# Multi-Fusion with Attention Mechanism for 3D Object Detection

Ning Wang, Ping Sun*
School of Software Engineering
Tongji University
Shanghai, China
*Corresponding author email: pingsun@tongji.edu.cn

***Abstract*—Artificial intelligence gradually plays the essential role in automatic driving, such as 3d object detection. Many state-of-the-art 3d detection frameworks fuse point cloud data and image data to perceive the surrounding environment of the vehicle. However, these approaches focus more on vehicle detections, and for objects with less point cloud sampling, such as pedestrians and cyclists, the performance is moderate. In this paper, we propose the multi-fusion framework with two kinds of attention mechanisms to solve the above problem and improve the detection accuracy of 3d objects. The proposed 3d attention mechanism with voxel sparse information is utilized in the framework. This framework contains two important modules: point fusion with 2d attention and voxel fusion with 3d attention. These modules firstly obtain the image features by projecting the lidar point or 8 vertices of the voxel to image feature maps. Then, these modules perform attentive fusion on the voxelized image features, point-wise image features and lidar data. Our evaluation on the challenging KITTI dataset, including 3d and bird's eye view metrics, demonstrates great improvements, especially at objects with less point cloud sampling.**

***Keywords-3d object detection; multi-sensor fusion; attention mechanism; convolutional neural network***

## I. INTRODUCTION

With the rapid development of artificial intelligence, great breakthrough has been made in the automatic driving. 3d object detection is an essential task in the automatic driving. Compared with 2d object detection, 3d object detection can obtain richer information such as the depth, position and volume, which helps to better perceive the surrounding environment of the vehicle. Lidar is the most used sensor for 3d object detection. Many early researches detect 3d objects from lidar point cloud [1, 2]. However, single sensor has its own disadvantages. For example, lidar cannot obtain intuitive image information. In this work, we focus on the multi-sensor data fusion for 3d object detection. On the basis of lidar point cloud data, the fusion of image data is helpful to give full play to the advantages of each sensor and improve the perception of multiple environments.

### *A. Challenges*

3d detection algorithms only driven by lidar suffer from the loss of texture information and the sparsity of point clouds. Missing texture information causes many false detections between objects of similar size. Very sparse point clouds of small or distant objects lead to missed detections.

DOI reference number: 10.18293/SEKE2021-115

To address these challenges, recent researches augment lidar point clouds with image features and learn to fuse features. Some researches [3, 4] utilize image features to generate 2d proposals, and then extract 3d features from the lidar points related to these 2d proposals. These approaches rely too much on reliable 2d detection results. In these methods, for the undetected object in the image, even if it has obvious features in the point clouds, it is difficult to detect it. Many algorithms [5, 6] project point clouds onto image features and then perform feature fusion. However, these approaches have high dependency on the reliability of high-resolution lidar point clouds and perform poorly when the lidar points are not sampled.

### *B. Our Contribution*

To deal with the above problems, the approach that reduces the reliance on high-resolution lidar point clouds, and increases the weight of image features when the lidar points are extremely sparse is expected. In this paper, we propose the multi-fusion framework with two kinds of attention mechanisms to achieve the above expectations. The proposed approach extends the recent algorithm Multimodal VoxelNet (MVX-Net) [6]. Specifically, this proposed framework contains two important modules: point fusion with 2d attention and voxel fusion with 3d attention. These modules obtain the image features by projecting the lidar point or vertices of the voxel to image feature maps. The combination of these two modules not only ensures the accurate association between the image features and the point clouds, but also reduces the dependence on the high-resolution lidar point clouds.

General attention mechanisms distribute attention according to image features and can't be directly applied to voxelized features. Inspired by the 2d attention mechanism, we propose the 3d attention mechanism for lidar point clouds. This mechanism takes dynamic voxelized data as the inputs, applies sparse 3d convolutions and produces a 3-dimensional spatial weight, which contributes to the selection of the effective voxelized features. What's more, considering that the sparsity information is weak before the attentive fusion, we apply the sparsity feature to voxelized image features.

