical FOV	360°/40°
		Detection distance accuracy [36]	≤3 cm
	Camera	Resolution	1920 × 1080
		Capture frequency	25 Hz
Dataset annotation [26]		Frames	38,845
		Object types	10 types (car, truck, van, bus, pedestrian, cyclist, tricyclist, motorcyclist, cart, traffic cone)
		2D box in image	Height, width
		3D box in point cloud	Height, width, length, location, rotation
		Calibration data	Extrinsic parameter matrix of RS and VS
		Coordinate conversion	LiDAR and camera coordinate of RS and VS, virtual world coordinate
		Other information	Time stamp, occluded state, truncated state

2.2. Method

In this paper, we propose a VIC feature fusion anchor-free traffic target detection method called PAFNet to enhance the target detection capability of autonomous vehicles. We propose a preprocessing method for spatial and temporal cooperative fusion. This method aligns the roadside LiDAR point cloud with the vehicle-side LiDAR point cloud to avoid the frame-to-frame and spatial differences of LiDAR from affecting the fusion results. A feature map fusion method based on spatial attention is proposed to solve the problem of information loss in feature fusion. For the problem of different densities of multiple source point clouds collected by LiDAR, we use the atrous spatial pyramid pooling (ASPP) module in the network to extend the receptive field. For the collaboratively fused feature maps, we use a center-based anchor-free detection head for vehicle, pedestrian, and cyclist target detection. The structure of the PAFNet is shown in Figure 2.

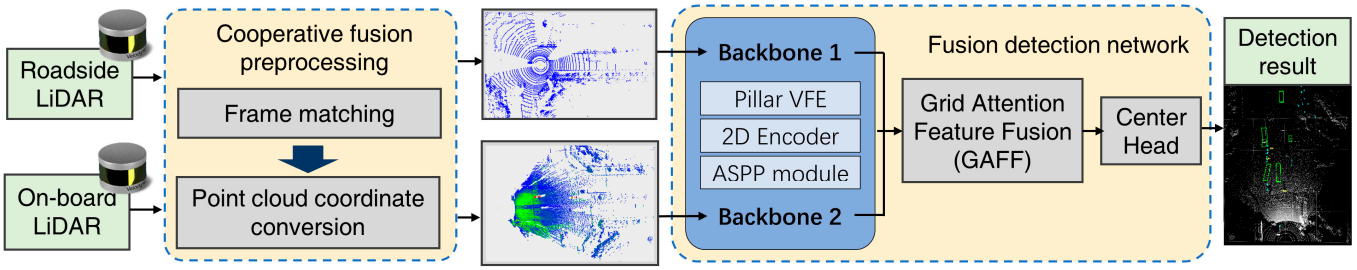

Figure 2. PAFNet structure.

2.2.1. Spatial and Temporal Cooperative Fusion Preprocessing

In practical VIC autonomous driving systems, roadside LiDAR is installed on roadside infrastructure, such as street light poles or traffic lights [39], and automotive LiDAR is mainly installed on the top of the vehicle. The models, frame rates, and point cloud densities of roadside LiDAR and automotive LiDAR often differ due to installation conditions. Therefore, it is necessary to perform frame matching between the roadside and vehicle point cloud data. We propose a preprocessing technique called Spatial and Temporal Cooperative Fusion Preprocessing (STCFP) to ensure consistency of frames and coordinates in VIC data. The formula for frame matching is as follows:

$$Frame_w^i(n) = Frame_i(j|\Delta t < \delta), \tag{1}$$

$$Frame_w^v(n) = Frame_v(k|\Delta t < \delta), \tag{2}$$

where $Frame_w^i(n)$ and $Frame_w^v(n)$ are the n-th frame after matching the roadside point cloud and vehicle-side point cloud, respectively. $Frame_i(j)$ and $Frame_v(k)$ denote the j-th and k-th frames at the roadside and vehicle side, respectively. $\Delta t = t_i(j) - t_v(k)$ is the time between the j-th frame of the roadside and the k-th frame of the vehicle side, and δ is a parameter. When Δt is minimized and less than the parameter δ, the matching between the j-th frame of the roadside and the k-th frame of the vehicle side is completed.

To facilitate feature fusion, the vehicle-side and roadside point clouds need to be unified from their respective sensor coordinate systems into a world coordinate system [40]. The position of an autonomous vehicle in motion relative to roadside LiDAR varies over time. The spatial coordinate system is transformed for each frame of the point cloud using the position information of both the vehicle-mounted LiDAR and roadside LiDAR. The raw point cloud obtained using LiDAR is represented as follows:

$$P = \left\{ [x_i, y_i, z_i, r_i]^T \middle| [x_i, y_i, z_i]^T \in \mathbb{R}^3 \right\}_{i=1,2,3\ldots,T}, r_i \in (0,1), \tag{3}$$

where $[x_i, y_i, z_i, r_i]$ is a single point p, $[x_i, y_i, z_i]$ is the position information of p, and r_i is the reflectivity of p, assuming P_v represents the vehicle-side point cloud data and P_i represents the roadside point cloud data. The translation vector from the sensor coordinate system to the world coordinate system is $T = [X, Y, Z]$, and the rotation vector is $\Gamma = [\Phi, \Theta, Y]$. The equation for transforming the point cloud from LiDAR sensor coordinates to the world coordinate system is:

$$P_w = \begin{bmatrix} M_Z & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} M_Y & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} M_X & 0 \\ 0 & 1 \end{bmatrix} \cdot P + \begin{bmatrix} T & 0 \end{bmatrix}^T, \tag{4}$$

