Multimodal Fusion with Multiple Attention Mechanisms for 3D Target Detection Algorithm

Xiucai Zhang , Lei He * , Junyi Chen , Baoyun Wang , Yuhai Wang , Yuanle Zhou

Posted Date: 28 July 2023

doi: 10.20944/preprints202307.1956.v1

Keywords: Multimodal fusion; Attention mechanism; 3D target detection; Deep learning

Article

Multimodal Fusion with Multiple Attention Mechanisms for 3D Target Detection Algorithm

Xiucai Zhang [1], Lei He [1,*] Junyi Chen [1]Baoyun Wang [1] Yuhai Wang [1] and Yuanle Zhou [1]

1 State Key Laboratory of Automotive Simulation and Control ; Jilin University, Changchun 130022, China; zhangxc21@mails.jlu.edu.cn(X.Z.);jlu_helei@jlu.edu.cn(L.H.);chenjy1101@foxmail.com(J.C.);wangby22@mails.jlu.edu.cn(B. W.);wangyuhai@jlu.edu.cn(Y.W.);zhouyl22@mails.jlu.edu.cn(Y.Z.)

* Correspondence: jlu_helei@jlu.edu.cn

Abstract: This paper proposes a multimodal fusion 3D target detection algorithm based on the attention mechanism to improve the performance of 3D target detection. The algorithm utilizes point cloud data and information from camera. For image feature extraction, the ResNet50+FPN architecture extracts features at four levels. Point cloud feature extraction employs the voxel method and FCN to extract point and voxel features. The fusion of image and point cloud features is achieved through regional point fusion and regional voxel fusion methods. After information fusion, the Coordinate attention mechanism and SimAM attention mechanism extract fusion features at a deep level. The algorithm's performance is evaluated using the DAIR-V2X dataset. The results show that compared to the Part-A2 algorithm, the proposed algorithm improves the mAP value by 7.9% in BEV view and 7.8% in 3D view at IOU=0.5 (cars) and IOU=0.25 (pedestrians and cyclist). At IOU=0.7 (cars) and IOU=0.5 (pedestrians and cyclist), the mAP value of the SECOND algorithm is improved by 5.4% in the BEV view and 4.3% in the 3D view, compared to other comparison algorithms.

Keywords: multimodal fusion; attention mechanism;3D target detection; deep learning

1. Introduction

The continuous development of driverless technology has led to increasingly complex vehicle environments. Consequently, 3D target detection has gained significant attention[1–3]. However, single-modal 3D target detection alone cannot handle such complex scenes. Therefore, the research focus has shifted towards multi-modal 3D target detection, a crucial aspect of driverless environment perception.

Convolutional Neural Network (CNN) based techniques have performed well on detection datasets of images[4]. Significant achievements have been made in 2D target detection in images[4–7]. However, these methods cannot be directly applied to 3D detection due to the different input modes. Because LiDAR can pinpoint objects in 3D space, detection techniques based on LiDAR data are often superior to camera-based 3D detection techniques. Some of these methods convert 3D point clouds into depth maps and bird's eye view (BEV) maps by manual processing and then process them in a 2D-CNN manner for vehicle detection and classification[8]. However, the manually extracted features must fully utilize the information from the point cloud and may lead to performance degradation when detecting fewer points or variable geometry objects. There are also approaches[9] that use a 2D detector to generate a 2D detection frame on the image, transform the 2D detection frame into a proposed region in 3D space, and then use the PointNet[10] architecture for target detection on the point cloud. However, this approach relies heavily on the performance of the 2D target detector and cannot take advantage of the 3D information to generate robust bounding boxes.

Recent research in 3D target detection has primarily focused on utilizing end-to-end trainable neural networks that can directly process point cloud data without requiring manual feature extraction, as seen in the case of BEV maps. Qi et al.[11] developed a neural network architecture that can directly utilize point clouds as inputs and output class labels. This allows for the learning of representations from raw data. However, due to limitations within the network architecture and the high storage cost, this approach cannot be applied to target detection and localization. To overcome this issue, Zhou and Tuzel[12] proposed VoxelNet, which involves voxelizing the point cloud and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

employing a series of voxel feature encoding (VFE) layers. This processing allows the VoxelNet network to directly extract the point cloud within the voxel using 3D convolution. Another model proposed by S. Shi et al., called PointRCNN[14], offers superior performance in 3D target detection of point clouds compared to the two-stage image target detection network, Fast-RCNN[13]. Despite its advantages, PointRCNN requires further clarification regarding the point cloud pooling strategy. The pooling of different proposals may lead to the pooling of the same set of points, resulting in the loss of geometric information encoding. To address this, a new point cloud pooling operation for regions of interest Part-A2[15] networks was proposed, which retains all information from non-empty and empty voxels within the proposals, eliminating ambiguity from previous point cloud pooling strategies. Although these methods have demonstrated improved performance, they all rely solely on a single modality, namely point cloud data. RGB images, on the other hand, offer denser texture, color, and additional information than point clouds, suggesting that both modalities can be leveraged to enhance detection performance.

This paper aims to solve the issues mentioned above by implementing a fusion method that combines LiDAR point cloud and RGB image for 3D target detection. The process involves utilizing ResNet50 as the backbone network for image feature extraction and the FPN structure to gather multi-level features from the images. Point cloud and image early and late fusion are achieved through regional point fusion and voxel fusion. The Coordinate and SimAM attention mechanisms further process the information extracted after point cloud image fusion. Finally, features are outputted using the SECOND-FPN structure. The network structure framework depicted in Figure 1 demonstrates the approach proposed in this paper.

