Journal of Physics: Conference Series

PAPER • OPEN ACCESS

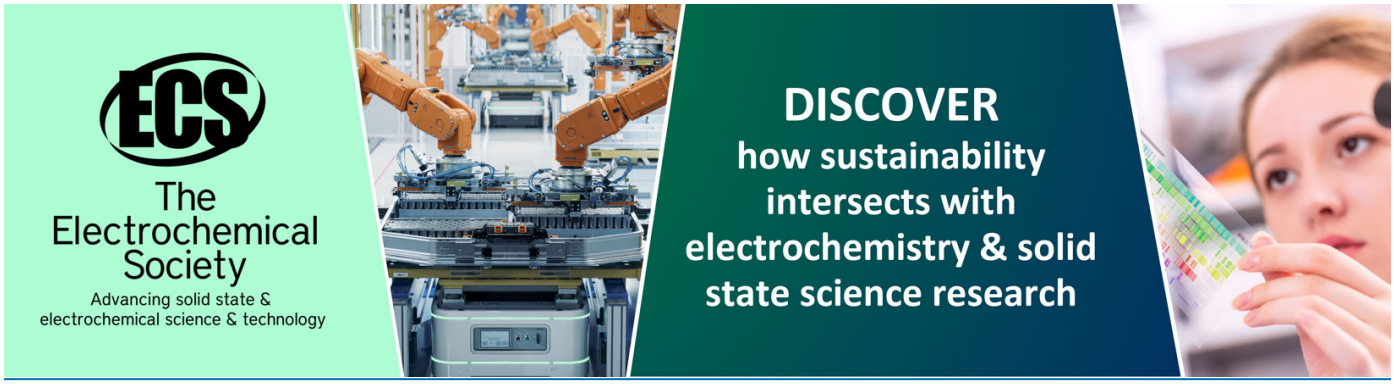
RGB and LiDAR Fusion-based 3D Semantic Segmentation for Autonomous Driving

To cite this article: Jianguo Liu *et al* 2023 *J. Phys.: Conf. Ser.* **2632** 012034

View the article online for updates and enhancements.

You may also like

- A non-iterative method for the electrical impedance tomography based on joint sparse recovery
Ok Kyun Lee, Hyeonbae Kang, Jong Chul Ye et al.
- Remote actuation based on magnetically responsive pillar arrays
Wei Jiang, Lanlan Wang, Bangdao Chen et al.
- Self-force and radiation reaction in general relativity
Leor Barack and Adam Pound



This content was downloaded from IP address 18.119.136.235 on 12/05/2024 at 08:24

RGB and LiDAR Fusion-based 3D Semantic Segmentation for Autonomous Driving

Jianguo LIU [1, 2], Zhiling JIA [1, 2*], Gongbo Li [1, 2], Fuwu YAN [1, 2], Youhua WU [1], and Yunfei SUN [3]

[1] Foshan Xianhu Laboratory of the Advanced Energy Science and Technology Guangdong Laboratory, Foshan 528200, Guangdong, China

[2] Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan 430070, China

[3] Ningbo Huade Automobile Parts Co., Ltd., Ningbo 315000, Zhejiang, China

*Corresponding author's e-mail: jzlwhut@163.com

Abstract: Projection-based multimodal 3D semantic segmentation methods suffer from information loss during the point cloud projection process. This issue becomes more prominent for small objects. Moreover, the alignment of sparse target features with the corresponding object features in the camera image during the fusion process is inaccurate, leading to low segmentation accuracy for small objects. Therefore, we propose an attention-based multimodal feature alignment and fusion network module. This module aggregates features in spatial directions and generates attention matrices. Through this transformation, the module could capture remote dependencies of features in one spatial direction. This helps our network precisely locate objects and establish relationships between similar features. It enables the adaptive alignment of sparse target features with the corresponding object features in the camera image, resulting in a better fusion of the two modalities. We validate our method on the nuScenes-lidar seg dataset. Our CAFNet achieves an improvement in segmentation accuracy for small objects with fewer points compared to the baseline network, such as bicycles (6% improvement), pedestrians (2.1% improvement), and traffic cones (0.9% improvement).

1. Introduction

Three-dimensional semantic segmentation is an essential visual task for scene understanding in autonomous driving. Its objective is to categorize each minimal unit of input modal information. The development of self-driving cars has been driven by advancements in artificial intelligence and an increasing number of sensors that are integrated into cars. Among these sensors, the camera is one of the most widely deployed ones, providing rich semantic and obstacle-shape information. However, camera images lack depth information and are susceptible to light conditions. The LiDAR sensor, also an advanced vehicle sensor, can capture precise spatial location and three-dimensional information about the environment and obstacles around the car. Nevertheless, the point cloud from LiDAR is sparse, resulting in insufficient details for obstacle representation, making it challenging to distinguish objects with similar appearances. This problem becomes more severe when dealing with small objects at long distances. Over the years, there has been a growing trend of equipping vehicles with both cameras and LiDAR sensors. Sensor fusion techniques have emerged as a research hotspot, aiming to integrate the information from these two types of sensors, leveraging their complementary strengths, and achieving

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

accurate and reliable environmental perception.