where P_w is the point cloud transformed to the world coordinate system, P is the original point cloud, $\begin{bmatrix} T & 0 \end{bmatrix}^T$ is the translation matrix, M_Z, M_Y, and M_X are the rotation matrices along the z-axis, y-axis, and x-axis, respectively. The formulae for M_Z, M_Y, and M_X are as follows:

$$M_Z = \begin{bmatrix} cos(Y) & -sin(Y) & 0 \\ sin(Y) & cos(Y) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5}$$

$$M_Y = \begin{bmatrix} cos(\Theta) & 0 & sin(\Theta) \\ 0 & 1 & 0 \\ -sin(\Theta) & 0 & cos(\Theta) \end{bmatrix}, \tag{6}$$

$$M_X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\Phi) & -sin(\Phi) \\ 0 & sin(\Phi) & cos(\Phi) \end{bmatrix}. \tag{7}$$

After matching the frames, the point cloud data for the roadside and vehicle side are computed using Equation (4), respectively, and then converted to the world coordinate

system. This completes the fusion preprocessing of the spatial and temporal cooperation between the roadside and the vehicle side, as shown in Figure 3.

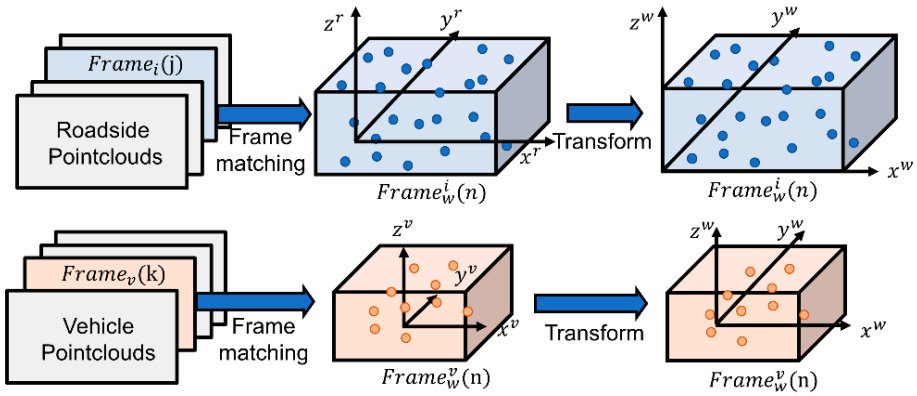

Figure 3. Spatial and temporal cooperative fusion preprocessing.

2.2.2. Feature Extraction Network

We use the Pillar Voxel Feature Encoding (Pillar VFE) [14] for 3D feature encoding. The point cloud P_w is divided into $K_m \times T_m$ pillars, where K_m and T_m represent the number of pillars in the x-axis and y-axis, respectively, and m represents the size of the pillar. All point clouds are aggregated in pillars, and an empty pillar is considered when there are no points in the pillar [41]. Each pillar is sampled at N points, with the coordinates, reflectance, and center position of each point encoded in D-dimensional vectors. The point cloud is represented by a tensor size of (P, N, D). The Set Abstraction module is used to extract features, and a multi-layer MLP is employed to transform the dimension of each point from D to C. The point cloud is converted from a tensor size of (P, N, D) to a point cloud features size of (P, C). The point cloud features are then expanded according to the pillar index to form a feature map size of (C, H, W), as shown in Figure 4. The pillar features contain the positional and local information of the point cloud in the voxel. As certain voxels may not contain point clouds, the features obtained by the pillar VFE network may contain empty features [42].

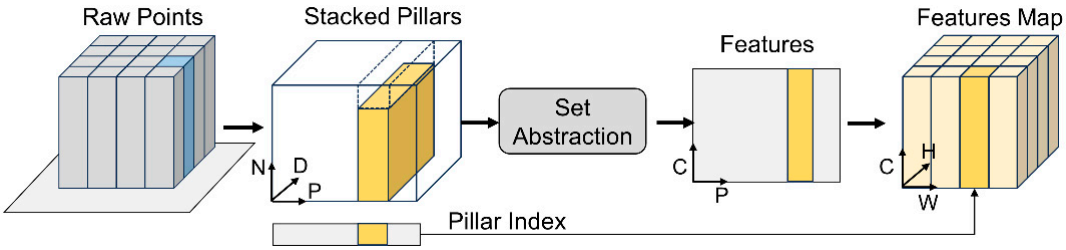

Figure 4. Pillar Voxel Feature Encoding.

To improve the feature extraction, PAFNet utilizes a four-layer residual network to create a straightforward 2D feature encoding network. This method has been shown to have a good feature extraction performance in CenterPoint and PillarNet. It is important to note that roadside LiDAR and automotive LiDAR collect point cloud densities and field of view ranges that differ from each other. Therefore, it is necessary to have a good extraction capability for both roadside and vehicle-side features in the fusion network. Feature pyramid networks (FPN) [43,44] and Bidirectional Feature Pyramid Network (BiFPN) [45,46] networks improve feature acquisition by capturing object and contextual information at different scales of an image through multi-scale feature representation. In contrast, 3D point cloud targets exhibit no changes in spatial scale, but only changes in the sparsity of the target point cloud. VoxelNeXt [12] showed that the edge features of the target point cloud have a significant impact on target detection. Even the ambient

point cloud around the target point cloud contributes to the detection of small targets, such as pedestrians. Therefore, expanding the sensory field is an effective way to improve the feature extraction capability. The issue of extracting the vehicle-side and roadside features in a single network is addressed by utilizing multi-scale features. The PAFNet network incorporates an atrous spatial pyramid pooling (ASPP) [47,48] module in its neck to combine BEV features, expand the receptive field, and fuse multi-scale contextual information. The ASPP module allows for the acquisition of receptive fields of different scales and the extraction of multi-scale information by controlling the expansion rate of the dilated convolution and adjusting the padding and dilation. It consists of three dilated convolution branches with different dilation rates, and the output feature maps of these branches are cascaded and fused by pooling for free multi-scale feature extraction [49], as shown in Figure 5.