The main contributions can be summarized as follows:

- The multi-fusion framework performs attentive fusion on voxelized image features, point-wise image features and lidar data. This framework preserves the detailed

image features without overly relying on the effectiveness of the lidar point clouds.

- 3d attention mechanism is proposed for lidar point clouds, which contributes to distributing the attention to voxelized features. What's more, the sparsity of lidar point clouds is utilized to enrich the voxelized features before the attention mechanism.
- Experiments on KITTI dataset demonstrates that our framework better handles error prone cases, and effectively reduces false detections caused by similar shape of point clouds, especially for objects with less point cloud sampling.

## II. Related Work

### A. 3D Object Detection From Multi Sensors

The multimodal 3d object detection fuses multi-sensor data, such as the LIDAR and RGB data. The realization of the multimodal fusion relies on the synchronization of multiple sensors in time and the transformation of spatial coordinates.

**Two-Stage Algorithm**: These algorithms can be divided into three categories: based on multiple views, based on 2d proposals and based on semantic segmentation methods.

*Multiple Views* Method: Chen et al. [7] proposed the MV3D algorithm, which firstly generates the 3d proposals by the lidar data and projects them to the bird's eye view, the front view and the image view. Then multi-view feature fusions are performed to refine the proposals. Later, many researches perform the multi-view fusion of different sensor data based on the 3d region proposals.

*2D Proposals* Method: Qi et al. [3] developed F-PointNets. This algorithm generates a 3d area for each 2d proposal, and applies PointNet++ [8] to obtain the point cloud features in the area. Zhao et al. [4] proposed the Point-SENet module to predict the scale factor and integrated the PointSIFT module to predict the direction.

S*emantic Segmentation* Method: In these methods, the existing semantic segmentation algorithm is used to eliminate most of the background points, and high-quality proposals are generated on the foreground points. Yang et al. [9] developed IPOD to remove most of the background points. Vora et al. [5] proposed the PointPainting, which appends the semantic features and the semantic prediction scores to the point cloud features. The accuracy performance is improved, but the inference speed is very slow.

The main disadvantage of Two-Stage Algorithms is that the two-stage operation slows the inference and training speeds, and requires higher computing resources for the computer. What's more, *Multiple Views* Method firstly generates 3d proposals, and then utilizes image information to refine the proposals. This causes these algorithms to rely heavily on 3d proposals generated from point clouds only. Meanwhile, *2D Proposals* Method pays more attention to reliable 2d detection results and weakens the effect of 3d point clouds. Therefore, an algorithm that can balance multimodal feature weights and complement each other is expected.

**One-Stage Algorithm**:

Sindagi et al. [6] proposed MVX-Net. This algorithm projects the non-empty voxels generated by VoxelNet [2] into the image, and uses a pre-trained network to obtain image features for each projected voxel feature. Then the combination of these image features and voxel features generates 3d detections. Though, the above method reduces the dependency on the availability of lidar points, the voxel projection reduces the accuracy of image features. In MVX-Net, authors also presented the point fusion. However, this method can't reasonably select effective image features from high dimensions and has poor performance in low point cloud sampling.

### B. Attention Mechanism

Attention plays an important role in human perception. Human vision obtains key areas by quickly browsing the whole picture, and then devotes more attention resources to the key area to obtain more detailed information, while suppressing other useless information. The attention mechanism in deep learning draws on the human attention and is widely used in various types of deep learning tasks such as natural language processing, image translation [10] and network pruning [11].

Recently, several researches have applied the attention mechanism to convolutional neural networks (CNN). Wang et al. [12] proposed Residual Attention Network, which stacks attention modules to generate attention-aware features. Hu et al. [13] proposed SENet which is generated by SE block. This architecture focuses on the channel relationship and uses global average pooling features to compute channel-wise attention. S Woo et al. [14] presented Convolutional Block Attention Module (CBAM). This module exploits both spatial-wise and channel-wise attentions and then the attention maps are multiplied to the feature map for adaptive refinement.