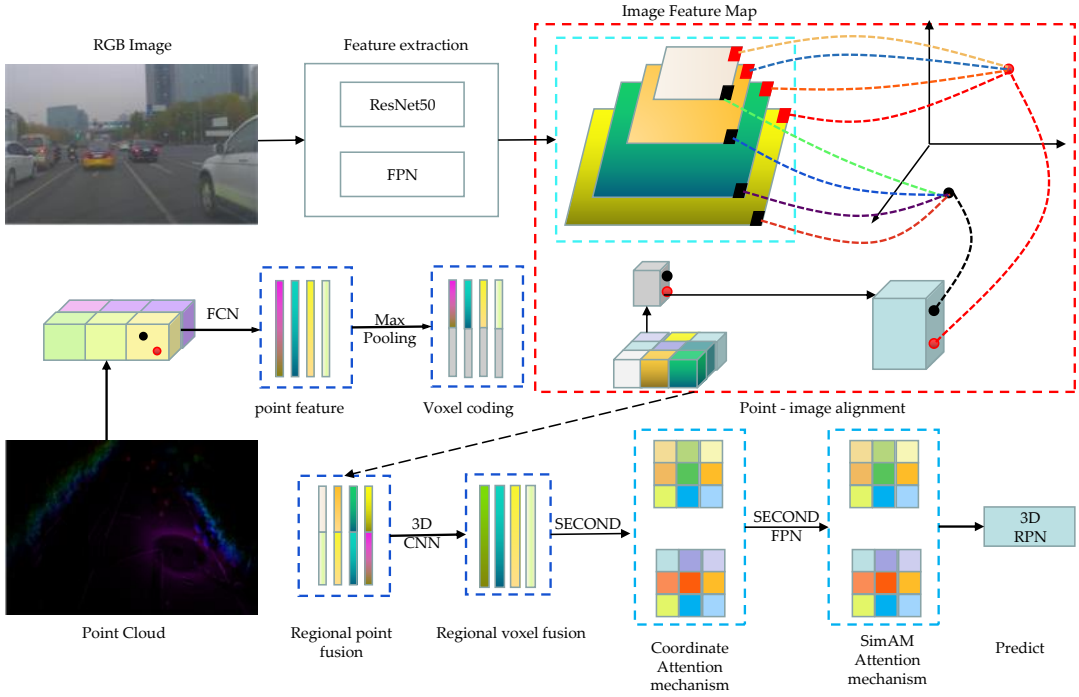

Figure 1. The framework of 3D object detection based on multimodal fusion.

2.Related Work

This section introduces 3D object detection techniques for vehicles, pedestrians, and cyclists using various technologies in literature research. Much research has been done on 3D object detection in autonomous driving and ADAS systems in recent years. It is divided into three categories: camera-based sensor, LiDAR sensor, camera, and LiDAR multi-mode fusion.

2.1 Target Detection Based on Camera Sensor

Currently, several methods exist to estimate 3D bounding boxes based on 2D image information[16–19]. [20,21] utilizes geometric constraints between 3D and 2D bounding boxes to restore the object detection attitude of 3D objects. [22–24] uses the similarity between the 3D object and the CAD model to restore the 3D object detection attitude. Chen el. [25,26] proposed to express

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

the three-dimensional geometric frame as an energy function and score the predefined three-dimensional box. More recently, [27,28] has explored stereoscopic images to improve the 3D detection performance of stereoscopic cameras. Due to the lack of accurate depth information, these pieces can only produce rough 3D inspection results and can be significantly affected by changes in appearance.

2.2 Target Detection Based on LiDAR Sensor

The research of 3D target detection based on LiDAR is a hot topic. In the early stages, the handmade feature method [29–33] was utilized successfully, but only under clear texture information and comprehensive 3D data. Subsequently, certain technologies have opted to employ voxel grid occupancy to represent 3D point clouds [34–36], which is then utilized for 3D bounding box calculations through 3D convolution. However, due to such methods' significant computational and memory demands, the BEV feature graph-based approach has been proposed [37,38]. This method assumes that point clouds are sparse in vertical height, although most scenarios do not meet such requirements.

In another approach, a two-stage object detection network is used for 3D object detection. In the first stage, regional proposals are generated first. In the second stage, point clouds and related semantic features in these candidate regions are reused to return to more accurate 3D bounding boxes. Some methods use mature two-dimensional detectors to obtain a two-dimensional Region of Interest (ROI) on the corresponding image and then transform these two-dimensional regions into three-dimensional space by inverse projection to obtain conical three-dimensional point clouds[39,40]. Finally, the conical three-dimensional point cloud was used as input to extract the region's features of interest through PointNet/ConvNet.

2.3 Target Detection Based on Camera and LiDAR Fusion

The existing research on multi-mode information fusion target detection based on LiDAR and RGB image data is limited[41,42]. Aiming at the problem that multi-view features are difficult to fuse effectively, Zhang et al [43] proposed a multi-view feature adaptive fusion 3D object detection framework. Chen et al. [1] proposed the Multi-View 3D Target Detection network (MV3D), which inputs LiDAR and image multi-mode data and combines regional features to generate 3D enclosing frames. Although this method uses multi-mode fusion and achieves good results, issues such as point cloud information loss and late-stage multi-mode information fusion limit the information exchange between data modes. Ku et al[44]. proposed a multi-mode fusion network combined with regional features, achieving better detection results by designing RPN structures with high-resolution feature mapping and improved performance in detecting and classifying small objects.

In another class of methods, Qi et al. introduced Frustum PointNets[40], a 3D target detection methodology combining LiDAR and image. The model first employs a 2D detector to generate a 2D detection box on the image, which is then transformed into a suggested area in 3D space. Finally, the PointNet architecture is utilized to conduct target detection on the point cloud. While this approach prioritizes image information, it does not utilize both sources of information to their fullest potential[45]. Unfortunately, the current state-of-the-art methods are inadequate to detect 3D objects in the broad field and multi-object view of infrastructure. We adopt an attention-based multi-mode fusion strategy to overcome this limitation to facilitate information exchange between multiple modes in the initial stages.

3. Image Feature Extraction Framework

In image processing, the performance of a convolutional neural network improves as the depth increases, leading to more advanced extracted features. However, traditional convolutional neural networks encounter issues such as network recession, gradient explosion, and gradient vanishing with increasing depth. Consequently, deeper networks tend to perform worse than shallower ones. To address these problems, this paper adopts ResNet50[46], which employs a residual module as the backbone network for image extraction (shown in Figure 2). The introduction of the residual module solves the issue of network recession, while the addition of a BN layer addresses gradient disappearance and explosion. ResNet50 consists of five stages, namely stage0 to stage5. The initial stage performs simpler tasks, primarily preprocessing the input data. The subsequent four stages

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

(stage1 to stage4) are composed of Bottleneck structures that extract high-level semantic features from the image.

