signals

MDPI

Article

3D Object Detection Using Frustums and Attention Modules for Images and Point Clouds

Yiran Li, Han Xie and Hyunchul Shin *

Department of Electrical Engineering, Hanyang University, Ansan 15588, Korea; liyiran07070@hanyang.ac.kr (Y.L.); xiehan@hanyang.ac.kr (H.X.)
* Correspondence: shin@hanyang.ac.kr

Abstract: Three-dimensional (3D) object detection is essential in autonomous driving. Three-dimensional (3D) Lidar sensor can capture three-dimensional objects, such as vehicles, cycles, pedestrians, and other objects on the road. Although Lidar can generate point clouds in 3D space, it still lacks the fine resolution of 2D information. Therefore, Lidar and camera fusion has gradually become a practical method for 3D object detection. Previous strategies focused on the extraction of voxel points and the fusion of feature maps. However, the biggest challenge is in extracting enough edge information to detect small objects. To solve this problem, we found that attention modules are beneficial in detecting small objects. In this work, we developed Frustum ConvNet and attention modules for the fusion of images from a camera and point clouds from a Lidar. Multilayer Perceptron (MLP) and tanh activation functions were used in the attention modules. Furthermore, the attention modules were designed on PointNet to perform multilayer edge detection for 3D object detection. Compared with a previous well-known method, Frustum ConvNet, our method achieved competitive results, with an improvement of 0.27%, 0.43%, and 0.36% in Average Precision (AP) for 3D object detection in easy, moderate, and hard cases, respectively, and an improvement of 0.21%, 0.27%, and 0.01% in AP for Bird's Eye View (BEV) object detection in easy, moderate, and hard cases, respectively, on the KITTI detection benchmarks. Our method also obtained the best results in four cases in AP on the indoor SUN-RGBD dataset for 3D object detection.

Keywords: 3D vision; attention module; fusion; point cloud; vehicle detection

Citation: Li, Y.; Xie, H.; Shin, H. 3D Object Detection Using Frustums and Attention Modules for Images and Point Clouds. *Signals* **2021**, *2*, 98–107. https://doi.org/10.3390/signals2010009

Received: 21 September 2020
Accepted: 6 January 2021
Published: 12 February 2021

1. Introduction

The detection of object instances in 3D sensory data has tremendous importance in many applications. Three-dimensional (3D) technology can receive more abundant and comprehensive environmental information. Therefore, it is widely used in robot navigation, automatic driving, Augmented Reality (AR), and industrial detection.

Point cloud and RGB image fusion can simultaneously extract 2D and 3D features by using a neural network. Objects can be detected with higher accuracy by simultaneously considering 2D and 3D information. With the progress of point clouds [1,2], 3D object detection methods [3,4] can resort to learning features directly from point clouds. For example, PointNets [3,4] are capable of classifying a whole point cloud or predicting a semantic class for each point in the point clouds. Three-dimensional (3D) point clouds are usually transformed into images or voxel grids [5] before PointNet [3,4]. It shows good performance in 3D object detection. However, the weakness of PointNet [3,4] and PointNet++ [4] is that a 3D bounding box cannot be estimated with direction. A new Frustum scheme was proposed by F-PointNets [2] and Frustum ConvNet [1], which use RGB-D data and a multilayer 2D region proposal to help the point clouds' segmentation form the 3D space. The global features are obtained from the local feature combination. F-PointNets used T-net [2] to determine the position and direction of a 3D bounding box. The disadvantage of the Frustum is that objects with unclear boundaries and small-scale instances are difficult to detect.

To solve this problem, we would like to refer to the attention modules used in 2D object detection methods. Guo et al. [5] proposed a method based on a Gaussian Mixture Model (GMM), in which attention modules were improved by colour, intensity, and orientation feature maps. The attention modules focused on interesting areas to enhance the features of the edge information and small objects. Fan et al. [6] proposed a Region Proposal Network (RPN) with an attention module, enabling the detector to pay attention to objects with high resolution while perceiving the surroundings with low resolution. These works inspired us to use attention modules for object detection in 3D point clouds.