However, these attention mechanisms operate on 2d convolutions and cannot be directly applied to 3d voxel operations of point clouds. In addition, the difference between the voxel and the pixel is that voxels have different densities, and the application of the previous attention mechanism will lack the consideration of the density of voxels.

## III. Multi-Fusion Framework with Attention Mechanism

We present a multi-fusion framework with two kinds of attention mechanisms to fuse the RGB and point cloud features. Inspired by MVX-Net, the presented framework contains two important modules: point fusion with 2d attention and voxel fusion with 3d attention. These modules firstly obtain the corresponding image features by projecting the lidar point or vertices of the voxel to image feature maps. Then, these modules perform the attentive fusion on the lidar data and the image features. The proposed 3d attention mechanism for lidar point clouds takes dynamic voxelized data as inputs and applies the sparse 3d convolution, which helps to generate the effective voxelized features. What's more, the sparsity distribution of voxels is exploited for the attention mechanism, which enriches the image features with the sparse information.

The overall architecture is illustrated in Fig. 1. First, we utilize the 2d convolutional neural network which takes RGB

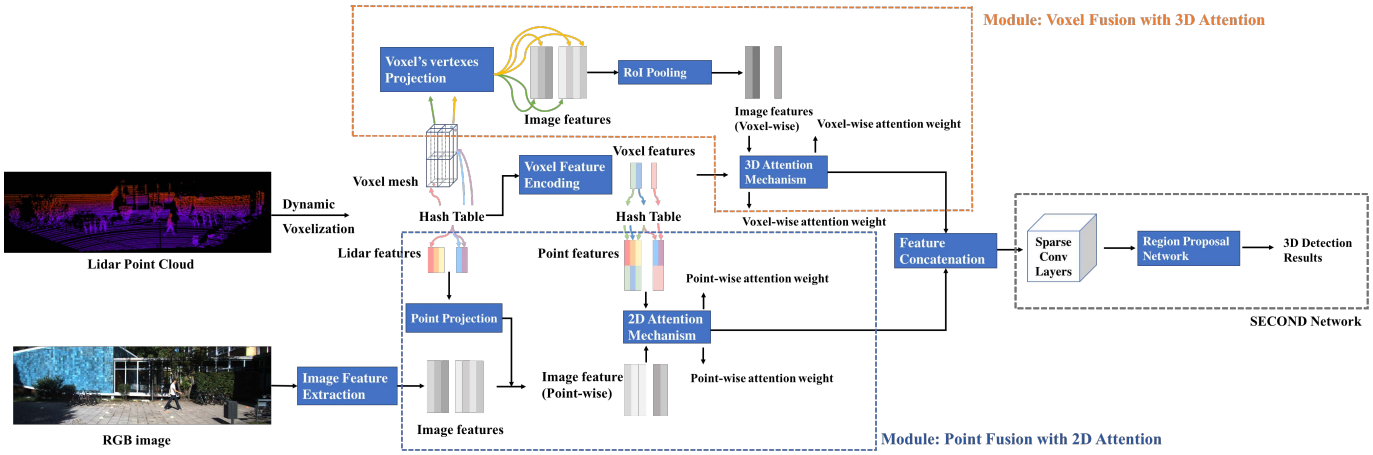


Figure 1. The Architecture of the Proposed Multi-fusion Framework

images as inputs and extracts multi-level image features. Next, voxel features are encoded from lidar point clouds and two attentive fusion modules are performed to generate fused features. Then, the 3d backbone network takes the concatenated features as inputs, and the head network outputs the 3d detection results.

### A. *Image Feature Extraction*

Residual Network (ResNet) [15] is made up from residual blocks with skip connections, which effectively increase the depth of network and the ability to extract features. Balancing the computing resource and model performance, we eventually adopt ResNet with 50 layers (ResNet50) as our image backbone.