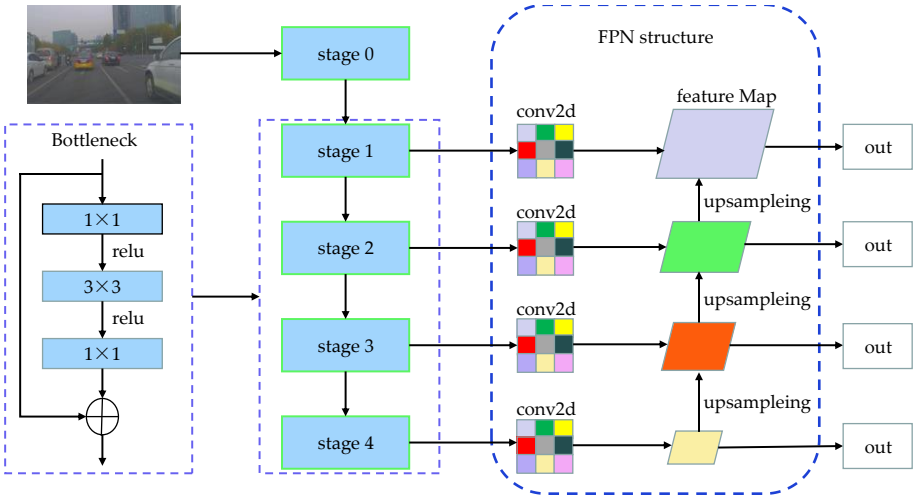

Figure 2. Image feature extraction network framework.

To enhance the ability to integrate low-level details and high-level semantics, it is necessary to expand the sensory field of the bottom layer and improve the detection performance of small targets. This paper adopts the multi-scale feature pyramid structure of the FPN in order to fuse the low-level detail information and the high-level semantic information, thereby increasing the sensory field of the bottom layer and enhancing the detection performance of small targets. The FPN structure, as depicted in Figure 2, consists of three lines: the self-low upward, the self-top downward, and the lateral link. The self-low-up module continuously pools the forward propagation feature maps, resulting in four feature maps of different sizes. The top-down module up-samples the small-size feature maps and performs splicing and fusion operations with the large feature maps from the low-up process. The horizontal link adjusts the output of different feature maps using a 1×1 convolutional kernel with 256 channels, facilitating subsequent fusion. Finally, the final prediction output is performed on the four feature maps, leading to the formation of a multi-level and multi-scale feature pyramid structure.

4. Point Cloud Feature Extraction

4.1 Point Cloud Voxelization

In order to facilitate feature extraction, this paper adopts a method similar to VoxelNet[12] to perform voxelization, grouping, and random extraction operations on the sparsely and haphazardly scattered point cloud throughout the space. Voxelization divides the 3D space into equally spaced voxels, with voxel sizes defined based on the range of the point cloud along the Z, Y, and X directions denoted as D, H, and W respectively. Points are then grouped based on the voxel where they are located. Due to factors such as distance, occlusion, and sparsity, the number of points in each voxel can vary, as shown in the leftmost voxelization of Figure 3.

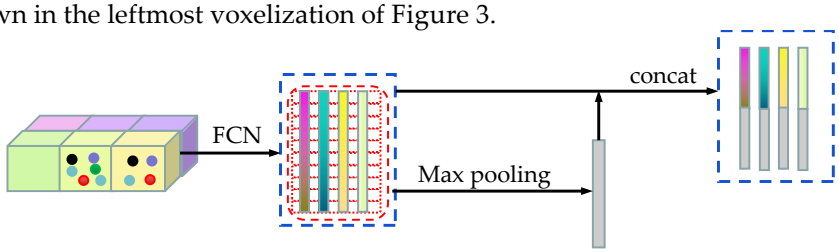

Figure 3. Point clouds voxelization and feature extraction.

In this paper, we address the issue of a memory burden on the computational platform that arises from processing the high-precision LiDAR point cloud, which consists of millions of points. Directly processing these point clouds can overwhelm the computational platform due to the high number of points. Another challenge is the variable density of the point cloud throughout the spatial height, which can impact the accuracy of the detection results. We propose randomly extracting a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

fixed number of point clouds in each voxel to mitigate these issues. This strategy reduces the memory burden on the computational platform and helps balance the voxel distribution, resulting in improved training diversification.

4.2 Point Cloud Feature Extraction

The point cloud feature extraction network, which is shown in Figure 3, first $p_i = [x_i, y_i, z_i, r_i]^T$ represents the coordinates of each point in the voxel. The voxel has four elements that represent the X, Y, and Z coordinates and reflectance, respectively. Before feature extraction, the initial feature of each point cloud is represented by the point coordinates and the centre position relative to the point coordinates. This initial feature can be expressed as $p_i = [x_i, y_i, z_i, r_i, x_i - v_x, y_i - v_y, z_i - v_z]^T$, where v_x, v_y, v_z denotes the centre position coordinates of voxels. The features of the points inside each voxel are then extracted by the FCN feature extraction network. After the feature extraction of each point inside the voxel is completed, the features are extracted as the voxel features using the maximum pooling method in the channel corresponding to each point. Finally, the extracted point features and voxel features of each voxel are spliced together as the final features. The FCN network comprises a linear layer, a batch normalization (BN) layer, and an activation function layer (ReLU). All the non-empty voxels are encoded in the same form and share all the parameter sets in the FCN network. Through the FCN structure, the input point cloud data is transformed into high-dimensional features. This structure encodes point interactions within voxels, allowing the final feature representation to learn to describe shape information. Therefore, point cloud features are extracted by stacking three layers of this structure in a point cloud feature extraction network.

5. Multimodal Fusion

This paper proposes two fusion techniques to improve the performance of 3D target detection in infrastructure view by extending the VoxelNet framework to fuse point cloud and image data. As mentioned, the VoxelNet[12] model is based on a single modality. However, this study enhances it by adding a multimodal fusion scheme, improving the network's performance.

5.1 Regional Point Fusion

This early fusion technique utilizes image features to aggregate 3D point clouds, enhancing the contextual information, as illustrated in Figure 4.

