energies

MDPI

Article

A Vehicle-to-Infrastructure beyond Visual Range Cooperative Perception Method Based on Heterogeneous Sensors

Tong Luo [1], Long Chen [1,*], Tianyu Luan [1], Yang Li [2] and Yicheng Li [1]

[1] Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China
[2] School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China
* Correspondence: chenlong@ujs.edu.cn

Abstract: With the development of autopilot, the performance of intelligent vehicles is constrained by their inability to perceive blind and beyond visual range areas. Vehicle-to-infrastructure cooperative perception has become an effective method for achieving reliable and higher-level autonomous driving. A vehicle-to-infrastructure cooperative beyond visual range and non-blind area method, based on heterogeneous sensors, was proposed in this study. Firstly, a feature map receptive field enhancement module with spatial dilated convolution module (SDCM), based on spatial dilated convolution, was proposed and embedded into the YOLOv4 algorithm. The YOLOv4-SDCM algorithm with SDCM module achieved a 1.65% mAP improvement in multi-object detection performance with the BDD100K test set. Moreover, the backbone of CenterPoint was improved with the addition of self-calibrated convolutions, coordinate attention, and residual structure. The proposed Centerpoint-FE (Feature Enhancement) algorithm achieved a 3.25% improvement in mAP with the ONCE data set. In this paper, a multi-object post-fusion matching method of heterogeneous sensors was designed to realize the vehicle-to-infrastructure cooperative beyond visual range. Experiments conducted at urban intersections without traffic lights demonstrated that the proposed method effectively resolved the problem of beyond visual range perception of intelligent vehicles.

Keywords: beyond visual range perception; intelligent infrastructure; multi-object detection; vehicle-to-infrastructure cooperation

check for updates

Citation: Luo, T.; Chen, L.; Luan, T.; Li, Y.; Li, Y. A Vehicle-to-Infrastructure beyond Visual Range Cooperative Perception Method Based on Heterogeneous Sensors. *Energies* **2022**, *15*, 7956. https://doi.org/10.3390/en15217956

Academic Editor: Andrea Mariscotti

Received: 24 September 2022
Accepted: 21 October 2022
Published: 26 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Environment sensing is one of the key modules in autonomous driving systems. However, the limited ability of a single agent creates a bottleneck in sensing performance. A single agent's sensing performance is limited by occlusion and sparse data. In order to overcome the limitations of single-agent sensing performance, multiple agents are required. Vehicle-to-infrastructure (V2I) communication has made significant advancements, and has become a crucial component in the development of intelligent driving systems. V2I consists of intelligent vehicles and intelligent roads, which can not only identify surrounding obstacles and road traffic information on their own, but also send information to other vehicles, in order to ensure the safety of traffic between intelligent vehicles and nearby vehicles.

Accurate road sensing is crucial for subsequent decision-making and control tasks, but the performance of common sensors in smart devices has both benefits and drawbacks. A camera that is equipped with a deep learning algorithm can accurately identify all types of objects [1], but it lacks depth information, and is highly environment-dependent [2,3]. LiDAR has the advantages of having high accuracy and depth information, but it is difficult to process sparse and irregular points, and it lacks textural information. Therefore, a variety of sensor fusion sensing schemes are now required for intelligent transportation facilities, in order to improve their environmental perception capabilities.

Compared to single-vehicle intelligence, vehicle–road collaborative sensing has more perception perspectives and a wider perception range, allowing intelligent vehicles to more

accurately assess the surrounding road conditions, which effectively improves commuting efficiency and driving safety. V2I can effectively compensate for the fact that the perception distance of a fast self-driving vehicle exceeds the vehicle's safe perception range, as well as the vehicle's ability to perceive turning, blocking, and other road conditions.

Due to the presence of at least two subject objects in V2I, the intelligent vehicle and the intelligent roadside must conduct environmental awareness and information interaction, respectively, in order to achieve matching and tracking of the surrounding targets, followed by the decision making and control. Therefore, road multi-target recognition and cross-view matching have become key problems necessary to solve in vehicle–road cooperative sensing.

This paper proposed a vehicle-to-infrastructure beyond visual range cooperative perception method, based on heterogeneous sensors, to address the aforementioned issues. The main contributions of this article are summarized as follows:

1. Proposed a spatial dilated convolution module (SDCM) that uses spatial dilated convolution. It was incorporated into the YOLOv4 algorithm, in order to improve the algorithm's detection accuracy, without increasing its computational load.
2. Based on the CenterPoint network, the backbone was enhanced by substituting the original traditional convolutional layer with a self-calibrated convolutions layer, and by adding coordinate attention and residual structure, which effectively improves the network's feature extraction capability.
3. Proposed and constructed a collaborative sensing platform with high flexibility for road vehicles; set cooperative perception modes and deployed perception algorithm on the basis of various scenes; and selected the intersection to conduct the beyond visual range perception experiment.

2. Related Studies

2.1. Multi-Object Detection Based on Vision

Vision sensors have strong environmental perception ability. The conventional multi-object detection algorithms consist of two and one stages [4].

In the model training phase, the two-stage detection algorithm can better balance the problem of unequal positive and negative samples, but its real-time performance struggles to meet the real-time requirements of intelligent driving vehicles. Therefore, the one-stage algorithm has become the prevalent visual detection algorithm for intelligent vehicles.

In the design of convolutional neural networks, small convolution kernels are generally used to realize the convolution operation. For example, in the common ResNet [5], convolution kernels of size 3×3 or 1×1 are used, except that the first convolutional layer uses a convolution kernel of size 7×7, in order to realize the dimensionality of the original image channel. A larger convolution kernel will improve the receptive field of the feature map [6], and the convolution layer's learned features will become more global.

Receptive fields have been proven to actually affect the performance of the entire convolutional neural network [7]. In convolutional neural networks, insufficient receptive fields are a common issue. Large convolutional kernels will necessitate a substantial amount of computation, degrading the algorithm's detection speed performance. In order to enhance the receptive field of the feature map and avoid additional computational overhead, it is a good method to effectively improve detection accuracy by expanding the receptive field.

Using dilated convolution should be considered, in order to augment the receptive field of the future map. The following are common methods for enhancing the receptive field of a feature map. They include increasing the number of convolutional layers, and pooling layers to deepen the depth of the neural network, in order to obtain larger-feature receptive fields. However, excessively deep neural networks result in the disappearance of gradients, which makes model training convergence difficult. The above problems can be optimized by designing a special convolution operation to replace the standard convolution operation, and thereby obtain larger receptive fields [8]. For example, the

dilated convolution used in this paper continuously improves the receptive fields of several feature maps of varying scales, following the feature fusion module.