Feature Pyramid Network (FPN) [16] is a feature extractor that combines multiple resolution features via a top-down pathway and lateral connections, which enriches the outputs with multi-dimensional information. We use FPN as the image neck network.

Given RGB images, the image backbone network generates multi-scale features. Then, these feature maps are merged by element-wise addition in the image neck network, which finally outputs several sets of image features with rich semantics.

### B. *Voxel Feature Encoding*

Voxel feature encoding (VFE) is a voxel feature learning network from VoxelNet [2]. The input of VFE is the point cloud data after the dynamic voxelization, which records the coordinates of the voxel where the point cloud is located and the raw features of the point cloud. The VFE network first obtains point-wise features through FCN learning, and then utilizes max pooling to generate the locally aggregated features. These features are regarded as the voxel global features, which are concatenate to each point-wise feature.

Stacks of such VFE layers transform low-dimensional point cloud features into high-dimensional voxel features, which will be the input of the Voxel Fusion with 3d Attention module. In order to obtain the input of the Point Fusion with 2d Attention module, the voxel features are discretized into the point cloud and connected with the initial point cloud feature.

### C. *Point Fusion with 2d Attention*

This module associates lidar point clouds to image features and perform the attentive fusion to obtain the point-wise features with additional image features. We adopt point fusion strategy for the accurate association information, which is described in MVX-Net [6]. Moreover, this module applies 2d attention to make fused features more expressive.

The details of this module are illustrated in Fig. 2. Given the multi-scale image features produced by the image backbone and point features produced by voxel feature encoding, this module outputs the attentive fusion features. In details, firstly 5 sets of 256-dimensional image features at different scales are input into the module. Then the point-wise image features are calculated:

$$I_{pw} = BL(-1 + 2 * \frac{M(T,P,C_{coor})}{[w,h]}, I) \quad (1)$$

$T$ denotes the transformation matrix, $P$ denotes preprocessing parameters, $C_{coor}$ denotes the 3d point cloud coordinates and $(w, h)$ is the width and height of the image. $M$ represents the coordinate transformation function and $BL$ represents the bilinear interpolation function. $I$ denotes the initial image features and $I_{pw}$ denotes the point-wise image features.

The 2d channel attention is performed in the above discrete point features, which is inspired by CBAM [14]. In details, as shown in Fig. 2, we respectively calculate the channel average feature and channel maximum feature of 640 dimensions, and use the sigmoid operation to obtain the 640-dimensional weight vector. The same attention mechanism is applied to the point-wise features obtained by VFE to generate the 64-dimensional weight vector. Then the fusion feature is generated by the concatenation of two attentive features.

The main advantage of this module is that for point cloud features with both raw features and voxel features, the attention mechanism can amplify effective features. In addition, the final

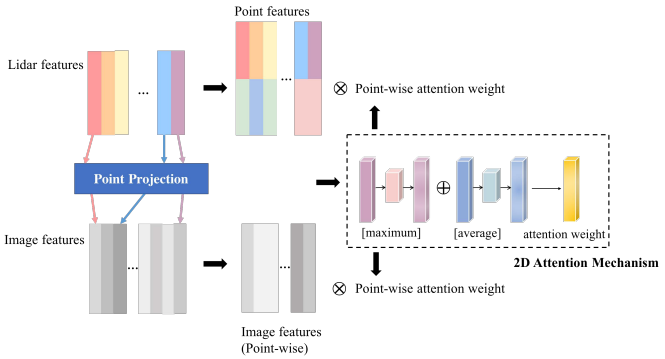


Figure 2. Point Fusion with 2D Attention Mechanism

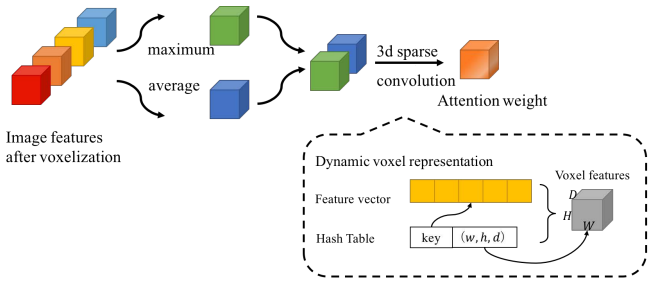


Figure 3. 3D Attention Mechanism

fusion directly performs on the raw point clouds, and this point-to-point mapping effectively reduces the quantization loss.