In our work, we developed the images and 3D point clouds fusion method to improve 3D object detection. A new attention module was designed with Frustum ConvNet [1] to enhance feature extraction and improve small object detection. We added attention modules to the input layer of Multilayer Perceptron (MLP) in PointNet [3,4]. We used the tanh activation function to extract and strengthen the attention of a small object, which can improve small object detection effectively.

In this paper, we propose a Frustum ConvNet with attention modules for 3D object detection, in which both images and point clouds are used. The contributions of the paper are as follows:

1. In the PointNet of Frustum ConvNet, we added the Convolutional Block Attention Module (CBAM) [7] attention module at the hidden layer of Multilayer Perceptron (MLP) to improve the detection accuracy. The CBAM attention module can improve the contrast between the object and the surrounding environment.
2. We propose an improved attention module by adding Multilayer Perceptron (MLP) and using the tanh activation function. The tanh function is used for average-pooling and max-pooling layers to extract features. The mean of the tanh activation function is 0. Furthermore, the tanh function can cope with cases when the feature values have big differences. Finally, the feature information of the pooling layers is fused through the sigmoid function.
3. We evaluated our approach on the KITTI [8] object detection benchmark. Compared with the state-of-the-art method, our method achieved competitive results, with an improvement of 0.21%, 0.27%, and 0.01% in Average Precision (AP) for 3D object detection in easy, moderate, and hard cases, respectively, and an improvement of 0.27%, 0.43%, and 0.36% in AP for Bird's Eye View (BEV) object detection in easy, moderate, and hard cases, respectively, on KITTI detection benchmarks. Our method also obtains the best results in four cases in AP on the indoor SUN-RGBD [9] dataset for 3D object detection.

The rest of this paper is organized as follows. Section 2 introduces the previous 3D object detection methods. Section 3 describes the architecture of Frustum ConvNet with Attention Module (FCAM). Section 4 presents the results of our experiments. We conclude in Section 5.

2. Related Works

This section briefly introduces the previous 3D object detection methods and related attention works. We organize our reviews into three categories of technical approaches, namely 3D object detection techniques from point clouds, attention modules in object detection, and activation functions in a neural network:

2.1. Three-Dimensional (3D) Object Detection from Point Clouds

Three-dimensional (3D) voxel patterns (3DVPs) [10] employ a set of Aggregate Channel Feature (ACF) [11] detectors to perform 2D detection and 3D pose estimation. A Multiview 3D Object Detection Network (MV3D) [12] proposed a sensory-fusion framework that takes both Lidar point clouds and RGB images as inputs and predicts oriented 3D bounding boxes. Different from the MV3Ds [12], Li et al. [13] and Song et al. [14] converted the features in point clouds into a voxel grid to improve accuracy at the cost of a large amount of computation. VoxelNet [15] proposed a generic 3D detection network

that unifies feature extraction and bounding box prediction into a single-stage, end-to-end trainable deep network. In this method, 3D object detection can operate directly on sparse 3D points and capture 3D shape information effectively.

2.2. Attention Module in Object Detection

Recently, some methods have been put forward to incorporate attention processing to improve the performance of CNNs in 2D-based large-scale classification tasks. Wang et al. [16] proposed a Residual Attention Network, which can incorporate state-of-the-art feed-forward network architecture in an end-to-end training fashion. This network can extract a large amount of attention information without interruption. Hu et al. [17] introduced a Squeeze-and-Excitation module that adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. This method has an improvement in calculation and speed. The Bottleneck Attention Module (BAM) [18] and Convolutional Block Attention Module (CBAM) [7] added spatial attention to increase accuracy. These attention models performed well for 2D object detection.