Based on the number of input modalities, there are two kinds of semantic segmentation methods: single-sensor methods and multi-sensor methods. The former includes camera-based methods and LiDAR-based methods. FCN [1] is used for image semantic segmentation and is the first full convolutional network for pixel-level prediction without the need for manual design. Wang et al. [2] enhance the semantic segmentation performance by incorporating dense upsampling and dilated convolutions. However, camera-based methods alone are unable to provide depth information of the surrounding environment for vehicles and are susceptible to changes in lighting conditions.

LiDAR sensor data can be represented in different data forms, including the following three methods:

(1) Point-based methods. This type of method directly utilizes the raw point cloud as input. Qi et al. [3] can learn features directly from a point cloud with multi-layer perceptron and realize point cloud invariance through pooling operation, which is suitable for classification, segmentation, and detection tasks. Qi et al. [4] introduce hierarchical point set abstraction layers to process more complex point cloud data, and these features from set abstraction layers are then used to complete specific prediction tasks. However, this series of methods struggle with handling sparse point clouds.

(2) Voxel-based methods. We perform semantic segmentation by using the obtained voxel representation. Tang et al. [5] propose a sparse point-voxel convolution module, and this module adds minimal computational overhead and addresses the issue of information loss. Zhou et al. [6] present a cylindrical segmentation for 3D point cloud representation. This method shifts the emphasis of outdoor point cloud segmentation from a two-dimensional projection to a three-dimensional structure, allowing for a deeper investigation into the intrinsic features of outdoor point clouds. Nevertheless, voxel-based methods often have problems with information loss during the conversion of raw input to voxels and typically have high time and space complexity.

(3) Projection-based methods. Zhang et al. [7] propose a new point cloud representation method. The method uses a polar coordinate system to encode point clouds, introducing a novel way of representing spatial information. Milioto et al. [8] convert point clouds into approximate 2D distance image representations through spherical projection. This method can employ any CNN backbone network for semantic segmentation and introduce a novel post-processing technique to restore consistent semantic information during inference. Multisensor methods combine LiDAR point clouds with camera images as input. Krispel et al. [9] establish point correspondences through RGB/LiDAR calibration and fuse feature representations from RGB and range images. The RGB features are warped based on the known correspondences to adapt them to the range image network. Zhuang et al. [10] project point clouds onto image coordinates and RGB images can have spatial information. It proposes a two-flow network that extracts features from point clouds and images and introduces additional perceptual losses. Yan et al. [11] propose a general training strategy for 2D prior-assisted semantic segmentation. It leverages multiscale fusion and knowledge distillation from auxiliary modalities to extract richer context information from multimodal data.

In conclusion, the fusion of multiple sensors in computer vision can enhance perception capabilities effectively by combining data from different sensors, and enhance the adaptability of autonomous driving vehicle perception systems in different environments. However, most multimodal-based semantic segmentation algorithms utilize image information as decoration for point clouds. Yet, there are heterogeneities and perspective variations between the two modalities, making it crucial to effectively align and fuse the features from the two different modalities. Therefore, this paper proposes a coordinated attention-based module for aligning and fusing LiDAR information and image features, building upon the 3D semantic segmentation network PMF [10]. The complementary information in the images allows the model to accurately segment small objects with sparse point cloud data. The proposed feature fusion and alignment network dynamically learns the correlation and connectedness between point cloud features and their corresponding image features, establishing an adaptive association mechanism between the two feature extraction networks, thus fully utilizing the complementary information from both modalities.

2. Method

2.1 Data processing

We employ a projection-based method, which differs from other methods as it preserves the original information and applies it to real-world autonomous driving scenarios. Moreover, the range-view image obtained by using the above method is dense, compared to the original point cloud data, which possesses characteristics similar to camera images. Therefore, we can use a generic convolutional network to extract deep-level features of the point cloud, and the transformed range images are more amenable to feature alignment and fusion with camera images.

Each point in the point cloud is composed of spatial coordinates (x, y, z) and reflection intensity r. As described in [8], the projection of point clouds into distance images can be represented as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}[1-\arctan(y,x)\pi^{-1}]\omega \\ [1-(\arcsin(z,r^{-1})+f_{down})f^{-1}]h \end{pmatrix} \tag{1}$$

In the proposed approach, every point of the point cloud is encoded with a 5-dimensional feature representation. The (u, v) coordinates represent the pixel coordinates corresponding to a point in the projected image. The variables ω and h represent the size of the projected image.

According to Equation (1), each point can be coded for a 5-dimensional feature representation, and the feature vector includes the distance (d), spatial coordinates (x, y, z), and reflectance strength (r), i.e., (d, x, y, z, r). By applying Equation (1), the point cloud is transformed into a range-view image with dimensions (5, w, h). Because point clouds are not as dense as RGB images, there may be pixels in the range image that do not have matching points in the point cloud. Hence, these pixels are initialized with a value of 0.