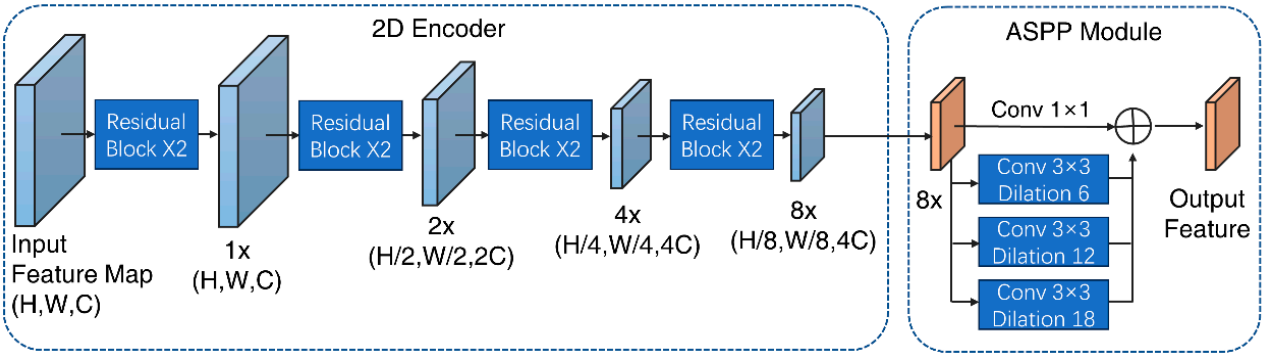

Figure 5. 2D feature extraction module.

2.2.3. Grid Attention Feature Fusion

Feature fusion is utilized to combine the BEV feature maps of the vehicle-side point cloud with the BEV feature maps of the roadside point cloud to create a feature map. This is a crucial component of PAFNet. The most straightforward method of feature fusion involves concatenating the two feature maps in the channel dimension to create a feature map with more channels. However, this can lead to an increase in parameters, which may introduce redundant information and complexity to the network, potentially interfering with the learning process of the network. To address this issue, Bai proposed Grid-wise Feature Fusion (GFF) [34], which utilizes a maximum pooling layer for feature fusion. This method has a simple structure. However, the maximally pooled feature fusion selects only the maximum value in the pooling window as input and discards other features. This may result in some useful features being ignored in the pooling, thus affecting the loss of feature information. We propose the Grid Attention Feature Fusion method to fuse vehicle-side feature maps and roadside feature maps using a spatial attention mechanism [50,51]. The GAFF model structure is shown in Figure 6.

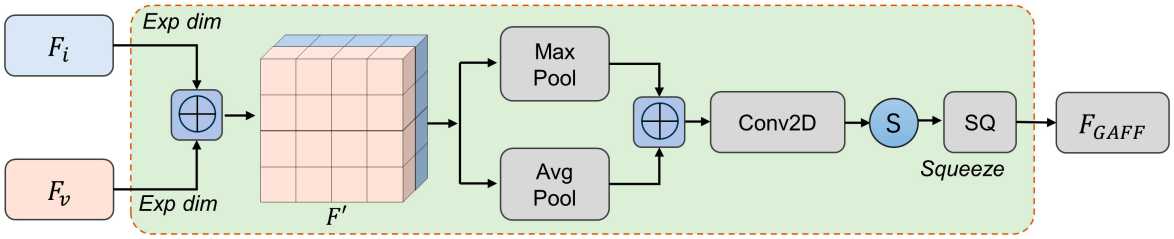

Figure 6. Grid attention feature fusion module.

GAFF first combines the vehicle-side BEV feature F_v with the roadside BEV feature F_i, and the formula is

$$F\prime(F_v, F_i) = Concat\left(f_{Exp\_dim}(F_v), f_{Exp\_dim}(F_i)\right), \tag{8}$$

where f_{Exp_dim} upscales the size of the feature from $((C, H, W)$ to $(C, H, W, 1)$, and the feature dimension of $F\prime$ is $(C, H, W, 2)$. Maximum pooling and average pooling are then used to operate on the last dimension. Feature fusion is then performed using Conv2D and a sigmoid function. The size of the feature is compressed into (C, H, W) using a squeeze operation to complete the attention-based feature fusion. The formula for GAFF is:

$$F_{GAFF}(F_v, F_i) = f_{Squeeze}(\sigma(f_{Conv2D}(Concat(MaxPool(F'(F_v, F_i)), AvgPool(F'(F_v, F_i)))))), \tag{9}$$

where σ denotes the sigmoid function. GAFF can adaptively determine the important regions in the input two feature maps and dynamically adjust the weight allocation according to the input feature maps. Maximum pooling and average pooling can suppress irrelevant or noisy information from the target features and make the model more focused on the target or object of interest [52]. This helps to reduce the interference of noise in the feature fusion process, improve the focus of the fused model on key features, as well as reduce the risk of the model being weakened by irrelevant information. Figure 7C shows the visualization of the center point feature heatmap after GAFF fusion. From Figure 7, it can be seen that the center point of the vehicle in the fused feature heatmap is more accurate, which indicates that the fused features using the GAFF module are more focused on the target itself, which improves the model's expressive and inference capabilities.