### D. Voxel Fusion with 3D Attention

This module extracts the image features projected by non-empty voxels. Then, we combine the voxel sparsity information with the 3d attention mechanism to take full advantage of multi-modal features. Voxel fusion effectively reduces the dependence on the high-resolution point clouds, as described in MVX-Net. For voxel operations, we propose a 3d attention mechanism and apply the sparsity information, which are conducive to extracting effective voxelized features and emphasizing image features when voxels are sparse.

The module is composed of 3 steps. (1) The extraction of voxelized image features. (2) The 3d attention mechanism is applied to obtain the attention vectors of multi-modal voxelized features respectively. (3) We calculate the voxel sparsity, and concatenate it with image features.

In detail, we first obtain all non-empty voxels' 8 vertex coordinates, and utilize the calibration matrix to project these point cloud coordinates to pixel coordinates in the image. Then, the largest rectangle obtained after the projection is utilized as the region of interest (RoI). Considering the different sizes of the RoIs, we use RoI Pooling to obtain 128-dimensional feature vectors from multi-scale image features.

From the discretization features obtained above, we design a 3d attention mechanism to obtain the weights of different voxels, which is shown in Fig. 3. Inspired by the spatial attention mechanism in CBAM, we respectively calculate the average and maximum features of all voxels and perform the concatenation operation. Then, combined with the voxel coordinates, the 3d sparse convolutions are performed to generate an $N$-dimensional attention vector, where $N$ represents the number of non-empty voxels. The same 3d attention mechanism is applied to the voxel features obtained by VFE.

In addition, we calculate the sparse value inside the voxel to optimize the image attention vector:

$$W_I = F(MLP\left[I^{d1}, \dots, I^{d128}, sigmoid\left(\frac{1}{N_{points}}\right)\right]) \quad (2)$$

$N_{points}$ represents the number of point clouds in a voxel, $I^d$ indicates the image feature, $W_I$ represents the attention weight for image data, $F$ denotes the operation of 3d attention and $MLP$ denotes the multilayer perceptron.

The main advantage of the proposed 3d attention mechanism is the selection of the effective voxelized features, thereby adaptively balancing the multimodal feature weights and emphasizing the image features when voxels are sparse. What's more, the fusion on voxels reduces the dependence on the lidar point clouds.

### E. SECOND Network

The SECOND [17] network improves VoxelNet and refines 3d convolution into 3d sparse convolution. First, we use the sparse conv layer and FPN to process the fused voxelized features. The structure of submanifold convolution is applied in this layer to limit the sparsity of the output, thereby greatly reducing the calculation of the convolution operation. Next, the region proposal network generates 3d proposals from the outputs of sparse conv layers. Then, after regression and refinement, the 3d detection results are generated.

## IV. Experiment Results

### A. Implementation Details

**Network Settings**: The image feature extraction takes images with the resolution of 1280 × 384 as inputs. We apply ResNet-50 to subsample the image features and output the feature maps of four blocks, of which the dimensions are 256, 512, 1024, 2048. Then, FPN is applied as the image neck network, which outputs five sets of 256-d multi-scale features. For lidar point clouds, the ranges are [0, 70.4], [-40, 40] and [-3, 1] meters respectively along the X, Y and Z axis, while the voxel size is [0.05, 0.05, 0.1]. The raw features of point clouds are xyz coordinates and reflectivity. The Dynamic VFE extracts 64-dimensional voxelized features from raw features. The anchor sizes of pedestrians, cyclists and cars are respectively [0.6, 0.8, 1.73], [0.6, 1.76, 1.73] and [1.6, 3.9, 1.56] meters.