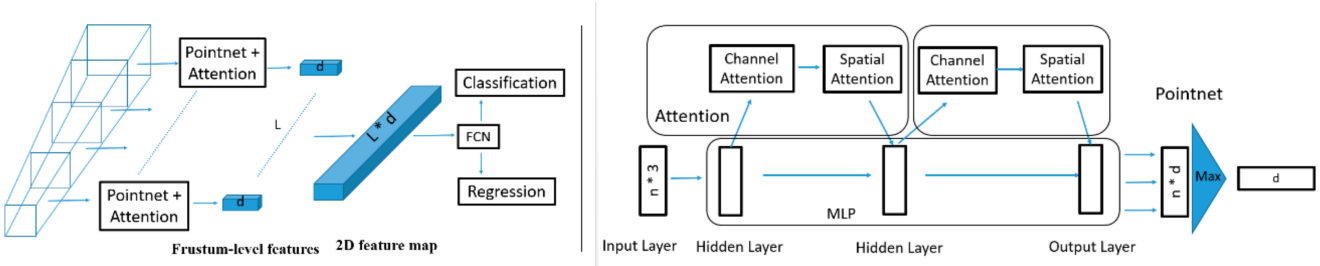
2.3. Activation Function in Neural Network

There is now a consensus that for deep networks, rectified linear units (ReLUs) are easier to train than logistic or tanh units, which were used for many years [19,20]. However, Le et al. [21] noticed that ReLUs seem inappropriate for RNNs because of the possibility that large output values may explode out of the bounded values. Ang-bo et al. [22] noticed that tanh alleviates the phenomenon of mean shift. Li et al. [23] noticed that the output of the tanh function can enhance the values activated by ReLU units. This inspired us to use the fusion activation function in the 3D object detection network.

Based on the Frustum architecture [1,2] and the attention module [7], we developed a new 3D object detection network by integrating the attention modules with Frustum ConvNet. Based on the advantages of the two functions, we fuse the ReLUs and tanh functions in the attention module to achieve higher accuracy. Our proposed method achieved competitive results in KITTI detection benchmarks.

3. Frustum ConvNet with Attention (FCAM)

The architecture of our 3D object detection network using Frustums and attention modules is shown in Figure 1. This network connects discrete disordered points from Frustums by using Fully Convolutional Networks (FCNs) [24], thus achieving 3D box-oriented estimation in a continuous 3D space. We first describe the structure of Frustum ConvNet in Section 3.1. Frustum ConvNet [1,2] uses PointNet [1,2] to extract and to aggregate point-wise features as Frustum-level feature vectors. Section 3.2 describes our improved attention modules by adding Multilayer Perceptron (MLP) and using the tanh activation function in the PointNet architecture.



(**a**) Frustum ConvNet (**b**) PointNet with attention modules

Figure 1. The architecture of the proposed 3D object detector by using Frustum ConvNet with attention modules. (**a**) The whole framework of Frustum ConvNet; (**b**) the structure of PointNet with attention modules.

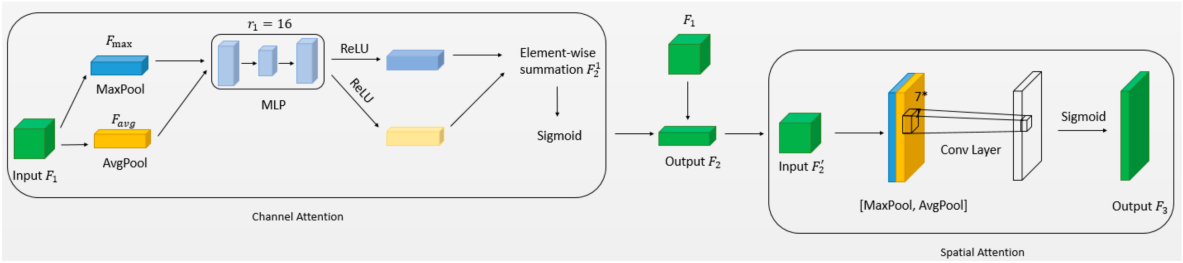
3.1. Frustum ConvNet

We designed the 3D object detector based on the framework of Frustum ConvNet, as shown in Figure 1a. At first, Frustum-level features are obtained from the Frustum through PointNet and attention modules, which are re-formed as a 2D feature map. Next, the PointNets are applied to each Frustum, and the PointNets with shared weights form the parallel streams of Frustum ConvNet [1]. For point cloud classification, the PointNet takes n points as the input, and the output comprises the classification scores for the d classes. The coordinates of the 3D space point clouds minus the centre coordinates of the Frustum to constitute a 1D vector and serve as the input of FCN. L-dimension vectors form a 2D feature map F with the size $L \times d$, which will be used as the input of a subsequent FCN [24]. Finally, the 2D feature maps are used as the input of Fully Convolutional Networks (FCNs) [24] for 3D prediction. Then, we use the detection head for classification and regression.