2.2. 3D Multi-Object Detection Based on LiDAR

LiDAR is capable of highly accurate three-dimensional sensing. Multi-object detection algorithm based on LiDAR is one of the core algorithms that is used in road environment sensing. However, the point cloud cannot be directly input into the neural network, due to its uneven spatial distribution; therefore, the point cloud must be regularly processed. LiDAR multi-object detection algorithms can be divided into two categories, based on the different processing methods of point cloud data. One of the categories is the point-based LiDAR detection algorithm; another category is the voxel-based LiDAR detection algorithm. The voxel-based LiDAR detection algorithm is derived from Voxel Net [9], which divides point clouds using regular voxels, and the sparse convolution proposed by Second [10]. Second is then utilized as the universal backbone in voxel-based methods. The single-stage algorithm is similar to SASSD [11], and the two-stage algorithm is similar to voxel RCNN [12]; both can achieve high detection accuracy. At the same time, with the development of anchor-free detection in the field of 2D vision, AFDet [13] and CenterPoint [14] introduced anchor Free, which greatly reduced the model volume and comprehensively improved the accuracy and speed of the algorithm. The voxel-based CenterPoint algorithm is one of the most advanced 3D object detection algorithms among these two categories. It employs advanced anchor-free detection headers to improve network overall performance of object identification accuracy. However, compared to the advanced Anchor Free, Backbone is relatively inferior at detecting head defects, which limits the algorithm's ability to achieve a breakthrough in accuracy, and prevents its performance from advancing further.

2.3. Vehicle-to-Infrastructure beyond Visual Range Cooperative Perception

In recent years, the re-identification of traffic participants, primarily pedestrians and vehicles [15,16], has developed rapidly. Several useful re-ID methods have been proposed, including those based on feature learning [17,18], metric learning [19,20], and spatial and temporal information [21,22].

Intelligent vehicles interact with various kinds of things around them, in order to exchange information which can improve perception performance, named V2X (vehicle to everything). Based on the various methods of information fusion, it can be divided into three categories: (1) early fusion [23], (2) intermediate fusion [24,25], and (3) late fusion [26,27]. This provides a solid basis for V2X cooperative perception.

Currently, the most common V2X solutions and validation are based on simulation platforms, such as CALAR [28]. It rarely undergoes actual testing. Therefore, it will be beneficial to develop a set of flexible intelligent roadside facilities.

3. Methodology

The structure of the vehicle-to-infrastructure beyond visual range cooperative perception algorithm is shown in Figure 1. Figure 1a is a vision-based multi-object detection network. The feature map receptive field enhancement module SDCM is described and embedded with the YOLOv4 algorithm. Figure 1b depicts the partial network for 3D object detection. The algorithm is based on CenterPoint, and has three components: a voxelization module, a 3D backbone with residual structure, and a 2D backbone with self-calibrated convolutions and coordinated attention. The detection results are matched using the Hungarian algorithm, and a Kalman filter is then utilized to predict and update the road multi-object. Finally, based on the relative position of the mobile platform and the roadside platform, the appropriate beyond visual range perception scheme is selected. The methodology is detailed below.

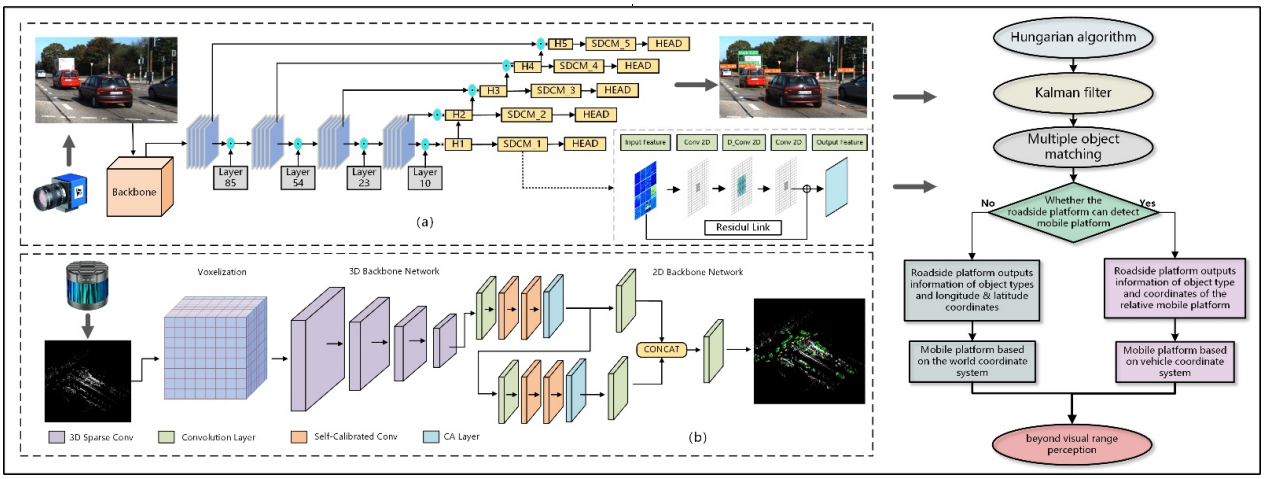


Figure 1. Structure of vehicle-to-infrastructure beyond visual range cooperative perception algorithm, (**a**) structure of 2D multi-object detection algorithm, (**b**) structure of 3D multi-object detection algorithm.

3.1. 2D Multi-Object Detection Algorithm YOLOv4-SDCM

A spatial dilated convolution module (SDCM) that is based on dilated convolution was proposed to address the issue of the absence of feature receptive fields in the YOLOv4 algorithm, as a consequence of the algorithm using many convolution layers with small kernels. Receptive field enhancement was performed on the feature map output by the feature fusion module in the YOLOv4 algorithm, which is named YOLOv4-SDCM.

YOLOv4-SDCM is an improvement algorithm that is based on YOLOv4. The SDCM module uses the convolution layer of 3×3, and the dilation rate of different scale feature is shown in Table 1; the activation function uses Relu function, and the hyperparameter in Head and other settings are the same as for YOLOv4.

The receptive field can be calculated as shown in Formula (1):

$$RF_n = \begin{cases} k_n, n = 1 \\ RF_{n-1} + (k_n - 1) \times S_{n-1}, n \geq 2 \end{cases} \quad (1)$$

where "n" means the number of convolution layers, "k_n" represents the size of the convolution kernel, "RF_{n-1}"is the receptive outlier of the previous layer, and "S_{n-1}"is the step size value in the previous convolution or pooling layer.