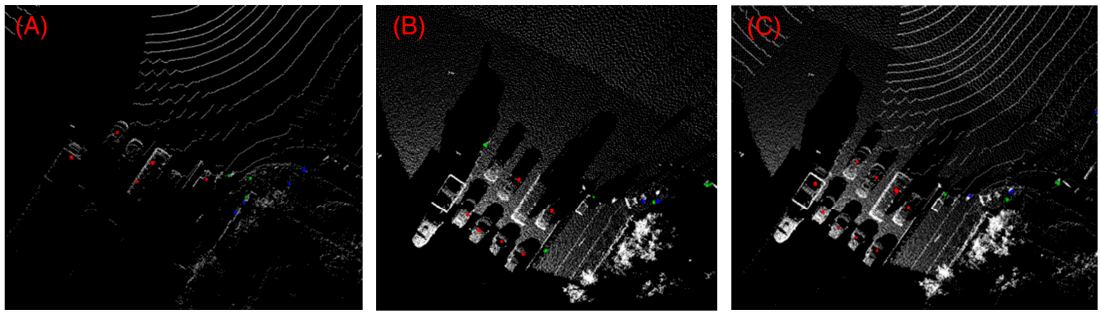

Figure 7. Visualization of center point heatmap. (**A**) Center point heatmap of vehicle side. (**B**) Center point heatmap of roadside. (**C**) Center point heatmap of VIC. Red, green, and blue points indicate class vehicles, class pedestrians, and class cyclists, respectively.

2.2.4. Detection Head

CenterPoint is a two-stage 3D point cloud target detection method that achieves a balance between detection accuracy and speed. Each target is represented by the centroid of the feature map in CenterHead, which achieves bounding box regression at each center location [18,53]. Due to its simple structure and excellent performance, we use the detection head in the anchor-free target detection method of CenterPoint. The center point of the CenterPoint feature map may be better for feature fusion and visualization. The structure of CenterHead is shown in Figure 8.

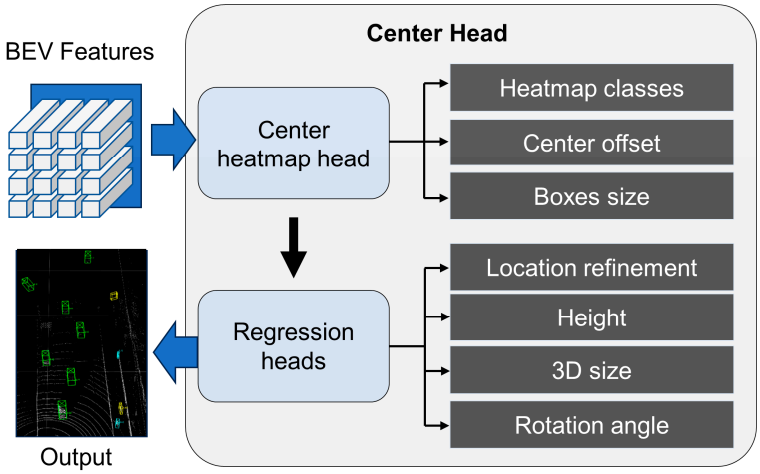

Figure 8. 2D feature extraction module.

3. Results

The proposed PAFNet network was trained and tested on a PC platform with a high-performance GPU. The PC hardware platform is equipped with a 12th Gen Intel® Core™ i5-12400F processor, an NVIDIA GeForce RTX 4080 graphics card, 32 GB DDR4 RAM, and a 2 TB SSD. The PC platform is equipped with the Ubuntu 20.04.6 LTS operating system, with the installation of CUDA 11.8 and CuDNN v8.6.0. We used the Pytorch 1.13.1 deep learning framework to build the PAFNet network model and mayavi 4.7.0 for point cloud and result visualization. We divided the DAIR-V2X-C dataset into a 70% training set and a 30% test set. The proposed PAFNet was trained and tested on the DAIR-V2X-C dataset, along with other algorithms for comparison experiments. Roadside data and vehicle-side data of DAIR-V2X-C were also utilized for training and testing.

The origin of the world coordinate system is chosen as the intersection of the center point of the roadside LiDAR with the plumb line of the ground plane and the ground plane. The data is preprocessed according to the world coordinate system. The target detection objects are vehicles, pedestrians, and cyclists. The following presents a visual analysis of the target detection results, as well as the results of ablation experiments.

3.1. Visualization Results and Analysis

In order to better visualize the experimental results, we used the random function in Python to select a set of data from all the test sets in the DAIR-V2X-C dataset. A visualization of the PAFNet target detection results is shown in Figure 9, with green bounding boxes for vehicles, blue bounding boxes for pedestrians, and yellow bounding boxes for cyclists. The red box indicates that the vehicle is completely occluded in the point cloud of the vehicle side and cannot be detected by automotive LiDAR. However, the roadside LiDAR can detect vehicles in the red box. Although the vehicle in the orange box is not detected on the roadside, it is detected in the target detection after feature fusion. The results demonstrate that our proposed feature fusion target detection algorithm, PAFNet, has better detection results.

In the feature fusion detection result, PAFNet is capable of detecting vehicle targets at 130.9 m, pedestrians at 79.3 m, and cyclists at 102 m. When comparing the detection results of using vehicle-side LiDAR alone and roadside LiDAR alone, PAFNet fused with both vehicle-side and roadside LiDAR detects a larger number of targets and detects them at a longer distance. The visualization of the target detection results shows that the network with fused vehicle-side and roadside features has a better target detection capability. After fusing the roadside features and vehicle-side features, PAFNet has a more accurate detection capability for vehicles, pedestrians, and cyclists at a long distance.