**Training Details**: Adam with decoupled weight decay is adopted to optimize the network. The learning rate and weight decay are set as 0.003 and 0.01. The momentum factors are 0.95 and 0.99. What's more, we utilize warm-up for the first 1000 steps with the initial learning rate $1e^{-5}$. The total epoch is 36 and the batch size is 2. All experiments are based on the open source 3d detection toolbox mmdetection3d [18] with GPU NVIDIA GeForce GTX 1080Ti.

### B. Results on KITTI Dataset

We evaluate our method on the KITTI Object Detection Benchmark [19]. This dataset contains both 2d and 3d annotations of cars, pedestrians and cyclists. There are 7481 training samples and 7518 testing samples. Following the common division rule in [7], the training samples are divided into 3712 samples as the training set and 3769 samples as the validation set. The evaluation is on the validation set for all three object categories.

We evaluate 3d object detection performance in accordance with the official KITTI evaluation protocol. For cars, 70% overlap of the 3d bounding box is required, while for pedestrians and cyclists, 50% overlap is required. Depending on different sizes, occlusions and truncations, the evaluation has three levels, that is easy, moderate and hard. The average precision (AP) at different levels are respectively calculated for the comparison.

Table I shows the performance of our method on the KITTI validation set, compared with other state-of-the-art methods. Considering that most methods only report the performance on the car category, we perform the comparison on the car category. Compared with the baseline MVX-Net, improvements in 3d and BEV are 6.9% and 6.4% respectively in hard mode. Compared with the 2d proposal method F-PointNet [3] and the 3d proposal method MV3D [7], the performance of our proposed framework has improvement in all three modes, especially in the hard mode.

In our analysis, the 2d proposal approaches focus on image features, which leads to weak processing capabilities in complex situations with more occlusions. The 3d proposal methods are overly dependent on point cloud data, resulting in the poor detection for the objects with less point cloud sampling. However, in our method, applying multiple attention mechanisms can reasonably select image data and point cloud data, and balance the effects of two features. In addition, the fusion of point-wise and voxel-wise methods can reduce the dependence on point cloud sampling while ensuring the accuracy of point cloud feature extraction. Therefore, our method has more advantages in difficult scenes with more occlusions or less point cloud sampling, which also enables our method to achieve better performance in hard mode.

Detection results are shown in Fig. 4. According to the comparison of column 2, 3, and 4, our method can better detect objects with low point cloud sampling, such as pedestrians in the distance. From the comparison in the first column, our method reduces false detections, which are caused by similar point cloud shapes.

### C. Ablation Study

Ablation experiments are conducted to evaluate the effects of the 2d attention and 3d attention modules. All ablation studies are conducted on the pedestrian and cyclist classes, considering that these modules have a great improvement for objects with less point cloud sampling, as demonstrated before.

We report the comparison results in Table II. We first incorporate the 2d attention mechanism on the point fusion module, which increases pedestrian detection by 3% and cyclist detection by about 4%, in easy mode. In addition, there are also improvements in moderate and hard modes. This shows the effectiveness of the 2d attention mechanism.

We observe that combining the point fusion and voxel fusion modules, the detection results have not been greatly improved, which demonstrates that simply combining the above two modules cannot effectively optimize the detection performance. However, the integration of voxel fusion and 3d attention mechanism performs notably better both in pedestrian and cyclist detections, manifesting the importance of 3d attention mechanism for the voxel fusion.

Then, we investigate the effect of fusing the above four modules, that is respectively applying the 2d attention and 3d attention to the point fusion and voxel fusion module. Table II shows that this approach gets the best result. Compared with the baseline experiment, the pedestrian detection is improved by 5% and cyclist detection is improved by 4%. This shows that 2d attention and 3d attention mechanisms are beneficial for the fusion of image features, voxel features and point cloud features.