According to Formula (1), the receptive field expands as the step size increases; however, a large step size will result in features that are too sparse. The dilated convolution represents an innovation in convolution kernel size. Dilated convolution is achieved by inserting zero values around each position of the convolution kernel. For a k*k convolution kernel, assuming that the cavity rate of the dilated convolution is d, the theoretical convolution size of the dilated convolution is k_i, which is calculated in (2):

$$k_i = k + (k-1) \cdot (d-1) \quad (2)$$

In the dilated convolution operation, the lower the dilated rate, the richer the local information learned by the convolution layer. The greater the dilated rate, the larger the feature receptive field and the more global the feature. Meanwhile, local details will be lost. The representation abilities of local information and global information of feature maps are contradictory.

The ASPP [29] algorithm processes images using a set of dilated convolutions with varying dilated rates, in an attempt to reconcile the aforementioned contradictions. The ASPP algorithm was proposed for image segmentation; however, the segmentation task is fundamentally distinct from that of object detection. Image-based segmentation algorithms have stricter requirements for global features, requiring that the features that are extracted

by convolutional neural networks have sufficient global receptive fields, whereas image-based object detection algorithms must also consider the local receptive fields in the region that contains a single target object.

Road multi-object detection in intelligent driving scenarios has the problem of the coexistence of differences between large and small objects, which requires feature maps of different scales to be extracted by the feature extraction network, in order to have receptive fields that correspond to object scales. In contrast to the ASPP structure, the SDCM proposed in this paper is designed to meet the image object detection algorithm's receptive field of the feature map requirements. Each scale feature map is specifically processed using dilated convolution with varying dilated rates. In order to prevent the operation of dilated convolution from focusing excessively on global features, the SDCM employs the concept of residual network structure and residual connection to recover the detailed features of the feature map that were processed by dilated convolution. The specific mathematical representation of SDCM is shown in Equations (3) and (4):

$$DC_i = Conv_2(DC_i(Conv_1(Input_i))) \tag{3}$$

$$X = In_i + DC_i \tag{4}$$

where $Input_i$ $(i \in [0,4])$ represents the feature input of each scale; DC_i convolution represents dilated layers with different dilated rates.

The structure of the YOLOv4-SDCM algorithm is shown in Figure 2. In the SDCM structure, the first standard convolution layer (Conv 2D_1) is primarily aimed at reducing the dimension of the channel number of the input feature map, while the second standard convolution layer (Conv 2D_2) is responsible for restoring the channel number of the feature map to its original value, following receptive field enhancement. Finally, the residual structure is used to combine the output of the original feature map and the output of the second convolutional layer, in order to complete the recovery of the feature map's detailed features, and to obtain the final feature output enhanced by the receptive field.

One of the most important problems in SDCM design is the selection of the dilated rate for the dilated d convolution. According to the low dilated rate, more emphasis is placed on the local and global features, which is more conducive to the detection of small objects. Large dilated rates pay more attention to the global features of the entire image, which is conducive to the large object detection principle. The "D_Conv 2D" in SDCM is the dilated convolution structure; the dilated rates corresponding to five different scales are shown in Table 1. The five feature map indexes correspond to the five layers in the SDCM structure. For example, the dilated rates of the feature map index are "0" in the SDCM_1 layer.

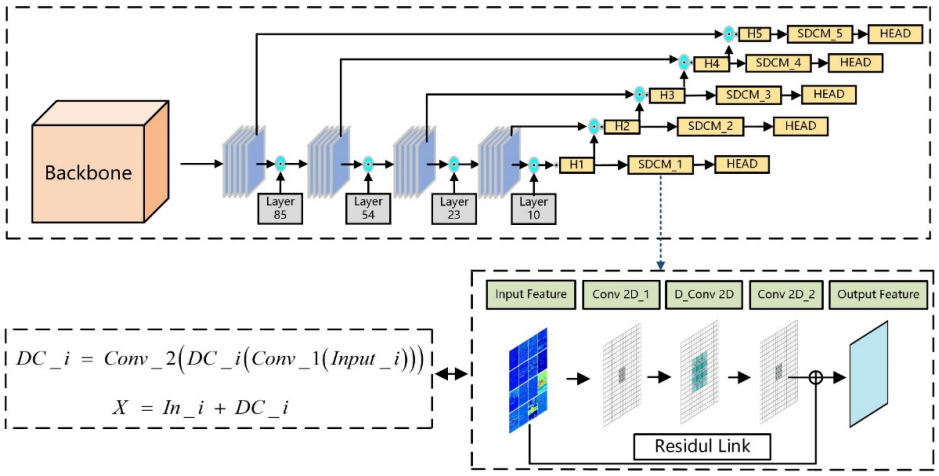


Figure 2. Schematic diagram of the overall structure of YOLOv4-SDCM.

Table 1. The dilation rates of different scale features.

Feature Map Index	Size	Dilation Ratio
0	19 × 19	4
1	38 × 38	3
2	76 × 76	2
3	152 × 152	1
4	304 × 304	1

3.2. 3D Multi-Object Detection Algorithm Centerpoint-FE

In this paper, the backbone of CenterPoint is improved, thereby improving the network feature extraction capability. The improved algorithm is referred to as Centerpoint-FE. The particular procedures are described in the following sections.

3.2.1. 2D Backbone Network Improvement

Through different convolutions of two groups of channels, CenterPoint's 2D backbone network extracts and fuses shallow structural features and deep semantic features, and its application in some networks [11,12] demonstrates that this structure is effective. Therefore, the structure of the 2D backbone network was not changed, but a more efficient network layer replaced the traditional convolutional layer.

Self-calibrated convolution SC-CONV, as shown in Figure 3a, was used to replace the original traditional convolutional layers. Self-calibrated convolution enables each spatial location to adaptively integrate the surrounding context information, greatly enhancing the network's receptive field.