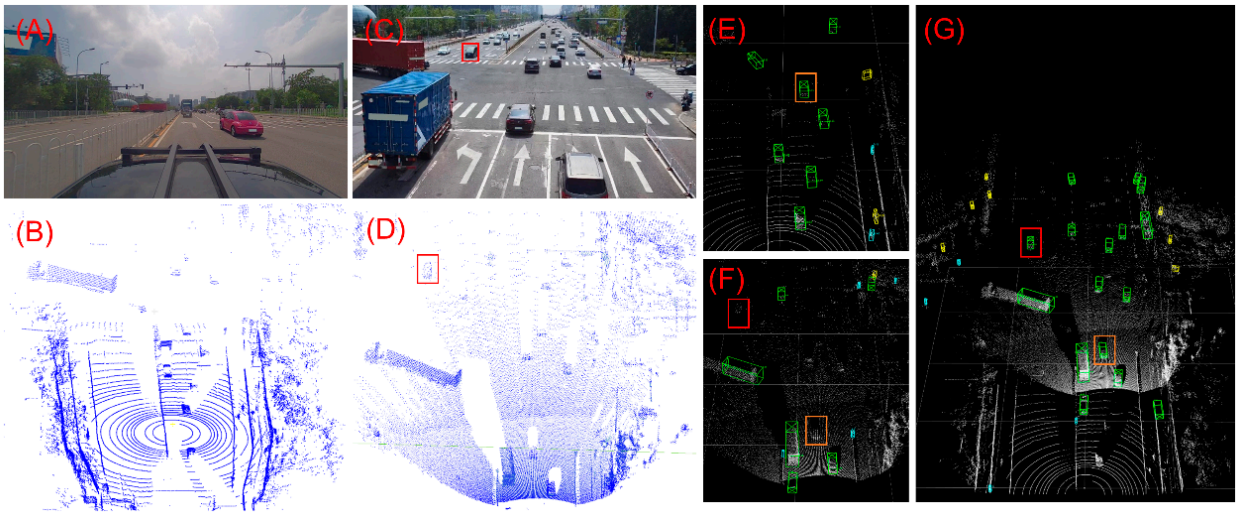

Figure 9. Visualization of target detection results. (**A**) Image collected by the vehicle-side camera. (**B**) Point cloud collected by the vehicle-side LiDAR. (**C**) Image collected by the roadside camera. (**D**) Point cloud collected by the roadside LiDAR. (**E**) Vehicle-side point cloud detection result. (**F**) Roadside point cloud detection result. (**G**) Feature fusion detection result. The red boxes and orange boxes indicate vehicles.

3.2. Detection Experiment

The experiment uses Average Precision (AP) and mean Average Precision (mAP) as performance metrics to evaluate the detectors. In order to balance the detection difficulty and evaluation criteria, to more accurately evaluate and compare the performance of different detection algorithms, we set the Intersection over Union (IoU) for vehicles to 70% and the IoU for pedestrians and cyclists to 50%. Figure 10 shows the 3D detection results of different algorithms for vehicles, pedestrians, and cyclists in the DAIR-V2X-C dataset. Our proposed PAFNet achieves 57.27%, 39.52%, and 41.95% detection accuracies with a mAP of 46.25%. PAFNet achieves the highest detection accuracy for feature fusion mode detection. The mAP of PAFNet is 5.91% higher than the second place FFNet [35], and 9.34% higher than PillarGrid [34]. Compared with all the detection results on automotive and roadside data, our proposed feature fusion target detection method, PAFNet, achieves the highest detection accuracy. While the detection accuracy of PAFNet is lower than that of the early fusion methods PointPillars [14] and VoxelNeXt [20], it is still comparable to the accuracy of PointPillars in early fusion.

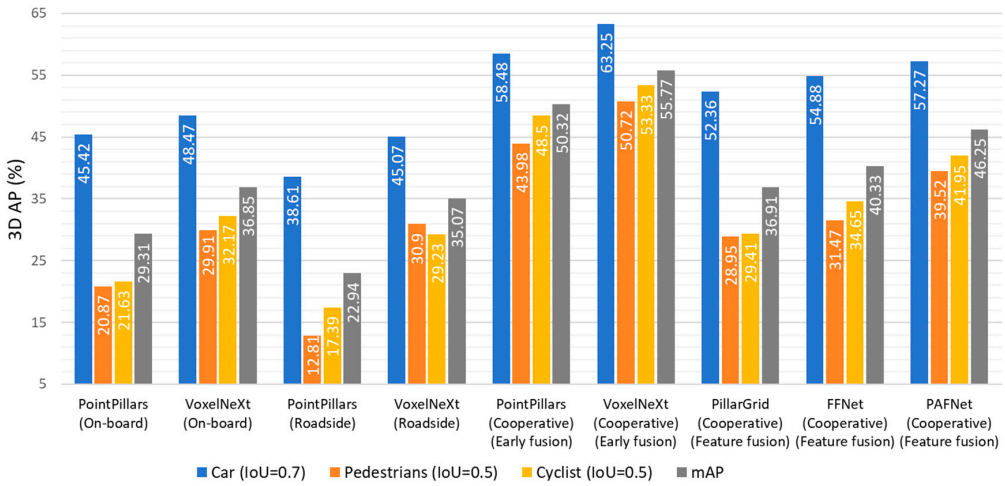

Figure 10. 3D detection results by different methods on DAIR-V2X-C.