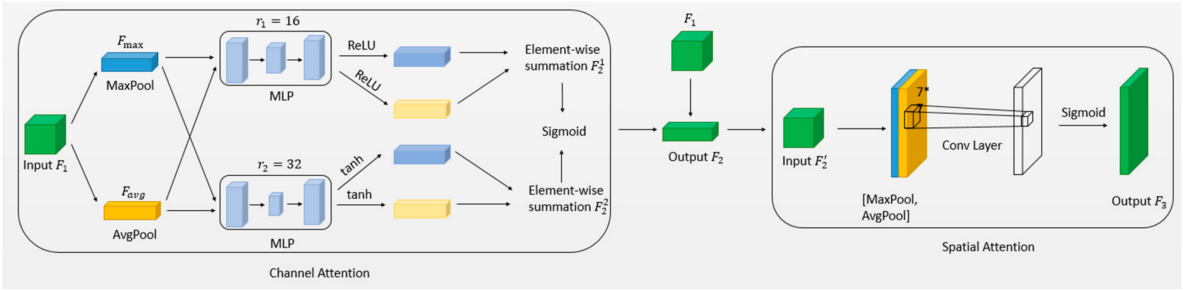
In the PointNet, we apply our improved Convolutional Block Attention Modules (CBAMs) in the hidden layer of Multilayer Perceptron (MLP), as shown in Figure 1b. Output features of the CBAM attention module are multiplied with the input feature of the MLP to obtain the final fused features.

3.2. The Improved CBAM Attention Model for Point Cloud Detection

The original CBAM attention model is shown in Figure 2a, and our improved CBAM attention model is shown in Figure 2b. In this section, we briefly introduce the original CBAM and then explain the improvement of our proposed attention model.



(**a**) Convolutional Block Attention Module (CBAM) architecture



(**b**) Improved CBAM architecture

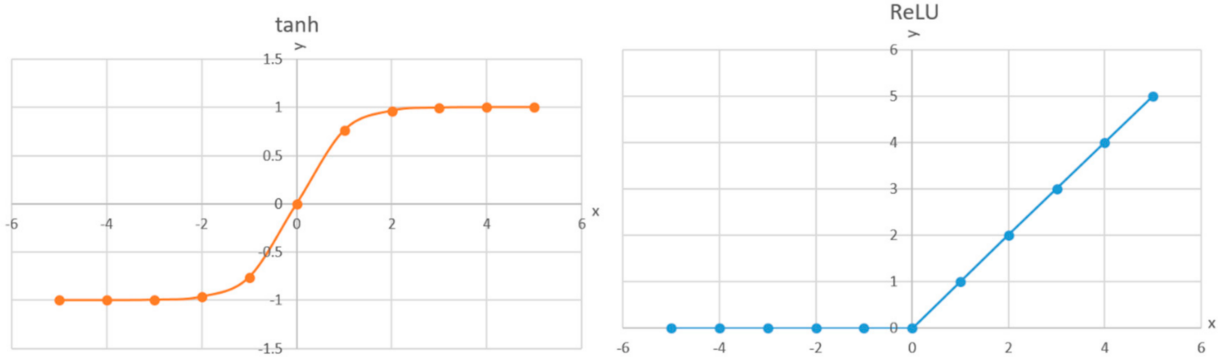
Figure 2. The architecture of attention modules. (**a**) The architecture of the original CBAM attention module [4]; (**b**) the architecture of our improved CBAM.

The original CBAM attention module consists of the channel attention and the spatial attention blocks. In the channel attention block, the input feature $\boldsymbol{F_1}$ passes max pooling and average pooling and then through the MLP with a reduction ratio $r_1 = 16$. They are followed by sigmoid activation to generate the final channel attention feature map. In the spatial attention block, channel-based global max pooling and global average pooling are connected on the input feature $\boldsymbol{F_2}$, and then $\boldsymbol{F_2}$ through a convolution layer with 7×7 kernel size. After a convolution layer, the dimension is reduced to one channel. Finally, the attention features are generated after calculation by the sigmoid function.