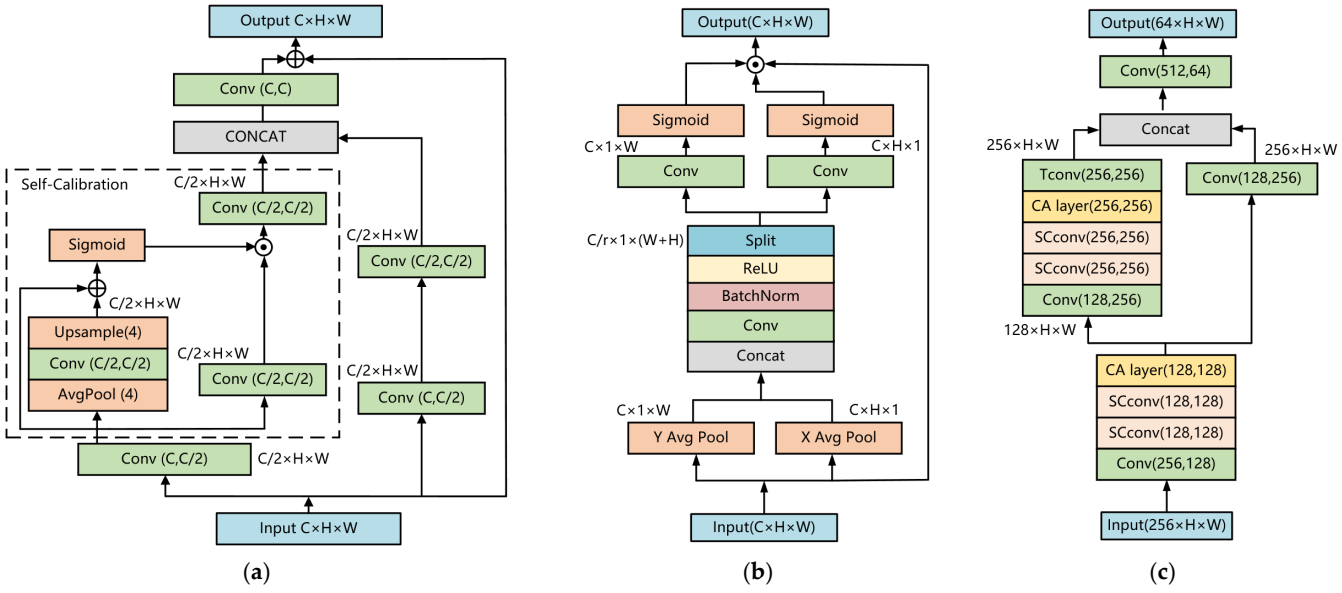


Figure 3. 2D structure of Centerpoint-FE. (**a**) SC conv module; (**b**) CA module; (**c**) 2D backbone.

At the same time, the development of the attention mechanism significantly expanded the feature extraction ability of the network. However, spatial feature information cannot be effectively modeled using multi-attention mechanisms such as SENET [30] and CBAM [31]. In order to solve this problem, we used coordinate attention [32], as shown in Figure 3b, to aggregate the features that are extracted by self-calibrated convolution (SC-CONV). CA is an efficient attention mechanism that integrates the space and channel information of the entire input feature via independent, one-dimensional global pooling in both horizontal and vertical directions, while efficiently preserving the feature location information. Finally, by combining the two one-dimensional features, the complete spatial location of the feature can be determined based on the feature channel dimensional encoding; thus, the attention

mechanism encoding that takes both the channel and the spatial feature into account is realized. Two layers of self-calibrated convolution and one layer of CA layer were combined to form a feature extraction module in the proposed 2D backbone, which replaced the common convolutional layer group in the original 2D backbone.

The proposed complete 2D backbone network of Centerpoint-FE is shown in Figure 3c.

3.2.2. 3D Backbone Network Improvement

The 3D backbone network of CenterPoint uses a combination of sparse convolutional and submanifold convolutional to complete the 8-fold down-sampling of 3D voxel features.

We observed that the residual structure was not applied in this process, so we added the residual structure to the submanifold convolution block, in order to improve the network's ability to extract features.

Figure 4 depicts the Centerpoint-FE submanifold convolution module with the residual structure.

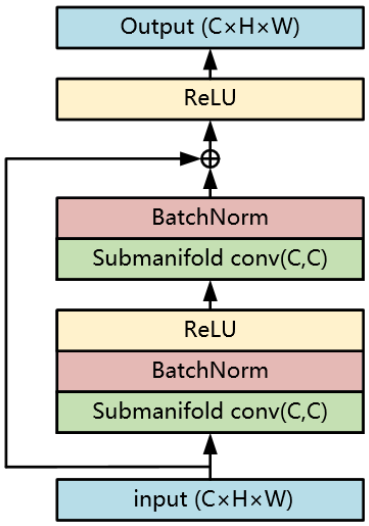


Figure 4. Submanifold convolutional of Centerpoint-FE.

The format of the output of the detection algorithm used in this study was (x, y, z, l, w, h, cla, θ), where (x, y, z) denotes the center point of the detection frame, (l, w, h) represents the length, width, and height information of the 3D detection frame, cla represents the type of object, and θ represents the frame's orientation angle.

In the actual environmental perception algorithm, the multi-target tracking algorithm was an indispensable part.

Due to the lack of textural information in the point cloud data, it was difficult to construct stable observation features and implement a tracking-by-detection strategy. However, the point cloud data have the ability to directly obtain the accurate 3D geometric measurement of the target, i.e., (x, y, z, l, w, h, cla, θ). Therefore, the kinematic characteristics of the target were exhaustively excavated in this study, i.e., the motion smoothness hypothesis and consistency hypothesis. Based on the aforementioned characteristics, we transformed the multi-object association problem into a bipartite graph global matching problem. We used the Hungarian algorithm to solve this problem, and then applied the Kalman filter time-domain filtering to each associative matching tuple, in order to improve the accuracy of the motion state estimation and achieve target tracking.

3.3. Algorithm of beyond Visual Range Perception Based on Vehicle-to-Infrastructure

Figure 5 shows the flowchart of the road vehicle collaborative sensing technique. The camera-based roadside and mobile platforms perform two-dimensional information perception of the surrounding environmental targets, while the LiDAR performs three-dimensional information perception of the surrounding targets. We obtained high-precision

latitude and longitude data from the GPS (Global Positioning System). When the mobile platform is outside the detection range of the roadside platform, the roadside platform transmits to the mobile platform the type of the detected target, as well as the corresponding positional information (latitude and longitude in the world coordinate system). When the mobile platform enters the detection range of the roadside platform, the roadside platform converts the detected target's coordinates into the mobile platform's relative positional information. The transformed information is sent to the mobile platform for realization beyond the visual range, based on road–vehicle cooperative sensing.

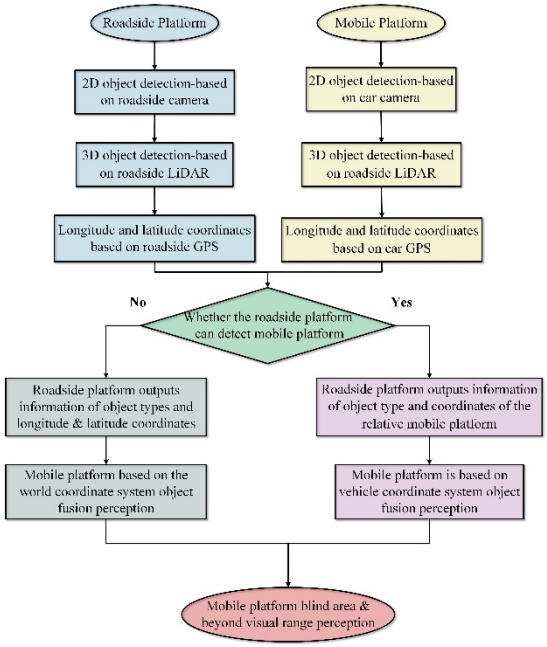


Figure 5. Flowchart of the road–vehicle collaborative perception technique.