Table 3 shows the parameters and detection time of the PAFNet network model. PAFNet utilizes an anchor-free detection method, which minimizes the number of parameters in the feature fusion process. The early fusion approach is faster than the feature fusion approach because it only needs to process either the roadside or vehicle-side point clouds. PAFNet employs an attention-based approach in feature fusion, which affects the inference speed.

Table 3. Model size and inference speed for different methods.

Method	Anchor	Detection Modality	Model Size	Speed
Pointpillars	Based	Early fusion	58.1 MB	69.85 ms
VoxelNeXt	Free	Early fusion	228.9 MB	103.21 ms
PillarGrid	Based	Feature fusion	114.8 MB	126.51 ms
FFNet	Based	Feature fusion	128 MB	131.86 ms
PAFNet	Free	Feature fusion	95.85 MB	133.52 ms

3.3. Ablation Experiment

Table 4 presents the results of the ablation experiments conducted on the STCFP, the ASPP, and the GAFF modules in PAFNet, along with the statistics of the 3D detection APs for vehicles, pedestrians, and cyclists. From Table 4, it can be seen that the ASPP module contributes a 1.88% increase in mAP to PAFNet. The ASPP module provides a wider receptive field for the feature extraction backbone network, and the multi-scale character also makes it perform better on targets of different scale types. In the GAFF module ablation experiments, GFF was used for feature fusion without the GAFF module. The experimental results demonstrate that the GAFF module, which utilizes the attention mechanism, provides a 5.22% increase in mAP for PAFNet. The improved fusion of vehicle-side and roadside feature information makes it easier to enhance the visibility of small targets such as pedestrians and cyclists. The accuracy of the fusion features is ensured by the STCFP through inter-frame matching and world coordinate transformation. Preprocessing in the DAIR-V2X-C dataset results in a 0.52% improvement in mAP for target detection.

Table 4. Ablation experiment of PAFNet.

STCFP Module	ASPP Module	GAFF Module	3D Detection AP (%)			
			Car (IoU = 0.7)	Pedestrian (IoU = 0.5)	Cyclist (IoU = 0.5)	mAP
×	×	×	51.56	30.37	31.70	37.88
√	×	×	52.62	30.54	32.03	38.40
√	×	√	56.08	36.42	38.37	43.62
√	√	×	53.80	32.73	34.29	40.27
√	√	√	57.27	39.52	41.95	46.25

4. Discussion

PAFNet achieves higher detection accuracy than the two feature fusion methods, PillarGrid and FFNet, for target detection on the DAIR-V2X-C dataset. PillarGrid uses the grid feature fusion method for fusion, but the detection accuracy is lower than that of PAFNet, which uses the attention mechanism fusion method. FFNet utilizes a 2D convolutional neural network with batch normalization for fusion of BEV pillar features. Additionally, it incorporates a feature pyramid for target detection. While this method is an improvement over the PillarGrid approach, there is still a gap in detection accuracy when compared to PAFNet. The GAFF feature fusion module adjusts the weight assignment adaptively to determine the fused feature values based on the roadside and vehicle-side features. It avoids directly selecting the maximum value in the features. This improves the efficiency of feature utilization and reduces the effect of noise in features on target detection. From the results of the comparison experiments and the ablation experiments, it can be

seen that the GAFF fusion method based on the spatial attention mechanism is an effective feature fusion method.

PAFNet has a higher detection accuracy than FFNet for vehicles, pedestrians, and cyclists by 2.39%, 8.05%, and 7.3%, respectively. This indicates that PAFNet has a higher detection accuracy for small targets such as pedestrians and cyclists. The ablation experiments also show that both the ASSP and GAFF modules contribute to the detection of small targets. We believe that the ASPP module in the feature extraction network enhances the edge features of small targets. Additionally, the multi-scale features provide more information about small targets. The GAFF module fuses the features from the vehicle side and the roadside, providing rich feature information that aids in the detection of small targets.

The early fusion method has a higher detection accuracy compared to the method using solely roadside LiDAR or solely vehicle-side LiDAR as shown in Figure 10, for both PointPillars and VoxelNeXt. Although the detection accuracy of the PAFNet feature fusion method is not as good as that of the early fusion method, the gap can be narrowed by optimizing the algorithm.

The ablation experiment shows that STCFP preprocessing improves vehicle, pedestrian, and bicyclist detection accuracy by 1.06%, 0.17%, and 0.33%, respectively. We believe that frame mismatch has a greater impact on vehicles than on pedestrians and cyclists. Vehicles move faster than pedestrians and cyclists, resulting in larger errors in inter-frame matching, which affects the accuracy of vehicle detection.

PAFNet maintains the advantage of the lightweight center point method. Compared to other feature fusion methods, PAFNet has fewer parameters, resulting in reduced computational costs. This contributes to the practicality of target detection for VIC autonomous driving.

5. Conclusions

In this paper, we propose a feature fusion target detection method named PAFNet for the cooperative sensing of roadside LiDAR and automotive LiDAR in autonomous driving. It is the first anchor-free target detection pipeline in VIC feature fusion 3D target detection. We introduce GAFF, a pillar-based attentional feature fusion method that uses a spatial attention mechanism to fuse features from the vehicle side and the roadside. This is the first time an attentional mechanism has been used in the feature fusion of roadside and vehicle-side point clouds. The feature extraction network utilizes the ASSP module to capture multi-scale information through dilated convolution with different sampling rates, enhancing the perception of targets of multiple sizes. We also propose a preprocessing method named STCFP for spatial and temporal cooperative fusion to ensure the matching of the fused features. Experiments on the DAIR-V2X-C dataset demonstrate that our proposed method, PAFNet, achieves higher detection accuracy than existing feature fusion target detection methods. Future research should investigate the impact of various feature fusion methods on the accuracy of 3D target detection using VIC feature fusion. Furthermore, it is suggested to optimize the inference time of feature fusion target detection and integrate these methods into the perception system of VIC autonomous driving. The VIC feature fusion 3D target detection methods are significant in enhancing the perception range of autonomous vehicles and reducing traffic accidents.