The challenge of 3D object detection using PointNet [3,4] is the feature missing problem, especially when the characteristics are quite different. In the whole Frustum ConvNet [1], when a 2D proposal is converted into a 1D vector, some feature information may be lost to some extent. To solve this problem, we designed a new kind of attention module to improve the detection capability of our proposed 3D object detector.

If an input falls into the region where x < 0, the gradient of the neuron becomes 0. This phenomenon is called the Dead ReLU problem [22]. It causes the regression of the model not to converge. To solve the Dead ReLU problem of the ReLU function and enhance the feature extraction ability of attention modules, the tanh function is used as an auxiliary optimization function in our new structure.

Based on the FCN [24] and the thought of the U-net [25] fusion, we added a parallel Multilayer Perceptron (MLP) architecture to the CBAM attention module and used the tanh activation function to enhance the contrast between the object and background. Multilayer Perceptron (MLP) can better fit the nonlinear region and performs well when dealing with deep networks and large amounts of information. To prevent overfitting due to too many parameters, we used different reduction ratios in two MLPs to reduce the input and output channels. Here, the reduction ratios are $r_1 = 16$ and $r_2 = 32$. Furthermore, the tanh function can alleviate the mean deviation problem in the ReLU function in [–1, 1], and the tanh function performs better when the feature values are quite different. However, the tanh function shows the gradient disappearance outside of [–1, 1], as shown in Figure 3a. The gradient disappearance problem can be solved by the ReLU function when x > 0, as shown in Figure 3b. To exploit the advantages of the two activation functions in the attention modules, we fused the two average pooling vectors by element-wise summation in the final step, as shown in Figure 3b. Because the sigmoid function has the scope of [0, 1], we used the sigmoid function as the output layer activation function to represent the prediction probability. It is used to filter the unimportant part and retain the important part feature information.



(**a**) The tanh function (**b**) The ReLU function

Figure 3. The activation function curves in attention modules.

In our improved CBAM attention module, feature maps of channel attention are obtained by using max pooling and average pooling. The feature map then passes through two MLPs consisting of two fully connected layers that use the ReLU activation function and tanh activation function, as shown in the channel attention block of Figure 2b. The output feature value $\boldsymbol{F_3}$ can be computed by using Equations (1)–(4), as follows:

$$\boldsymbol{W_{c1}}\left(\boldsymbol{F_2^1}\right) = \boldsymbol{W_1}(\boldsymbol{W_0}(\boldsymbol{F_{avg}}) + \boldsymbol{W_1}(\boldsymbol{W_0}(\boldsymbol{F_{max}})) \quad (1)$$

$$\boldsymbol{W_{c2}}\left(\boldsymbol{F_2^2}\right) = \boldsymbol{W_3}(\boldsymbol{W_2}(\boldsymbol{F_{avg}}) + \boldsymbol{W_3}(\boldsymbol{W_2}(\boldsymbol{F_{max}})) \quad (2)$$

$$\boldsymbol{F_2} = \mathit{Sigmoid}\left(\boldsymbol{W_{c1}}\left(\boldsymbol{F_2^1}\right) + \boldsymbol{W_{c2}}\left(\boldsymbol{F_2^2}\right)\right) \quad (3)$$

$$F_3 = Sigmoid\left(f^{7*7}[Maxpool; Avgpool]\right) \quad (4)$$

where $W_0 \epsilon \mathbb{R}^{\frac{C}{r_1} * c}$, $W_1 \epsilon \mathbb{R}^{C * \frac{C}{r_1}}$, $W_2 \epsilon \mathbb{R}^{\frac{C}{r_2} * c}$ and $W_3 \epsilon \mathbb{R}^{C * \frac{C}{r_2}}$. W_0, W_1, W_2, W_3 are the weights of the four fully connected layers in the two MLP networks. W_{c1}, W_{c2} are the weights of the F_2^1 and F_2^2. C is the number of channels. $\mathbb{R}$ is the real number field. F_2^2, F_2^1 are the element-wise summation features by two different MLP networks. f^{7*7} represents a convolution operation with a filter size of 7×7.