We deployed the differential positioning system separately on mobile platforms and roadside platforms, in order to obtain high-precision positioning information, course angles, and time stamps.

The roadside platform obtains accurate position information in the world coordinate system, using high-precision positioning. By combining this information with the relative position information of the detected target, absolute and relative position coordinates of multiple targets within its detection range are obtained.

When the mobile platform is outside the detection range of the roadside platform, the roadside platform and the mobile platform use high-precision positioning to obtain their precise positioning in the world coordinate system. The roadside platform detects the road targets within its coverage area, using a camera and LiDAR. By acquiring the detection information, the relative coordinates of the obstacle in relation to the roadside platform can be computed. In addition, the type and size of the obstacle are obtained. We transformed the value relative to the roadside platform into the world coordinate system, in order to obtain the world coordinates of the target within the detection range of the roadside platform. The roadside platform sends the mobile platform detection results along with latitude and longitude information. Thus, the mobile platform obtains road conditions prior to entering the detection range of the roadside platform, and achieves perception that is beyond the visual range.

When the mobile platform enters the detection range of the roadside platform, the roadside platform captures the type and size of the mobile platform, along with the relative position, by considering the roadside platform as the coordinate origin. Using a coordinate transformation, all detected target position information in the roadside platform

is converted into relative coordinate information with the mobile platform as the coordinate origin, using edge calculation.

4. Experiments and Results

4.1. 2D Multi-Object Detection Algorithm YOLOv4-SDCM

In this paper, KITTI and BDD100K data sets were used to train and validate the algorithm. For the BDD100K data set with ten categories, in order to eliminate the influence of long mantissa data as much as possible, the BDD100K data set was combined with labels. The final labels that were used for the training set were car, bus, truck, person, traffic Light, traffic Sign, and non-motorized vehicle. The training set used 70K images as the training set, and 10K images were used as the validation set. The experiment was conducted on an NVIDIA GTX 2080Ti with CUDA 10.0. Ubuntu 18.04 was used as the system, and the programming language was Python, and the machine learning package used was PyTorch.

Table 2 shows the object detection accuracy comparison of YOLOv4-SDCM, YOLOv4 and Faster R-CNN [4], on the basis of BDD100K data sets. It can be seen from the table that, on the BDD100K test set, the mAP values that were obtained by the YOLOv4 and Faster R-CNN algorithm evaluations were 0.6590 and 0.6385, respectively. However, the mAP of the YOLOv4-SDCM object detection algorithm was 0.6755, indicating that the road multi-object detection algorithm YOLOv4-SDCM proposed in this study, with an enhanced receptive field, was superior to the baseline model algorithm YOLOv4; the comprehensive detection accuracy was improved by 1.65%. The frames per second (FPS) by the YOLOv4 and Faster R-CNN algorithm evaluations were 10.8 and 52.2, respectively. The FPS of the YOLOv4-SDCM object detection algorithm was 50.4, which is very fast, and meets real-time requirements.

Table 2. Object detection accuracy comparison of Faster R-CNN, YOLOv4 and YOLOv4-SDCM on BDD100K data sets.

	Faster R-CNN	**YOLOv4**	**SDCM-YOLOv4**
Car	0.8012	0.7902	0.8036 (+1.34%)
Person	0.6501	0.6387	0.6422 (+0.35%)
Bus	0.6754	0.6525	0.6936 (+2.51%)
Truck	0.6589	0.6739	0.6928 (+1.90%)
Traffic light	0.5628	0.5830	0.5807 (−0.23%)
Traffic sign	0.7018	0.7216	0.7218 (+0.02%)
Bike	0.5098	0.5128	0.5140 (+0.12%)
mAP@0.5	0.6385	0.6590	0.6755 (+1.65%)
FPS	10.8	52.2	50.4

Figure 6 depicts YOLOv4-SDCM and YOLOv4 object detection with the KITTI data set, where the IoU threshold for the detection score was set to 0.5. There were many kinds and different sizes of objects in the intelligent driving traffic detection scene at the intersection. This type of object-dense detection scene can effectively test the algorithm's detection capabilities. As shown in Figure 6, the red circle indicates that YOLOv4-SDCM was able to detect objects that YOLOv4 could not, which demonstrated the algorithm's superior object detection performance. The comparative analysis demonstrated that the YOLOv4-SDCM algorithm significantly improves object detection accuracy, under the assumption of almost no loss of small object detection accuracy, demonstrating the effectiveness of the improvements presented in this study.

4.2. 3D Multi-Object Detection Algorithm Centerpoint-FE

Data set: Huawei's open source ONCE data set [33] was used as the training set. ONCE provides 16,000 frames of annotated data, including sunny, rainy, and cloudy weather conditions, in addition to various morning and evening time periods, of which

5000 frames are the training set, 3000 frames are the validation set, and 8000 frames are the test set.

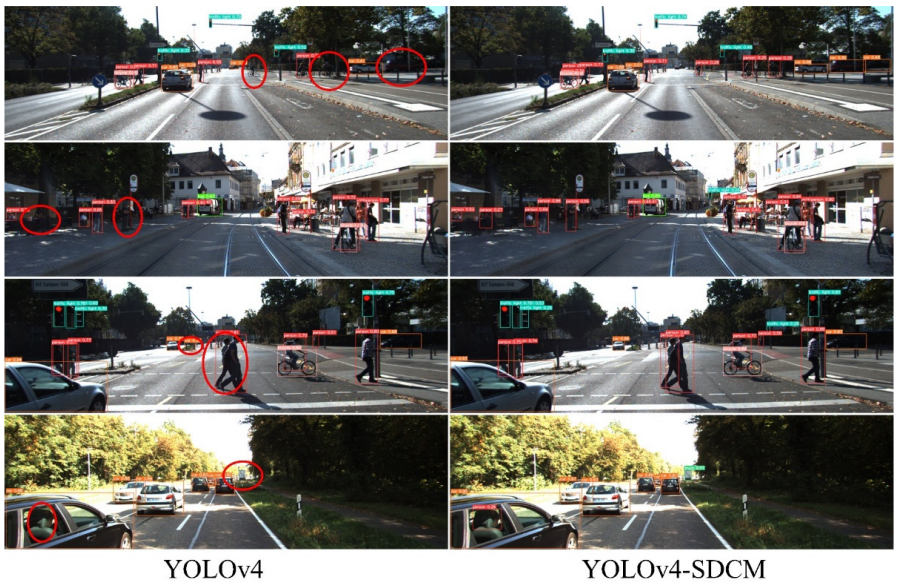


Figure 6. Visualization of YOLOv4-SDCM and YOLOv4 object detection with the KITTI data set.