Author Contributions: L.W. and J.L. are co-first authors with equal contributions. Conceptualization, L.W. and J.L.; methodology, L.W.; validation, L.W.; writing, L.W.; investigation, L.W. and M.L.; supervision, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the 14th Five-Year Plan Funding of China, grant number 50916040401, and in part by the Fundamental Research Program, grant number 514010503-201.

Data Availability Statement: The DAIR-V2X-C dataset mentioned in this paper is openly and freely available at https://thudair.baai.ac.cn/roadtest (accessed on 29 January 2024 in Beijing, China).

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

References

1. Royo, S.; Ballesta-Garcia, M. An Overview of Lidar Imaging Systems for Autonomous Vehicles. *Appl. Sci.* **2019**, *9*, 4093. [CrossRef]
2. Li, Y.; Ibanez-Guzman, J. Lidar for Autonomous Driving: The Principles, Challenges, and Trends for Automotive Lidar and Perception Systems. *IEEE Signal Process. Mag.* **2020**, *37*, 50–61. [CrossRef]
3. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep Learning for 3d Point Clouds: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4338–4364. [CrossRef]
4. Fernandes, D.; Silva, A.; Névoa, R.; Simões, C.; Gonzalez, D.; Guevara, M.; Novais, P.; Monteiro, J.; Melo-Pinto, P. Point-Cloud Based 3d Object Detection and Classification Methods for Self-Driving Applications: A Survey and Taxonomy. *Inf. Fusion* **2021**, *68*, 161–191. [CrossRef]
5. Zhikun, W.; Jincheng, Y.; Ling, Y.; Sumin, Z.; Yehao, C.; Caixing, L.; Xuhong, T. Improved Hole Repairing Algorithm for Livestock Point Clouds Based on Cubic B-Spline for Region Defining. *Measurement* **2022**, *190*, 110668. [CrossRef]
6. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
7. 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*, 1–10.
8. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
9. Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; Ghanem, B. Pointnext: Revisiting Pointnet++ with Improved Training and Scaling Strategies. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23192–23204.
10. Ma, X.; Qin, C.; You, H.; Ran, H.; Fu, Y. Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual Mlp Framework. *arXiv* **2022**, arXiv:2202.07123.
11. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11040–11048.
12. 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, 18–23 June 2018; pp. 4490–4499.
13. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337. [CrossRef]
14. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
15. Kuang, H.; Wang, B.; An, J.; Zhang, M.; Zhang, Z. Voxel-FPN: Multi-Scale Voxel Feature Aggregation for 3d Object Detection from Lidar Point Clouds. *Sensors* **2020**, *20*, 704. [CrossRef]
16. 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. [CrossRef]
17. Shi, G.; Li, R.; Ma, C. Pillarnet: Real-Time and High-Performance Pillar-Based 3d Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 35–52.
18. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3D Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
19. Wang, G.; Wu, J.; Tian, B.; Teng, S.; Chen, L.; Cao, D. Centernet3D: An Anchor Free Object Detector for Point Cloud. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 12953–12965. [CrossRef]
20. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. Voxelnext: Fully Sparse Voxelnet for 3d Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Oxford, UK, 18–22 June 2023; pp. 21674–21683.
21. Li, J.; Luo, C.; Yang, X. Pillarnext: Rethinking Network Designs for 3d Object Detection in Lidar Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Oxford, UK, 18–22 June 2023; pp. 17567–17576.
22. Wang, G.; Wu, J.; Xu, T.; Tian, B. 3D Vehicle Detection with RSU Lidar for Autonomous Mine. *IEEE Trans. Veh. Technol.* **2021**, *70*, 344–355. [CrossRef]
23. Schinagl, D.; Krispel, G.; Possegger, H.; Roth, P.M.; Bischof, H. Occam's Laser: Occlusion-Based Attribution Maps for 3d Object Detectors on Lidar Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1141–1150.
24. Wu, J.; Xu, H.; Tian, Y.; Pi, R.; Yue, R. Vehicle Detection under Adverse Weather from Roadside LiDAR Data. *Sensors* **2020**, *20*, 3433. [CrossRef]
25. Wang, J.; Wu, Z.; Liang, Y.; Tang, J.; Chen, H. Perception Methods for Adverse Weather Based on Vehicle Infrastructure Cooperation System: A Review. *Sensors* **2024**, *24*, 374. [CrossRef]
26. Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3d Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 21361–21370.