4. Experimental Results

We evaluated our 3D object detector on KITTI benchmarks [8] for 3D car detection. We performed the experiment on a TITAN X GPU and developed the code with PyTorch version 1.1. For each image, evaluation is required for 0.005 s. Our experiment was based on Frustum ConvNet [1] and tested the vehicles using the KITTI dataset [8]. We applied the attention module in PointNet and added a parallel Multilayer Perceptron (MLP) architecture in the CBAM attention module with the tanh activation function. The number of parameters in Frustum ConvNet [1] is 3,340,089, and the number of parameters in FCAM is 3,351,353. To reduce the network's model size and prevent overfitting, we used two attention modules and increased the reduction ratios from 16 to 32. A larger reduction ratio reduces parameter overhead and improves the speed of our method.

KITTI: The KITTI dataset [8] contains 7481 training pairs, 7518 testing pairs of RGB images, and corresponding point clouds. Following an existing work [15], we split the original training set into the new training and validation sets of 3712 and 3769 samples, respectively. Learning rates start from 0.001 and decay by a factor of 10 every 22nd epoch of the total 50 epochs.

Metrics: We evaluated 3D object proposals using 3D box recall as the metric. For 3D localization, we projected the 3D boxes to the ground plane. We used 3D object detection AP to evaluate the accuracy of the 3D object detection. We used BEV object detection AP to evaluate the accuracy of the BEV object detection.

We will now explain the experimental ablation results. Tables 1 and 2 show the detection performance of the improved CBAM on the KITTI validation set for 3D object detection and Bird's Eye View (BEV) object detection, respectively. T means the running time to process each image. Compared with Frustum ConvNet [1] + CBAM [7], our results showed improved AP by 0.09% in the BEV object detection of easy categories. For the 3D object detection, our results showed improved AP by 0.15%, 0.13% in moderate and hard categories. Because the attention model combines two activation functions, our attention module has slightly improved the accuracy, but the running time is longer than before. Figure 4 shows the convergence curves of Frustum, Frustum + CBAM, and Frustum + Improved CBAM. Frustum + Improved CBAM can achieve the highest accuracy on average among these three methods.

Table 1. Bird's Eye View (BEV) object detection Average Precision (AP) (%) on the KITTI dataset. The best result are in bold.

Method	T(s)	Easy	Moderate	Hard
Frustum ConvNet [1]	**0.002**	90.23	88.79	86.84
Frustum ConvNet [1] + CBAM [7]	0.003	90.35	89.06	**86.88**
Frustum ConvNet [1] + Improved CBAM	0.004	**90.44**	**89.06**	86.85

Table 2. Three-dimensional (3D) object detection AP (%) on the KITTI dataset. The best result are in bold.

Method	T(s)	Easy	Moderate	Hard
Frustum ConvNet [1]	**0.002**	89.02	78.80	77.09
Frustum ConvNet [1] + CBAM [7]	0.003	**89.35**	79.08	77.32
Frustum ConvNet [1] + Improved CBAM	0.004	89.29	**79.23**	**77.45**