Data augmentation: As the volume occupied by vehicles, pedestrians, and other objects in the real point cloud space is very small, the learning of features of the network is limited. In order to increase the number of multiple target augmentation samples for each frame of training data, a GT sampling data augmentation strategy was employed to randomly select 8, 2, and 2 truth samples for car, pedestrian, and cyclist targets, respectively. In addition, we randomly flipped the data on the X-axis and Y-axis, and randomly rotated by $[-\frac{\pi}{4}, +\frac{\pi}{4}]$ on the Z-axis. Finally, we performed global random scaling of [0.95,1.05] on the global data.

Training: We trained our network on a GPU (Graphics Processing Unit) server with four RTX 3090 GPUs, a set epoch of 80, and a batch size of 32. The AdamW optimizer employed a one-cycle learning strategy with a maximum learning rate of 0.001, a decay factor of 0.01, and a momentum between 0.85 and 0.95.

On the validation set, the accuracy of the trained model was evaluated, and the results are listed in Table 3.

Table 3. Comparison of object detection accuracy of CenterPoint-FE algorithm results.

Method	Vehicle				Pedestrian				Cyclist				mAP
	Overall	0–30 m	30–50 m	50 m-inf	Overall	0–30 m	30–50 m	50 m-inf	Overall	0–30 m	30–50 m	50 m-inf	
CenterPoint (Org)	66.79	80.10	59.55	43.39	49.90	56.24	42.61	26.27	63.45	74.28	57.94	41.48	60.05
CenterPoint (Retrain)	76.11	86.81	70.44	56.62	49.01	57.81	42.73	22.55	64.93	76.54	59.95	41.60	63.35
CenterPoint (ours)	78.12	87.72	72.23	62.05	52.73	61.76	45.88	26.47	68.93	80.29	62.41	47.39	66.60
Improvement	2.01	0.91	1.89	5.43	3.72	3.95	3.15	3.92	4.00	3.75	2.46	5.79	3.25

As shown in Table 3, the mAP of the retrained basic model was 63.35%, which is higher than the official mAP of 60.05%; therefore, the retrained version was used as the baseline model.

The table outlines the detection accuracy of the model at various distances for vehicles, pedestrians, and bicycles. On the validation set mAP, the Centerpoint-FE model improved by 3.25% to 66.60%.

For different classes, the accuracy of the Centerpoint-FE model improved at all distance segments. Among them, the total mAP for the category of vehicles was enhanced by

2.01%. In short-range, medium-range, and long-range object detection, mAP increased by 0.91%, 1.89%, and 5.43%, respectively. The results indicate that the enhancement of Backbone significantly enhanced the feature extraction capability of long-range vehicle targets, resulting in a significant improvement in precision.

The total mAP of the pedestrian model increased by 3.72%, and the AP increased by 3.95%, 3.15%, and 3.92%, in short-, medium-, and long-range object detection, respectively. The total mAP of the cyclist class increased by 4.00%, while the AP increased by 3.75%, 2.46%, and 5.79% for short-, medium-, and long-range object detection, respectively. For the pedestrian and cyclist classes, unlike the vehicle class, the AP of the long-distance category had the largest improvement, which is because the point clouds of pedestrian and cyclist targets are sparse; therefore, Backbone's AP promotion was more uniform across all distance segments for these classes than for the vehicle category.

Ablation experiments: Ablation experiments were performed on the various proposed methods, in order to verify the effectiveness of the improvements; the results are shown in Table 4.

Table 4. Ablation study results for different components.

SC Conv	CA Layer	Residual Block	mAP	FPS
			63.35	27.17
✓			64.33	23.28
✓	✓		64.89	21.63
✓	✓	✓	66.60	21.12

As shown in Table 4, the 2D backbone network with 3-layer SC Conv increased mAP by 0.98%. The last layer of the 3-layer SC Conv was replaced by the CA attention mechanism, which brought a 0.56% mAP improvement for the model. The proposed 2D model backbone yielded an overall mAP improvement of 1.54%. In the 3D backbone network structure, the addition of the residual structure increased the network's mAP by 1.71%. In terms of network speed, Centerpoint-FE reached 21.12 FPS, which still meets real-time requirements.

As depicted in Figure 7, the CenterPoint and the proposed CenterPoint-FE were visually evaluated using the ONCE data set. The left represents the CenterPoint detection results; the right represents the Centerpoint-FE detection results; the green box represents the vehicle; the blue box represents the cyclist; and the yellow box represents the pedestrian.

In the figure on the left, the area denoted by the red circle represents a CenterPoint detection failure; meanwhile, the Centerpoint-FE algorithm correctly detected the object. It is evident from Figure 7 that the proposed method significantly improves the accuracy of detecting distant vehicle targets and small targets, further validating its effectiveness.

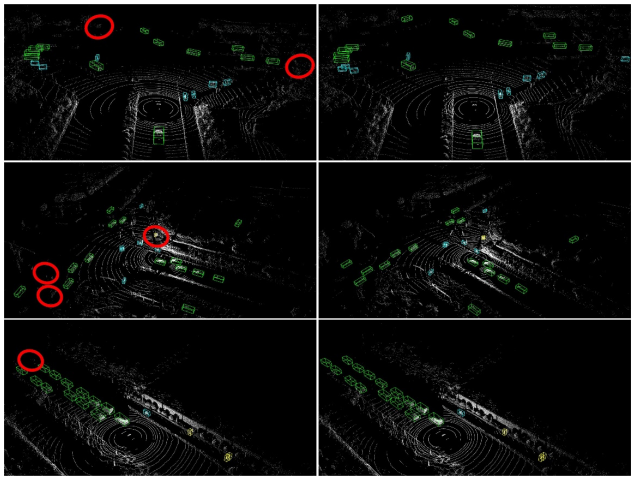


Figure 7. Visual detection results for CenterPoint (**left**), and Centerpoint-FE (**right**).