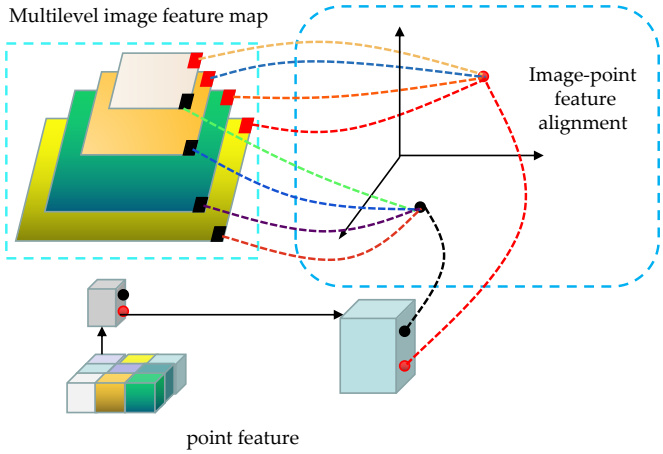

Figure 4. The framework of regional point fusion.

The method initially employs ResNet50 and FPN structure extraction networks to extract high-level feature maps from images with multi-level semantic encoding. These feature maps are then utilized to project each 3D point onto different layers' feature maps using a calibration matrix, thereby identifying the corresponding position on the image for each point. Subsequently, a 3×3 convolution is applied to extract centralized features of the small region associated with each point. These features are combined with the point cloud features derived from the previous feature extraction network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

Following the splicing of features, a set of FCN layer network structures is employed for further processing. Ultimately, these processed features are employed in the subsequent detection stage.

The ability to connect multi-level image features to point cloud features at an early stage is the advantage of this approach. This provides information on the location of the image corresponding to each point, as well as features within a small area of that point. Subsequently, the network can learn useful information in both modalities through the FCN layer.

5.2 Regional Voxel Fusion

Region voxel fusion employs a relatively late fusion strategy compared to region point fusion features. After the 3D convolution has extracted the features, the 3D space is transformed into a 2D space. Subsequently, an expansion convolution with an expansion coefficient of 1 is used to further the fusion of a larger range of information to increase the sensory field. After region point fusion, semantic features of the image are attached to the voxel level. Each voxel contains both point cloud and image features within it. In order to fully consider the information interaction between voxel contexts, this paper adopts three 3D convolutions for the fusion and extraction of regional features on 3D information.

Regional voxel fusion is a relatively late fusion strategy but offers certain advantages. Firstly, this approach enables the fusion of features projected from the image onto the point cloud at the voxel level, thereby enhancing the combination of feature information in multiple regions near the point cloud. Secondly, expanding convolution can enhance the receptive field and facilitate the detection of smaller objects.

6. Fusion Feature Extraction and Attention Mechanism

6.1 Coordinate Attention Mechanism

After fusing point cloud and image information, this paper utilizes the Coordinate attention mechanism for enhancing the feature map processing, as depicted in Figure 5.

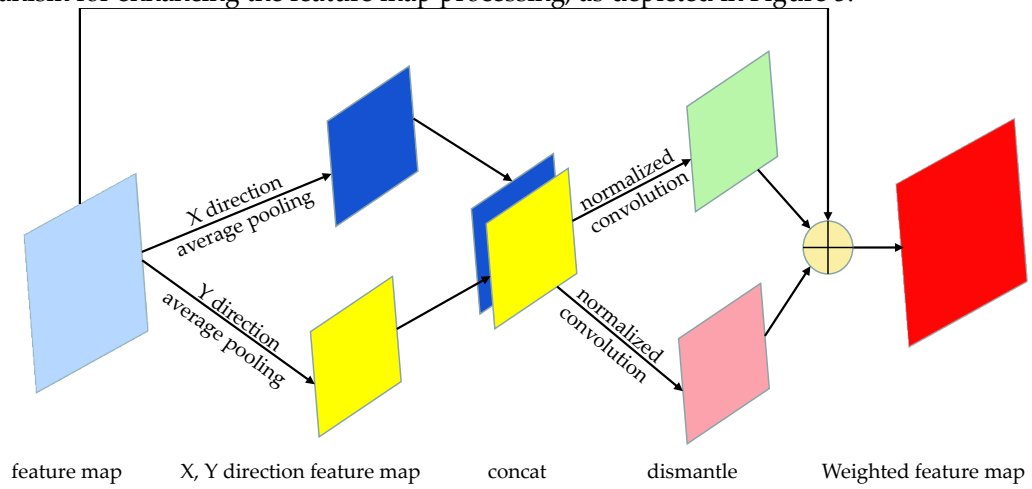

Figure 5. The schematic diagram of coordinated attention mechanism.

In this paper, an attention mechanism is proposed that takes into account both channel aspect information and position information in the feature map's horizontal and vertical dimensions. To ensure that spatial information is not compressed into the channel and to enable spatial interaction of the captured information, average pooling is performed separately in the X-axis and Y-axis directions. The pooled results are then concatenated and subjected to convolution operation, allowing interactions between the positions in the two axes. Once the information interaction is completed, the respective weights are sparsely computed along the two axes, resulting in a final feature that retains both positional and channel-specific information.

6.2 SimAM Attention Mechanism

After completing the fusion feature extraction with the Coordinate attention mechanism, this paper focuses on utilizing the FPN structure in the point cloud for two-level feature extraction. This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

approach aims to ensure that the deeper-level feature map contains richer semantic information, while the shallow-level feature map preserves more complete geometric details. Figure 6 demonstrates the SimAM parameter-free attention mechanism employed in this study to better assess the significance of each neuron in the network. This mechanism accomplishes the differentiation by defining the energy function's form, where neurons with higher energy functions are assigned greater weights due to their increased importance, while those with lower energy functions are assigned lower weights.