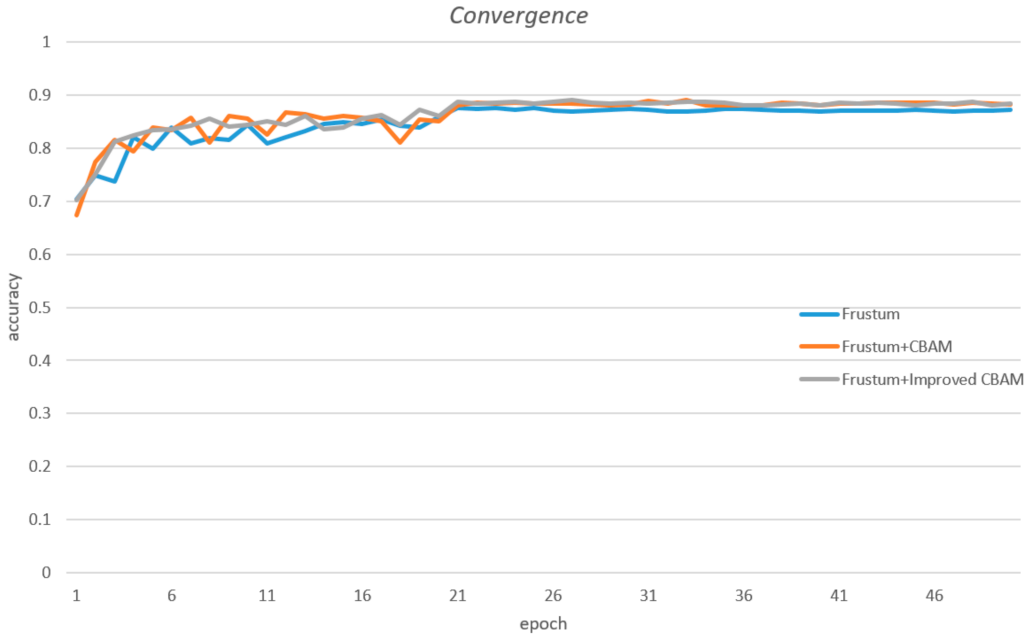


Figure 4. The validation accuracy of Frustum, Frustum + CBAM, and Frustum + Improved CBAM.

Tables 3 and 4 show the detection performance of our proposed FCAM on the KITTI validation set for 3D object detection and Bird's Eye View (BEV) object detection, respectively. Compared with Frustum ConvNet [1], our results showed improved AP by 0.21%, 0.27%, and 0.01%, respectively, in the BEV object detection of easy, moderate, and hard categories. For the 3D object detection, our results showed improved accuracy in easy, moderate, and hard cases by 0.27%, 0.43%, and 0.36%, respectively.

Table 3. BEV object detection AP (%) on the KITTI dataset. The best result are in bold.

Method	Easy	Moderate	Hard
MV3D [12]	86.55	78.10	76.67
VoxelNet [15]	89.60	84.81	78.57
F-PointNet [2]	88.16	84.92	76.44
IPOD [26]	88.3	86.4	84.6
Frustum ConvNet [1]	90.23	88.79	86.84
FCAM (Ours)	**90.44**	**89.06**	**86.85**

Table 4. Three-dimensional (3D) object detection AP (%) on the KITTI dataset. The best result are in bold.

Method	Easy	Moderate	Hard
MV3D [12]	71.29	62.68	56.56
VoxelNet [15]	81.97	65.46	62.85
F-PointNet [2]	83.76	70.92	63.65
IPOD [26]	84.1	76.4	75.3
AVOD-FPN [27]	84.41	74.44	68.65
PointRCNN [28]	88.88	78.63	77.38
Frustum ConvNet [1]	89.02	78.80	77.09
FCAM (Ours)	**89.29**	**79.23**	**77.45**

However, it can be seen that our method did not achieve the best results in 3D hard detection. The reason is that the attention model is targeted at objects with obvious image features. When the features are not obvious and occluded, the detection results can be

affected. By improving the channel attention block, accuracy is improved at the cost of running time. However, our method is fast enough for real-time applications, being able to process 200 images per second.

We also evaluated our 3D object detector on the indoor SUN-RGBD [9] test set for 3D object detection. Table 5 shows the detection performance of our proposed FCAM on the SUN-RGBD test set for 3D object detection. We tested 5198 images and compared them with Frustum ConvNet [1]. Our method achieved competitive results, with an improvement of 0.46% in Average Precision (AP) for 3D object detection. In four cases, our method achieved the best results in Average Precision (AP), such as the bed (1.67%), chair (0.92%), dresser (1.41%), and sofa (0.3%) for 3D object detection. On average, our method shows the best results, as can be seen in the last column of Table 5.

Table 5. Three-dimensional (3D) object detection AP (%) on the SUN-RGBD test set (IoU 0.25). The best result are in bold.