4.3. Vehicle-to-Infrastructure beyond Visual Range Cooperative Perception Test

The crosses and T-shaped intersections are the most common and important use scenarios in the urban traffic environment. At intersections, there are a large number of vehicles, complicated road conditions, and many perceptual blind spots. In order to achieve beyond visual range object perception based on road–vehicle coordination at the intersection, we moved and set the highly flexible roadside platform in the center of the road at a closed campus T-junction for the mobile and roadside platform collaborative perception experimental prototype vehicle. This is presented in Figure 8.



Figure 8. The road–vehicle collaborative perception mobile and roadside prototype platform.

Figure 9 demonstrates the real-time results of the roadside platform's perception at a T-shaped intersection. The moment is the same as the mobile platform's perception presented in Figure 10. The mobile platform accurately detected pedestrians, bicycles, and vehicles. However, due to the sensor's performance of detection range and the road division at the intersection, it was unable to obtain information about the road conditions ahead of the intersection.

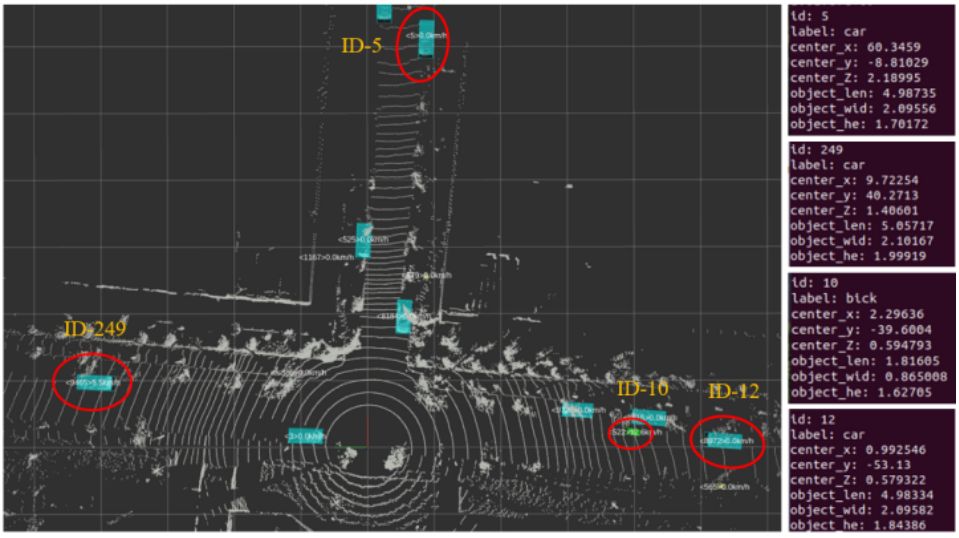


Figure 9. Visualized output of roadside platform perception results.



Figure 10. Visualized output of the mobile platform perception results.

The roadside unit accurately recognized various targets at the intersection. Figure 9 depicts ID-12 as the mobile platform, and ID-10 as the electric vehicle that was driving in front of the vehicle. When the roadside platform recognizes and tracks objects within its sensing range, it will give the objects an ID, and output the relative position information of the object and the roadside platform. ID-5 and ID-249 are taken as examples in Figure 9. The X-axis coordinate of the mobile platform is 0.99, and the Y-axis coordinate is -53.13 in relation to the roadside. By coordinate transformation, the coordinates of ID-5 relative to the mobile platform, based on the right-hand coordinate system, are (-8.71,93.33) meters. The coordinates of ID-249 relative to the mobile platform, based on the right-hand coordinate system, are (-59.31, 44.33) meters. The mobile platform realized the perception of beyond-horizon targets in the straight and turning states of the intersection ahead of it, with the help of road–car cooperative driving.

Information is transmitted by the road side unit (RSU) and onboard unit (OBU) modules, which may be based on 4G, 5G, dedicated short-range communication (DSRC), or other technologies. Based on road–vehicle cooperative sensing and coordinate conversion, the sensing range of the mobile platform, when driving in a straight line, increased from the original 60 m to achieve beyond the visual range distance of 93.33 m. When the mobile platform entered the detection range of the roadside unit, road–vehicle cooperative sensing allowed it to achieve up to 180 m beyond the visual range.

Figure 11 presents the visual output of the perception results of the mobile platform. The visual interface makes it clear that when the mobile platform is obstructed by a vehicle in front, it only detects the obstructing vehicle and the left electric vehicle in front, regardless of the road conditions of the obstructed portion. Neither vision nor LiDAR could complete the perception. Figure 12 presents the visualized interface of the roadside platform at the intersection. In the figure, ID-12 denotes the mobile platform, ID-169 denotes the front obstructing vehicle, and ID-159 denotes the electric vehicle on the left front of the mobile platform. According to the perception results of the roadside platform, there were multiple vehicles and pedestrians at the intersection, and the information on the front scene could not be obtained, since the mobile platform's view was obstructed. The roadside platform captured the mobile platform, and applied the same method of coordinate conversion, in order to detect targets within its range. The mobile platform acquired the target type and position information of the intersection ahead of the road, enabling perception beyond the visual range.



Figure 11. Visualized output of mobile platform perception results.

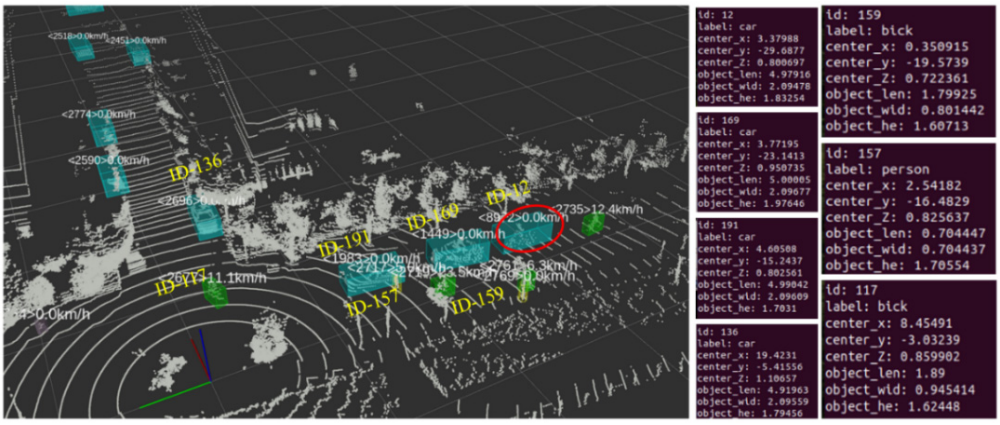


Figure 12. Visualized output of roadside platform perception results.