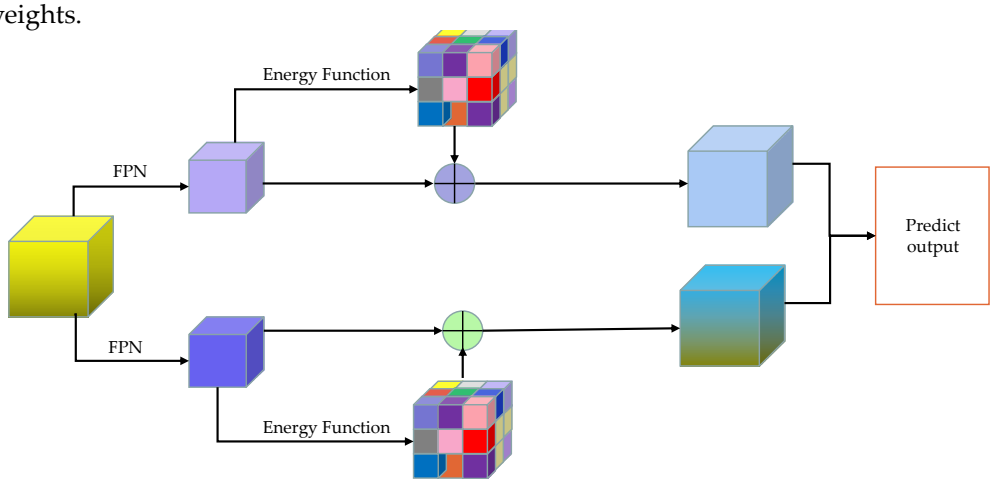

Figure 6. The schematic diagram of SimAM attention mechanism.

7. Loss Functions

This paper firstly parameterizes the 3D truth frame as follows: $(x_c^g, y_c^g, z_c^g, l^g, w^g, h^g, \theta^g)$. x_c^g, y_c^g, z_c^g represent the centre of the 3D truth frame, l^g, ω^g, h^g represent the length, width, and height of the 3D truth frame, and θ^g represent the rotation angle along the Z-axis. At the same time, the paper parameterizes the Anchor designed by ourselves in the target detection as $(x_c^a, y_c^a, z_c^a, l^a, w^a, h^a, \theta^a)$, and it defines seven residual coefficients of regression. These coefficients represent the offset relative to the centre coordinate, the elongation or shortening ratio of the three dimensions of length, width, and height, as well as the deviation around the direction of the Z-axis. The calculations of these seven coefficients are shown as follows.

$$\begin{aligned} &\Delta x = \frac{x_c^g - x_c^a}{d^a}, \Delta y = \frac{y_c^g - y_c^a}{d^a}, \Delta z = \frac{z_c^g - z_c^a}{h^a}, \\ &\Delta l = \log(\frac{l^g}{l^a}), \Delta w = \log(\frac{w^g}{w^a}),, \Delta h = \log(\frac{h^g}{h^a}), \\ &\Delta\theta = \theta^g - \theta^a \end{aligned} \tag{1}$$

In the designed anchor frame, equation (1) l^a represents the diagonal length of the 3D frame base. In order to estimate the oriented 3D detection frame directly, Δx 、 Δy and the diagonal of the 3D frame d^a is normalized. This differs from the method provided by Li et al.[47]. Finally, the loss function is defined in this paper, as shown below.

$$\begin{aligned} L = \alpha \frac{1}{N_{pos}} \sum_i L_{cls}(p_i^{pos}, 1) + \beta \frac{1}{N_{neg}} \sum_j L_{cls}(p_j^{neg}, 0) + \\ \frac{1}{N_{pos}} \sum_i L_{reg}(u_i, u_i^*) \end{aligned} \tag{2}$$

In equation (2), N_{pos} represents the number of positive anchor frames and N_{neg} represents the number of negative anchor frames. p_i^{pos} represents the probability that the i th anchor frame is predicted to be a true label, and p_j^{neg} represents the probability that the j anchor frame is predicted to be a false label. u_i represents the bounding box regression parameter of the i th anchor box, and u_i^* represents the bounding box regression parameter of the i th anchor box

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

corresponding to the truth box. L_{cls} denotes cross-entropy loss, α , β denotes balanced positive and negative sample parameters, and L_{reg} denotes regression parameters.

8. Training Details

8.1 Datasets

In this paper, we propose a multimodal fusion with multiple attention mechanisms for 3D target detection algorithm. The algorithm will be evaluated on the DAIR-V2X dataset, which contains 15,285 image data and 15,285 frames of point cloud data. The dataset is further decomposed into a training set and a validation set in the ratio of 7:3. After the split, the training set consists of 10700 samples, and the validation set is 4,585. The evaluation will analyze the effectiveness of the proposed multimodal approach by comparing it with previously published methods for 3D target detection tasks. The evaluation will consider three difficulty levels, easy, medium, and hard, based on object size, visibility (occlusion), and truncation.

8.2 Data Enhancement

This paper aims to address the issue of overfitting during the network training process by enhancing the data in two aspects, namely image and point cloud. Regarding images, we first scaled the size of the image into two sizes (640,192) and (2560,768) and then took [103.53,116.28,123.675] as the average value. Take [1.0,1.0,1.0] as the variance to regularize, and finally flip the horizontal side with a flip ratio of 0.5. For the point cloud, it is flipped in a specified angle range and scaled in the range of [0.95,1.05]. Like the image, the point cloud is also flipped horizontally with a flip ratio 0.5. Moreover, global scaling is applied to all truth boxes bi and the entire point cloud M. Specifically, the XYZ coordinates and 3D space of each bi, as well as the XYZ coordinates of all points in M, are multiplied by a random variable uniformly distributed in [0.95,1.05]. This paper also introduces global scaling in the image-based classification [48] and detection task [13] to enhance the network's ability to detect objects of different sizes and distances, thereby improving its overall robustness.

8.3 Experimental Parameters

8.3.1 Image Detection Networks

This paper uses the ResNet50+FPN structure for feature extraction in image target detection. The training dataset uses the DAIR-V2X dataset with data augmentation. In the training process, the image's shortest edge was rescaled to 600 pixels. The paper uses four scale anchors {4,8,16,32} and three aspect ratios {0.5,1,2} on the final output layer. Anchor points were labelled as positive anchor points if the intersection and concurrency ratio (IOU) with the ground truth frame was greater than 0.7 and negative anchor points if the IOU was less than 0.3. The network was trained using stochastic gradient descent with a learning rate 0.0005 and a momentum of 0.9. After completing the training of the image detection network, the training parameters of the image network were frozen, and the weight coefficients of the image backbone network were kept unchanged during the multimodal fusion training process.