TABLE I. PERFORMANCE COMPARISON OF OBJECT DETECTION WITH STATE-OF-THE-ART METHODS ON CAR CLASS OF KITTI VALIDATION SET

| Method | $AP_{3D}$ (Car) | | | $AP_{BEV}$ (Car) | | |
|---|---|---|---|---|---|---|
| | *Easy* | *Moderate* | *Hard* | *Easy* | *Moderate* | *Hard* |
| MV3D (L) [7] | 71.2 | 56.6 | 55.3 | 86.2 | 77.3 | 76.3 |
| MV3D (L&I) [7] | 71.3 | 62.7 | 56.6 | 86.6 | 78.1 | 76.7 |
| F-PointNet (L&I) [3] | 83.8 | 70.9 | 63.7 | 88.2 | 84.0 | 76.4 |
| VoxelNet (L) [2] | 82.0 | 65.5 | 62.9 | **89.6** | 84.8 | 78.6 |
| MVX-Net (L&I) [6] | 85.5 | 73.3 | 67.4 | 89.5 | 84.9 | 79.0 |
| Our Proposed Method (L&I) | **86.4** | **76.3** | **74.3** | 89.2 | **86.4** | **85.4** |

TABLE II. ABLATION STUDY IN KITTI VALIDATION SET

| Point Fusion | 2d attention | Voxel Fusion | 3d attention | $AP_{3D}$ (Pedestrian) | | | $AP_{3D}$ (Cyclist) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *Easy* | *Moderate* | *Hard* | *Easy* | *Moderate* | *Hard* |
| √ | - | - | - | 59.2110 | 55.5011 | 50.6946 | 67.7570 | 51.2467 | 48.5160 |
| √ | √ | - | - | 62.7588 | 57.7813 | 51.9383 | 71.5124 | 52.5746 | 49.4193 |
| √ | - | √ | - | 59.2497 | 54.6029 | 50.3430 | 67.7148 | 53.0462 | 50.1806 |
| √ | - | √ | √ | 61.2489 | 56.6993 | 51.9426 | 70.9903 | **54.1027** | 50.6354 |
| √ | √ | √ | √ | **64.2624** | **58.2686** | **52.6567** | **71.6036** | 53.9948 | **51.4825** |