5. Conclusions

In this study, a feature map receptive field enhancement module, SDCM, based on a spatial dilated convolution module, was first proposed. It was integrated into the YOLOv4 algorithm. Compared with YOLOv4, the YOLOv4-SDCM algorithm utilized the SDCM module to improve object detection accuracy by 1.65% with the BDD100K test set, and its detection rate reached 50.4 FPS. Subsequently, the Backbone of CenterPoint was improved by incorporating self-calibrated convolutions, coordinate attention, and residual structures. The improved Centerpoint-FE algorithm achieved 66.60% mAP accuracy with the ONCE data set, 3.25% mAP improvement over the CenterPoint algorithm, and a detection rate of 21.1 FPS. We proposed a highly flexible road–vehicle cooperative perception scheme beyond the visual range, and employed the post-fusion matching method and detection algorithms in order to implement it. When a mobile platform entered the perception range of a roadside platform, it was able to obtain target type and position information within the detection range of the roadside platform, regardless of whether it was facing straight, turning, or obstructing conditions; it realized beyond visual range and blind area perception, resulting in a more than 200% increase in perception range. This is a scheme that can effectively improve the perception performance of object detection accuracy, as well as the safety performance of intelligent vehicles.

Author Contributions: Methodology, T.L. (Tong Luo), T.L. (Tianyu Luan), and Y.L. (Yang Li); software, T.L. (Tong Luo), T.L. (Tianyu Luan), and L.C.; validation, T.L. (Tong Luo) and Y.L. (Yang Li); data curation, T.L. (Tong Luo), L.C. and Y.L. (Yichen Li), writing—original draft preparation, T.L. (Tong Luo) and T.L. (Tianyu Luan). All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by the National Natural Science Foundation of China (U20A20333, 52072160, 61906076), Key Research and Development Program of Jiangsu Province (BE2019010-2, BE2020083-3).

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

References

1. Wang, H.; Xu, Y.; He, Y.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. YOLOv5-Fog: A Multiobjective Visual Detection Algorithm for Fog Driving Scenes Based on Improved YOLOv5. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2515612. [CrossRef]
2. Wang, H.; Chen, Y.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. SFNet-N: An Improved SFNet Algorithm for Semantic Segmentation of Low-Light Autonomous Driving Road Scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, 1–13. [CrossRef]
3. Liu, Z.; Cai, Y.; Wang, H.; Chen, L.; Gao, H.; Jia, Y.; Li, Y. Robust Target Recognition and Tracking of Self-Driving Cars With Radar and Camera Information Fusion Under Severe Weather Conditions. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 6640–6653. [CrossRef]
4. Luo, T.; Wang, H.; Cai, Y.; Chen, L.; Wang, K.; Yu, Y. Binary Residual Feature Pyramid Network: An Improved Feature Fusion Module Based on Double-Channel Residual Pyramid Structure for Autonomous Detection Algorithm. *IET Intell. Transp. Syst.* **2022**. [CrossRef]
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
6. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
7. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
8. Cai, Y.; Luan, T.; Gao, H.; Wang, H.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. YOLOv4-5D: An Effective and Efficient Object Detector for Autonomous Driving. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 4503613. [CrossRef]
9. Zhou, Y.; Tuzel, O. Voxelnet: End-to-End Learning for Point Cloud Based 3d Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
10. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337. [CrossRef] [PubMed]
11. He, C.; Zeng, H.; Huang, J.; Hua, X.-S.; Zhang, L. Structure Aware Single-Stage 3d Object Detection from Point Cloud. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11873–11882.
12. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; Li, H. Voxel R-Cnn: Towards High Performance Voxel-Based 3d Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 1201–1209.
13. Ge, R.; Ding, Z.; Hu, Y.; Wang, Y.; Chen, S.; Huang, L.; Li, Y. Afdet: Anchor Free One Stage 3d Object Detection. *arXiv* **2020**, arXiv:2006.12671.
14. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3d Object Detection and Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
15. Dong, C.; Chen, X.; Dong, H.; Yang, K.; Guo, J.; Bai, Y. Research on Intelligent Vehicle Infrastructure Cooperative System Based on Zigbee. In Proceedings of the 2019 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, UK, 14–17 July 2019; pp. 1337–1343.
16. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). In Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 480–496.
17. Meng, D.; Li, L.; Liu, X.; Li, Y.; Yang, S.; Zha, Z.-J.; Gao, X.; Wang, S.; Huang, Q. Parsing-Based View-Aware Embedding Network for Vehicle Re-Identification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7103–7112.
18. Liu, X.; Liu, W.; Mei, T.; Ma, H. *A Deep Learning-Based Approach to Progressive Vehicle Re-Identification for Urban Surveillance*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 869–884.
19. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609. [CrossRef]
20. Luo, H.; Jiang, W.; Zhang, X.; Fan, X.; Qian, J.; Zhang, C. AlignedReID++: Dynamically Matching Local Information for Person Re-Identification. *Pattern Recognit.* **2019**, *94*, 53–61. [CrossRef]
21. Wu, C.-W.; Liu, C.-T.; Chiang, C.-E.; Tu, W.-C.; Chien, S.-Y. Vehicle Re-Identification with the Space-Time Prior. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 121–128.
22. Jiang, N.; Xu, Y.; Zhou, Z.; Wu, W. Multi-Attribute Driven Vehicle Re-Identification with Spatial-Temporal Re-Ranking. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 858–862.
23. 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.
24. Vadivelu, N.; Ren, M.; Tu, J.; Wang, J.; Urtasun, R. Learning to Communicate and Correct Pose Errors. *arXiv* **2020**, arXiv:2011.05289.
25. Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; Urtasun, R. *V2vnet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 605–621.
26. Rauch, A.; Klanner, F.; Rasshofer, R.; Dietmayer, K. Car2x-Based Perception in a High-Level Fusion Architecture for Cooperative Perception Systems. In Proceedings of the 2012 IEEE Intelligent Vehicles Symposium, Madrid, Spain, 3–7 June 2012; pp. 270–275.

27. Rawashdeh, Z.Y.; Wang, Z. Collaborative Automated Driving: A Machine Learning-Based Method to Enhance the Accuracy of Shared Information. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3961–3966.
28. Wang, H.; Yu, Y.; Cai, Y.; Chen, X.; Chen, L.; Liu, Q. A Comparative Study of State-of-the-Art Deep Learning Algorithms for Vehicle Detection. *IEEE Intell. Transp. Syst. Mag.* **2019**, *11*, 82–95. [CrossRef]
29. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
31. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
33. Mao, J.; Niu, M.; Jiang, C.; Liang, H.; Chen, J.; Liang, X.; Li, Y.; Ye, C.; Zhang, W.; Li, Z. One Million Scenes for Autonomous Driving: Once Dataset. *arXiv* **2021**, arXiv:2106.11037.