27. Abdelazeem, M.; Elamin, A.; Afifi, A.; El-Rabbany, A. Multi-Sensor Point Cloud Data Fusion for Precise 3D Mapping. *Egypt. J. Remote Sens. Space Sci.* **2021**, *24*, 835–844. [CrossRef]
28. Zhou, Y.; Sun, P.; Zhang, Y.; Anguelov, D.; Gao, J.; Ouyang, T.; Guo, J.; Ngiam, J.; Vasudevan, V. End-to-End Multi-View Fusion for 3d Object Detection in Lidar Point Clouds. In Proceedings of the Conference on Robot Learning, Virtual, 16–18 November 2020; pp. 923–932.
29. Yu, H.; Yang, W.; Ruan, H.; Yang, Z.; Tang, Y.; Gao, X.; Hao, X.; Shi, Y.; Pan, Y.; Sun, N. V2x-Seq: A Large-Scale Sequential Dataset for Vehicle-Infrastructure Cooperative Perception and Forecasting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Oxford, UK, 18–22 June 2023; pp. 5486–5495.
30. Sun, P.; Sun, C.; Wang, R.; Zhao, X. Object Detection Based on Roadside LiDAR for Cooperative Driving Automation: A Review. *Sensors* **2022**, *22*, 9316. [CrossRef]
31. Chen, Q.; Tang, S.; Yang, Q.; Fu, S. Cooper: Cooperative Perception for Connected Autonomous Vehicles based on 3d Point Clouds. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 514–524.
32. Tang, Z.; Hu, R.; Chen, Y.; Sun, Z.; Li, M. Multi-Expert Learning for Fusion of Pedestrian Detection Bounding Box. *Knowl.-Based Syst.* **2022**, *241*, 108254. [CrossRef]
33. Hurl, B.; Cohen, R.; Czarnecki, K.; Waslander, S. Trupercept: Trust Modelling for Autonomous Vehicle Cooperative Perception from Synthetic Data. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 341–347.
34. Bai, Z.; Wu, G.; Barth, M.J.; Liu, Y.; Sisbot, E.A.; Oguchi, K. Pillargrid: Deep Learning-Based Cooperative Perception for 3d Object Detection from Onboard-Roadside Lidar. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 1743–1749.
35. Yu, H.; Tang, Y.; Xie, E.; Mao, J.; Luo, P.; Nie, Z. Flow-Based Feature Fusion for Vehicle-Infrastructure Cooperative 3d Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 1–9.
36. Raj, T.; Hanim Hashim, F.; Baseri Huddin, A.; Ibrahim, M.F.; Hussain, A. A Survey on Lidar Scanning Mechanisms. *Electronics* **2020**, *9*, 741. [CrossRef]
37. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. Carla: An Open Urban Driving Simulator. In Proceedings of the 1st Annual Conference on Robot Learning, Proceedings of Machine Learning Research (PMLR), Amsterdam, The Netherlands, 13–15 November 2017; pp. 1–16.
38. Beck, J.; Arvin, R.; Lee, S.; Khattak, A.; Chakraborty, S. Automated Vehicle Data Pipeline for Accident Reconstruction: New Insights from Lidar, Camera, and Radar Data. *Accid. Anal. Prev.* **2023**, *180*, 106923. [CrossRef]
39. Zhou, S.; Xu, H.; Zhang, G.; Ma, T.; Yang, Y. Leveraging Deep Convolutional Neural Networks Pre-Trained on Autonomous Driving Data for Vehicle Detection from Roadside Lidar Data. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22367–22377. [CrossRef]
40. Xie, S.; Gu, J.; Guo, D.; Qi, C.R.; Guibas, L.; Litany, O. Pointcontrast: Unsupervised Pre-Training for 3d Point Cloud Understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 574–591.
41. Fei, J.; Peng, K.; Heidenreich, P.; Bieder, F.; Stiller, C. Pillarsegnet: Pillar-Based Semantic Grid Map Estimation Using Sparse Lidar Data. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; pp. 838–844.
42. Yuan, Z.; Song, X.; Bai, L.; Wang, Z.; Ouyang, W. Temporal-Channel Transformer for 3d Lidar-Based Video Object Detection for Autonomous Driving. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2068–2078. [CrossRef]
43. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended Feature Pyramid Network for Small Object Detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [CrossRef]
45. Zhu, L.; Deng, Z.; Hu, X.; Fu, C.; Xu, X.; Qin, J.; Heng, P. Bidirectional Feature Pyramid Network with Recurrent Attention Residual Modules for Shadow Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 121–136.
46. Gao, J.; Geng, X.; Zhang, Y.; Wang, R.; Shao, K. Augmented Weighted Bidirectional Feature Pyramid Network for Marine Object Detection. *Expert Syst. Appl.* **2024**, *237*, 121688. [CrossRef]
47. Lian, X.; Pang, Y.; Han, J.; Pan, J. Cascaded Hierarchical Atrous Spatial Pyramid Pooling Module for Semantic Segmentation. *Pattern Recognit.* **2021**, *110*, 107622. [CrossRef]
48. Qiu, Y.; Liu, Y.; Chen, Y.; Zhang, J.; Zhu, J.; Xu, J. A2sppnet: Attentive Atrous Spatial Pyramid Pooling Network for Salient Object Detection. *IEEE Trans. Multimed.* **2023**, *25*, 1991–2006. [CrossRef]
49. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder-Decoder Network and Structural Similarity Loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
50. Guo, M.; Xu, T.; Liu, J.; Liu, Z.; Jiang, P.; Mu, T.; Zhang, S.; Martin, R.; Cheng, M.; Hu, S. Attention Mechanisms in Computer Vision: A Survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
51. Yan, J.; Peng, Z.; Yin, H.; Wang, J.; Wang, X.; Shen, Y.; Stechele, W.; Cremers, D. Trajectory Prediction for Intelligent Vehicles Using Spatial-Attention Mechanism. *IET Intell. Transp. Syst.* **2020**, *14*, 1855–1863. [CrossRef]

52. Chen, J.; Chen, Y.; Li, W.; Ning, G.; Tong, M.; Hilton, A. Channel and Spatial Attention Based Deep Object Co-Segmentation. *Knowl.-Based Syst.* **2021**, *211*, 106550. [CrossRef]
53. Xue, Y.; Mao, J.; Niu, M.; Xu, H.; Mi, M.B.; Zhang, W.; Wang, X.; Wang, X. Point2seq: Detecting 3d Objects as Sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8521–8530.