8.3.2 Point Cloud Detection Networks

This paper primarily adopts the VoxelNet architecture for point cloud detection. The range of consideration for the point clouds is within [-3,1]×[-40,40]×[0,70.4] meters along the Z, Y, and X axes, respectively. To ensure accuracy, points projected outside the image boundary are eliminated[25]. We choose the voxel size of v_D = 0.05, v_H = 0.05, v_W = 0.1 m, and get D'= 80, H' = 1600, W'= 1408. Let T = 35 be the maximum number of random sampling points in each non-empty voxel. Furthermore, two FCN layers, FCN-1(7,32) and FCN-2(32,128) are employed. The final FCN maps the VFE-2 output to R128. Subsequently, feature extraction between the point clouds is achieved using three conv3D convolutional fusion.

8.3.3 Multimodal Fusion

Two 128-dimensional FCN modules extract the features after fusion when the image features are projected correspondingly to each point. The information after point fusion is then voxel fused using 3-layer conv3D convolution. Eventually, the RPN structure outputs the final result. For training

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

purposes, this paper utilizes the RTX3060 device. The Adam optimizer is employed with a learning rate 0.0003, a weight decay coefficient of 0.01, and a momentum parameter beta of (0.95, 0.99).

9. Experimental Results and Discussion

9.1 DAIR-V2X Dataset Evaluation

This paper evaluates the detection performance using the standard DAIR-V2X evaluation protocol (IOU=0.7 and IOU=0.5 for vehicles, and IOU=0.5 and IOU=0.25 for cyclist and pedestrians). In Tables 1 to 4, the algorithms proposed in this paper are compared with commonly used algorithms in terms of AP in both 3D and Bird's Eye View (BEV) views. The results demonstrate that multimodal fusion with multiple attention mechanisms for 3D target detection algorithm proposed in this paper significantly improves the detection performance compared to the commonly used algorithms. Notably, the effectiveness of fusion is more apparent in 3D view scores than in BEV view scores. It is also worth mentioning that the proposed fusion technique outperforms the original voxel networks with more powerful RPNs and the use of additional data enhancements. Furthermore, the method consistently outperforms other recent top-performing methods[49–52]. Example detection results of the proposed method are illustrated in Figure 7.

Table 1. Performance comparison of AP value algorithms with IOU=0.5 and IOU=0.25 from the BEV perspective.

method	car			pedestrian			cyclist		
	IOU=0.5			IOU=0.25			IOU=0.25		
	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
SECOND[49]	85.5	86.1	83.9	73.3	64.1	63.6	70.3	65.8	64.3
PointPillars[50]	85.5	83.9	83.7	66.9	55.3	54.8	62.8	59.8	58.7
Part-A2[51]	85.1	85.4	83.4	65.9	59.8	59.3	66.6	61.7	60.5
3D-SSD[52]	84.2	83.9	82.0	N/A	N/A	N/A	N/A	N/A	N/A
Ours	**88.4**	**86.6**	**86.5**	**81.5**	**71.0**	**70.7**	**71.6**	**70.8**	**70.9**

Table 2. Performance comparison of AP value algorithms with IOU=0.7 and IOU=0.5 from the BEV perspective.

method	car			pedestrian			cyclist		
	IOU=0.5			IOU=0.25			IOU=0.25		
	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
SECOND[49]	81.9	83.0	83.6	57.5	49.3	48.9	65.0	59.6	58.7
PointPillars[50]	82.0	80.9	78.4	57.7	46.3	45.8	50.5	55.1	53.4
Part-A2[51]	80.2	80.7	79.7	55.7	47.6	47.1	62.8	56.1	55.2
3D-SSD[52]	79.3	79.4	79.5	N/A	N/A	N/A	N/A	N/A	N/A
Ours	**84.8**	**83.5**	**83.8**	**70.2**	**60.2**	**59.9**	**66.0**	**64.4**	**63.1**

Table 3. Performance comparison of AP value algorithms with IOU=0.5 and IOU=0.25 from the 3D perspective.

method	car	pedestrian	cyclist

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

	IOU=0.5			IOU=0.25			IOU=0.25		
	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
SECOND[49]	82.9	83.6	83.3	72.9	63.6	62.4	70.2	65.6	64.3
PointPillars[50]	83.0	83.5	81.1	66.7	55.0	54.1	62.7	59.8	57.9
Part-A2[51]	84.2	83.2	82.4	65.5	59.5	59.0	65.9	60.9	60.2
3D-SSD[52]	82.3	81.9	72.5	N/A	N/A	N/A	N/A	N/A	N/A
Ours	**87.8**	**86.2**	**83.8**	**81.1**	**70.6**	**70.3**	**70.5**	**70.6**	**70.7**

Table 4. Performance comparison of AP value algorithms with IOU=0.7 and IOU=0.5 from the 3D perspective.

method	car			pedestrian			cyclist		
	IOU=0.5			IOU=0.25			IOU=0.25		
	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
SECOND[49]	75.8	73.9	70.9	52.8	43.4	42.9	63.2	56.5	55.7
PointPillars[50]	75.6	73.6	70.8	52.5	40.6	40.1	56.5	50.8	49.0
Part-A2[51]	77.1	74.7	71.5	52.8	41.8	41.5	61.6	54.0	52.4
3D-SSD[52]	76.1	71.0	65.7	N/A	N/A	N/A	N/A	N/A	N/A
Ours	**79.9**	**76.1**	**73.5**	**62.9**	**51.7**	**51.4**	**63.4**	**58.7**	**57.6**

Figure 7. Diagram of algorithm test results.