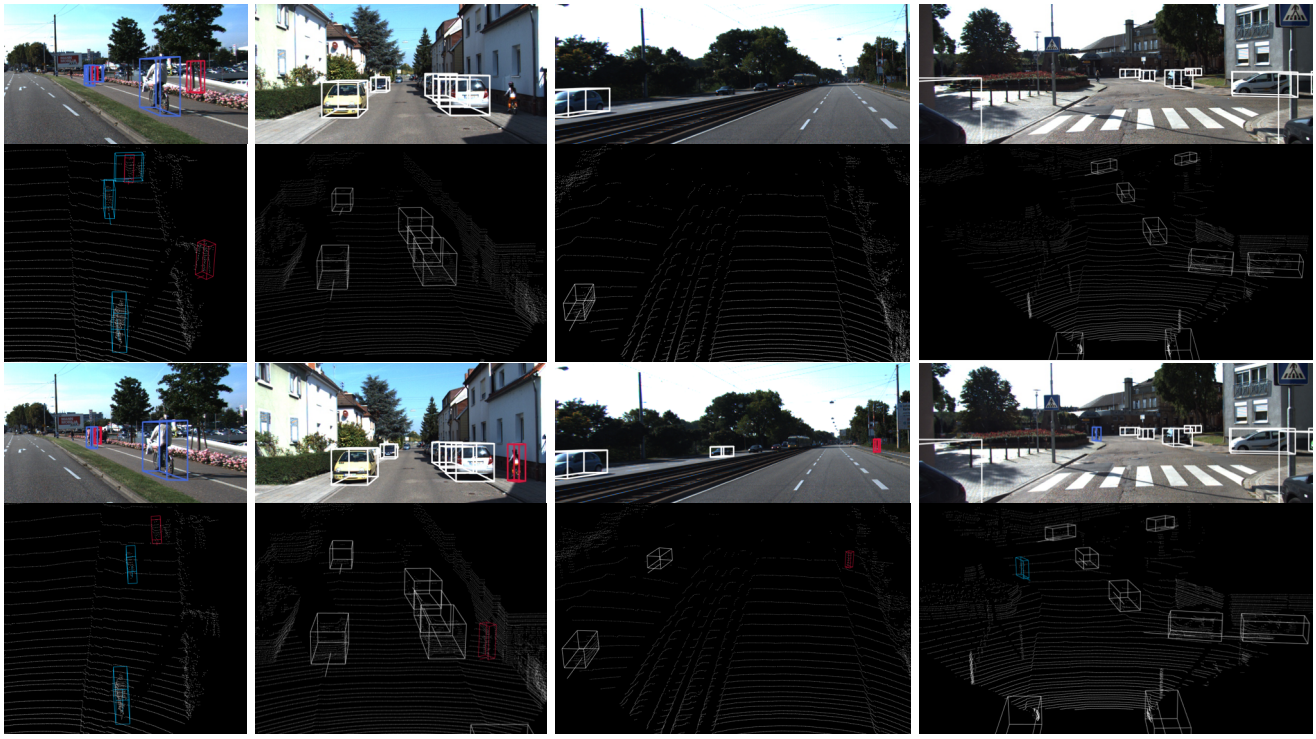

Figure 4. Comparison results on the KITTI validation dataset. For each comparison, the upper is the output of MVX-NET [18], while the under is the output of our proposed model. The color of car, cyclist and pedestrian are respectively white, blue and red.

## V. Conclusion

A multi-fusion framework is proposed in this paper, which implements the attentive fusion on image features and lidar data. We propose the 3d attention mechanism for point cloud data to amplify the effective voxelized features and contributes to emphasizing the image features when voxels are sparse. This framework retains detailed image features without overly relying on the effectiveness of lidar point clouds. Experiments show that the framework can better detect the distant or small objects and effectively reduce false detections caused by similar point cloud shapes.

## References


[1] Shi S, Guo C, Jiang L. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10529-10538.

[2] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4490-4499.

[3] Qi C R, Liu W, Wu C. Frustum pointnets for 3d object detection from rgb-d data[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 918-927.

[4] Zhao X, Liu Z, Hu R. 3D object detection using scale invariant and feature reweighting networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 9267-9274.

[5] Vora S, Lang A H, Helou B. Pointpainting: Sequential fusion for 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4604-4612.

[6] Sindagi V A, Zhou Y, Tuzel O. Mvx-net: Multimodal voxelnet for 3d object detection[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 7276-7282.

[7] Chen X, Ma H, Wan J. Multi-view 3d object detection network for autonomous driving[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 1907-1915.

[8] Qi C R, Yi L, Su H. PointNet++ deep hierarchical feature learning on point sets in a metric space[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 5105-5114.

[9] Yang Z, Sun Y, Liu S. Ipod: Intensive point-based object detector for point cloud[J]. 2018.

[10] Kim, Jin Yong, Lee, Myeong Oh, Jo, Geun Sik. DCBlock: Efficient module for unpaired image to image translation using GANs[C]//SEKE. 2020: 13-18.

[11] Wang X J, Yao W B, Fu H. A Convolutional Neural Network Pruning Method Based On Attention Mechanism[C]//SEKE. 2019: 343-452.

[12] Wang F, Jiang M, Qian C. Residual attention network for image classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3156-3164.

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[14] Woo S, Park J, Lee J Y. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[16] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[17] Yan Y, Mao Y, Li B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[18] CUHK Multimedia Laboratory. https://github.com/open-mmlab/mmdetection3d, 2020.

[19] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.