Method	Bathtub	Bed	Bookshelf	Chair	Desk	Dresser	Nightstand	Sofa	Table	Toilet	Mean
DSS [14]	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [29]	58.26	63.67	31.80	62.17	**45.19**	15.47	27.36	51.02	51.29	70.07	47.63
2Ddriven3D [30]	43.45	64.48	31.40	48.27	27.93	25.92	41.92	50.39	37.02	80.40	45.12
PointFusion [31]	37.26	68.57	**37.69**	55.09	17.16	23.95	32.33	53.83	31.03	83.80	45.38
Ren et al. [32]	**76.2**	73.2	32.9	60.5	34.5	13.5	30.4	60.4	**55.4**	73.7	51.0
F-PointNet [2]	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	**90.9**	54.0
Frustum ConvNet [1]	61.32	83.19	36.46	64.4	29.67	35.10	**58.42**	66.61	53.34	86.99	57.55
FCAM (Ours)	57.18	**84.86**	36.04	**65.32**	32.40	**36.51**	57.62	**66.91**	54.87	88.36	**58.01**

5. Conclusions and Future Works

This paper proposed using Frustum ConvNet with an improved CBAM attention model for 3D object detection. We propose an improved attention module by adding Multilayer Perceptron (MLP) and using the tanh activation function to improve the contrast between the object and the surrounding environment. We evaluate the proposed Frustum ConvNet with the attention model (FCAM) in the KITTI dataset and achieve competitive results with the state-of-the-art methods. This Frustum ConvNet with attention architecture can provide applications such as autonomous driving and robotic object manipulation.

In the future, we plan to further improve the performance of our 3D object detector. Our proposed attention model does not perform well when the network architecture is relatively complex. It is difficult for the attention model to focus on occluded objects in a complex environment. We plan to change the network architecture, reduce parameters, and further improve the adaptability of attention modules.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in [KITTI and SUN-RGBD dataset] at [doi: 10.1109/CVPR.2012.6248074 and 10.1109/CVPR.2015.7298655], reference number [8,9].

Acknowledgments: This material is based on work supported by the Ministry of Trade, Industry and Energy (MOTIE, Korea) under the Industrial Technology Innovation Program (10080619).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Wang, Z.; Jia, K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. Computer Vision and Pattern Recognition (CVPR). *arXiv* **2019**, arXiv:1903.01864.
2. 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.
3. Qi, C.R.; Liu, W.; Wu, C.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
4. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv* **2017**, arXiv:1706.02413.
5. Guo, W.; Xu, C.; Ma, S.; Xu, M. Visual attention based small object segmentation in natual images. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 1565–1568.
6. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 15–20.
7. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
8. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
9. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576.
10. 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.
11. Yang, B.; Yan, J.; Lei, Z.; Li, S.Z. Aggregate channel features for multi-view face detection. In Proceedings of the IEEE International Joint Conference on Biometrics, Clearwater, FL, USA, 29 September–2 October 2014; pp. 1–8.
12. 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.
13. 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, 24–28 September 2017; pp. 1513–1518.
14. 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.
15. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
16. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
18. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
19. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on Machine Learning (ICML), Haifa, Israel, 22–24 June 2010.
20. Zeiler, M.D.; Ranzato, M.; Monga, R.; Mao, M.; Yang, K.; Le, Q.V.; Hinton, G.E. On rectified linear units for speech processing. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3517–3521.
21. Le, Q.V.; Jaitly, N.; Hinton, G.E. A simple way to initialize recurrent networks of rectified linear units. *arXiv* **2015**, arXiv:1504.00941.
22. Ang-bo, J.; Wei-wei, W. Research on optimization of ReLU activation function. *Trans. Microsyst. Technol.* **2018**, *2*. Available online: https://en.cnki.com.cn/Article_en/CJFDTotalCGQJ201802014.htm (accessed on 11 February 2021).
23. Li, X.; Hu, Z.; Huang, X. Combine Relu with Tanh. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 51–55.
24. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
26. Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. Ipod: Intensive point-based object detector for point cloud. Computer Vision and Pattern Recognition (CVPR). *arXiv* **2018**, arXiv:1812.05276.
27. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3d proposal generation and object detection from view aggregation. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8.
28. Shi, S.; Wang, X.; Wang, H. PointRCNN Li. 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 15–20.

29. Ren, Z.; Sudderth, E.B. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1525–1533.
30. Lahoud, J.; Ghanem, B. 2d-driven 3d object detection in rgbd images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4622–4630.
31. Xu, D.; Anguelov, D.; Jain, A. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
32. Ren, Z.; Sudderth, E.B. 3d object detection with latent support surfaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 937–946.