9.2 Performance Analysis of the AP Value Algorithm

9.2.1 Vehicle Target Detection Analysis

The algorithm proposed in this paper has been found to perform best compared to other algorithms at this stage, as indicated by Table 1-Table 4. In the case of BEV view with an IOU of 0.5, the AP value has increased by 2.9%, 0.5%, and 2.4% over the SECOND algorithm, which is currently the best performing algorithm, in the easy, medium, and difficult modes, respectively (Table 1). Furthermore, in the case of the BEV view with an IOU of 0.7, the AP values have improved by 2.9%, 0.5%, and 0.2% in the easy, medium, and difficult modes, respectively (Table 2). Similarly, in the case of the 3D view with an IOU of 0.5, the AP value has increased by 4.9%, 2.6%, and 0.5% in the easy, medium, and hard modes, respectively (Table 3). Lastly, Table 4 shows that in the case of 3D view with an IOU of 0.7, the AP values have increased by 4.1%, 2.2%, and 2.6% in the easy, medium, and hard modes, respectively.

9.2.2 Pedestrian Target Detection Analysis

The algorithm proposed in this paper demonstrates superior performance compared to existing pedestrian target detection algorithms, as indicated by Table 1-Table 4. Table 1 presents results for the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2023 doi:10.20944/preprints202307.1956.v1

BEV view with an IOU of 0.5, showing an improvement in the AP value of 8.2%, 6.9%, and 7.1% in the easy, medium, and difficult modes, respectively, compared to the SECOND algorithm, which is currently the leading algorithm in this field. Table 2 reveals that for the BEV view with an IOU of 0.7, there are improvements in the AP values of 2.7%, 10.9%, and 11% in the easy, medium, and hard modes, respectively. Similarly, Table 3 presents results for the 3D view with an IOU of 0.5, indicating an increase in the AP value of 8.2%, 7.0%, and 7.9% in the easy, medium, and hard modes, respectively. Finally, Table 4 demonstrates that in the case of the 3D view with an IOU of 0.7, there is an increase in the AP value of 10.1%, 8.3%, and 8.5% in the easy, medium, and hard modes, respectively.

9.2.3 Cyclist Target Detection Analysis

The algorithm proposed in this paper is the best-performing algorithm compared to other excellent cyclist target detection algorithms, as indicated by Table 1 to Table 4. In Table 1, when looking at the BEV view with IOU=0.5, the AP value shows an increase of 1.3%, 5%, and 6.6% in the easy, medium, and difficult modes, respectively, compared to the SECOND algorithm, which is also considered an excellent performer at this stage. Similarly, Table 2 illustrates that when considering the BEV view with IOU = 0.7, the AP values improve correspondingly by 1%, 4.8%, and 4.4% in the easy, medium, and difficult modes. In Table 3, the case of the 3D view with IOU=0.5 shows an increase of 0.3%, 5.0%, and 6.4% in the easy, medium, and hard modes, respectively. Lastly, Table 4 indicates that in the case of the 3D view with IOU=0.7, the AP value increases by 0.2%, 2.2%, and 1.9% in the easy, medium, and hard modes, respectively.

9.3 Stability Analysis

The stability of the algorithm is represented by calculating the mAP value of each method in this paper. A relatively high mAP value indicates better performance in the detection of vehicles, pedestrians, and bicyclists, making the algorithm suitable for multi-size multi-target detections. Conversely, a low mAP value suggests that the algorithm is less effective in detecting one of the objects in the target detection. In Table 5, it can be observed that the proposed algorithm in this paper exhibits a 7.9% improvement in mAP value in the BEV view and a 7.8% improvement in the 3D view compared to the Part-A2 network, which is an exceptional performer among the compared algorithms, with an IOU of 0.5 for cars and 0.25 for pedestrians and bicycles. Additionally, Table 6 demonstrates that when the IOU is set to 0.7 for cars and 0.5 for pedestrians and bicycles, the proposed algorithm shows a 5.4% improvement in the mAP value in the BEV view and a 4.3% improvement in the 3D view, compared to the SECOND algorithm, which is another remarkable performer among the compared algorithms.

Table 4. mAP results at IOU=0.5 and IOU=0.25.

method	IOU=0.5(car) IOU=0.25（pedestrian、cyclist）	
	BEV	3D
SECOND[49]	66.0	72.1
PointPillars[50]	67.9	67.1
Part-A2[51]	69.7	69.0
Ours	**77.6**	**76.8**

Table 5. mAP results at IOU=0.5 and IOU=0.25.

method	IOU=0.5(car) IOU=0.25（pedestrian、cyclist）	
	BEV	3D
SECOND[49]	65.3	59.4
PointPillars[50]	61.1	56.6
Part-A2[51]	62.9	58.6

Ours	**70.7**	**63.9**

10. Conclusion

The proposal presented in this paragraph is a multimodal fusion with multiple attention mechanisms for 3D Target Detection Algorithm aimed at addressing the limitations of single-modal target detection. The first step involves utilizing the ResNet50+FPN network framework to extract image features, resulting in the extraction of four-level features. Simultaneously, the point cloud feature extraction employs the voxel grid method and FCN to extract point features and voxel features from each voxel. These extracted features are then considered as the final features of the point cloud. Following this, regional point fusion and voxel fusion techniques are employed to combine the image and point cloud features. Once the fusion process is completed, the fused features undergo depth extraction using the SECOND network. Furthermore, the Coordinate attention mechanism and the SimAM attention mechanism are implemented during this stage. Finally, the RPN is applied to obtain the output. To validate and compare the proposed algorithm with other state-of-the-art algorithms, it is tested on the DAIR-V2X dataset. The results demonstrate that the proposed algorithm surpasses other algorithms in terms of detection performance.

Author Contributions: Conceptualization, X.Z.,Y.W. and L.H.; software, X.Z.; validation, J.C., B.W., data curation, Y.Z.; writing—original draft preparation, X.Z.; writing—review and editing, Y.W. and Y.Z.; visualization, J.C. and B.W.; supervision, L.H.; project administration, L.H.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Industry and Information Technology of Jilin Province, China under grant number TC210H02S.

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

References

1. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
2. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep Continuous Fusion for Multi-sensor 3D Object Detection. In Proceedings of the Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, 8–14 September 2018; Volume 11220, pp. 663–678. [CrossRef]
3. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection From Point Clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
4. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
5. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
6. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
7. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
8. Yang, B.; Luo, W.; Urtasun, R. PIXOR: Real-Time 3D Object Detection From Point Clouds. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7652–7660. [CrossRef]
9. Wang, Z.; Jia, K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1742–1749.
10. Garcia-Garcia, A.; Gomez-Donoso, F.; Garcia-Rodriguez, J.; Orts-Escolano, S.; Cazorla, M.; Azorin-Lopez, J. Pointnet: A 3d Convolutional Neural Network for Real-Time Object Class Recognition. In Proceedings of the 2016 International Joint Conference on Neural Networks, Vancouver, BC, Canada, 24–29 July 2016.
11. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85. [CrossRef]
12. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499. [CrossRef]
13. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
14. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
15. Shi, S.; Wang, Z.; Wang, X.; Li, H. Part-a2 net: 3d part-aware and aggregation neural network for object detection from point cloud. arXiv 2019, arXiv:1907.03670.
16. Song, S.; Xiao, J. Deep sliding shapes for amodal 3D object detection in rgb-d images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 808–816.
17. Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Data-driven 3d voxel patterns for object category recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1903–1911.
18. Xue, B.; Tong, N. Real-World ISAR Object Recognition Using Deep Multimodal Relation Learning. IEEE Trans. Cybern. 2020, 50, 4256–4267. [CrossRef]
19. Timofte, R.; Zimmermann, K.; Van Gool, L. Multi-View Traffific Sign Detection, Recognition, and 3D Localisation. Mach. Vis. Appl. 2014, 25, 633–647. [CrossRef]
20. Li, B.; Ouyang, W.; Sheng, L.; Zeng, X.; Wang, X. GS3D: An Effifficient 3D Object Detection Framework for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1019–1028.
21. Mousavian, A.; Anguelov, D.; Flynn, J.; Košecká, J. 3D bounding box estimation using deep learning and geometry. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5632–5640.
22. Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; Chateau, T. Deep MANTA: A Coarse-to-fifine Many-task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2040–2049.
23. Zhu, M.; Derpanis, K.G.; Yang, Y.; Brahmbhatt, S.; Zhang, M.; Phillips, C.; Lecce, M.; Daniilidis, K. Single Image 3D Object Detection and Pose Estimation for Grasping. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA 2014), Hong Kong, China, 31 May–7 June 2014.
24. Manhardt, F.; Kehl, W.; Gaidon, A. ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2069–2078.
25. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2147–2156.
26. Chen, X.; Kundu, K.; Zhu, Y.; Berneshawi, A.G.; Ma, H.; Fidler, S.; Urtasun, R. 3d object proposals for accurate object class detection. In Advances in Neural Information Processing Systems; NIPS: San Diego, CA, USA, 2015; pp. 424–432.
27. Li, P.; Chen, X.; Shen, S. Stereo r-cnn based 3d object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7644–7652
28. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8437–8445. [CrossRef]
29. Chua, C.S.; Jarvis, R. Point signatures: A new representation for 3d object recognition. Int. J. Comput. Vis. 1997, 25, 63–85. [CrossRef]
30. Ba, J.; Mnih, V.; Kavukcuoglu, K.J. Multiple object recognition with visual attention. arXiv 2014, arXiv:1412.7755.
31. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3d object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artifificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11677–11684.
32. Dorai, C.; Jain, A.K. COSMOS-A representation scheme for 3D free-form objects. IEEE Trans. Pattern Anal. Mach. Intell. 1997, 19, 1115–1130. [CrossRef]
33. Tuzel, O.; Liu, M.-Y.; Taguchi, Y.; Raghunathan, A. Learning to rank 3d features. In Proceedings of the

European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part I. pp. 520–535.
34. Wang, D.Z.; Posner, I. Voting for voting in online point cloud object detection. In Proceedings of the Robotics: Science and Systems, Rome, Italy, 13–17 July 2015; pp. 10–15607.
35. Engelcke, M.; Rao, D.; Wang, D.Z.; Tong, C.H.; Posner, I. Vote3deep: Fast object detection in 3d point clouds using effificient convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Marina Bay Sands Convention Centre, Singapore, 29 May–3 June 2017; pp. 1355–1361.
36. Li, B. 3d fully convolutional network for vehicle detection in point cloud. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, September 24–28 2017; pp. 1513–1518.
37. Premebida, C.; Carreira, J.; Batista, J.; Nunes, U. Pedestrian detection combining rgb and dense lidar data. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 4112–4117.
38. Gonzalez, A.; Villalonga, G.; Xu, J.; Vazquez, D.; Amores, J.; Lopez, A. Multiview random forest of local experts combining rgb and lidar data for pedestrian detection. IEEE Intell. Veh. Symp. (IV) 2015, 356–361. [CrossRef]
39. Wang, Z.; Jia, K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1742–1749.
40. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum pointnets for 3d object detection from rgb-d data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 918–927.
41. Zhang, Z.H.; Liang, Z.D.; Zhang, M.; Zhao, X.; Li, H.; Yang, M.; Tan, W.M.; Pu, S.L. RangeLVDet: Boosting 3D Object Detection in LIDAR with Range Image and RGB Image. IEEE Sens. J. 2022, 22, 1391–1403. [CrossRef]
42. Premebida, C.; Carreira, J.; Batista, J.; Nunes, U. Pedestrian detection combining rgb and dense lidar data. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 4112–4117.
43. Zhang Y, Wu H. 3D Object Detection Based on Multi-view Adaptive Fusion[C]//2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). IEEE, 2022: pp.743-748.
44. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S. Joint 3D Proposal Generation and Object Detection from View Aggregation. arXiv 2017, arXiv:1712.02294.
45. Rozenberszki D, Litany O, Dai A. UnScene3D: Unsupervised 3D Instance Segmentation for Indoor Scenes[J]. arXiv preprint arXiv:2303.14541, 2023.
46. 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.
47. Li, B.; Zhang, T.; Xia, T. Vehicle detection from 3D lidar using fully convolutional network. arXiv 2016, arXiv:1608.07916.
48. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2014, arXiv:1409.1556.
49. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. Sensors 2018, 18, 3337. [CrossRef]
50. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
51. Shi, S.; Wang, Z.; Wang, X.; Li, H. Part-a2 net: 3d part-aware and aggregation neural network for object detection from point cloud. arXiv 2019, arXiv:1907.03670.
52. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11873–11882.