2.2 Network structure

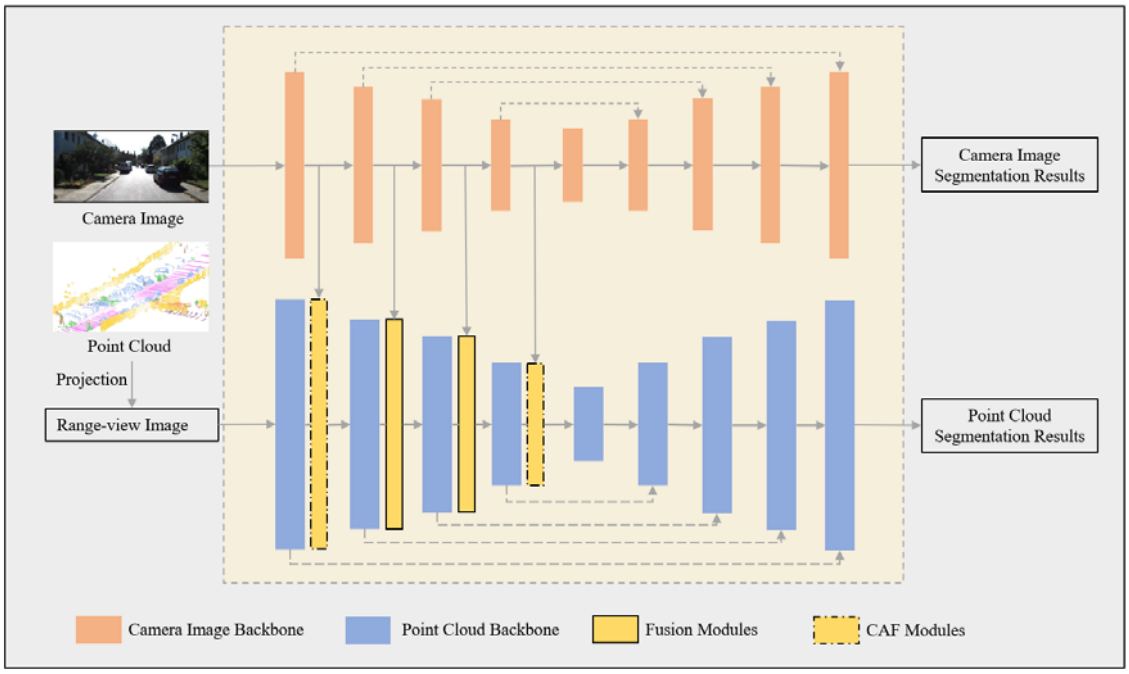


Figure 1. Network structure

In this paper, the Projection-based Multimodal Fusion PMF [10] network is used as the basic network, which performs 3D semantic segmentation by taking camera images and LiDAR point clouds as input. It is a dual-stream network that extracts features from both modalities and fuses them by using a fusion

module. Additionally, a perceptual adversarial loss is employed to calculate the differences between two modalities of information.

The overall CAFNet structure is shown in Figure 1, consisting of two parts: the SalsaNext [12] network which uses the range-view image as input, and the ResNet network for image segmentation. Both the range-view image segmentation network and the camera image segmentation network adopt an encoder-decoder architecture. The encoder units incorporate a set of residual connection blocks, while the decoder part combines the upsampled features from the residual blocks through skip connections.

To begin with, the unordered point cloud data is converted to a dense range-view image representation that allows for standard convolutional operations. The 3D semantic segmentation network's encoder is then employed to encode and extract features from the range-view image, while the ResNet is used as the camera image processing stream. Next, the features extracted from network layers of different dimensions are combined by using an attention-based feature fusion module. This fusion process generates fused features. Subsequently, the encoder outputs are upsampled through a decoder network to restore the original size, and class labels are generated for each pixel, producing the final result of semantic segmentation.

2.3 Feature fusion module

In the encoder module, every layer conducts downsampling by using convolutional and activation layers to extract features from various receptive fields. Additionally, adaptive pooling is used to resize the camera image features and range-view image features to the same size. Then, the extracted features are fused and aligned by the CAF module. The feature fusion module can be repeated multiple times in the encoder to achieve different levels of feature fusion.

Due to the disparity in the field of view between LiDAR and the camera, there is no precise one-to-one mapping between the target objects. When fusing the features from both modalities in the encoder, not all pixel features are equally important, and the information from the LiDAR depth features is unevenly aligned with each camera pixel. To better align the features from LiDAR with the most relevant camera features and complement the point cloud information with camera image information, we propose the Coordinate Attention Fusion (CAF) module. It utilizes coordinate attention mechanisms [13] to dynamically capture the correlation between the two modalities. The aforementioned structure is illustrated in Figure 2.

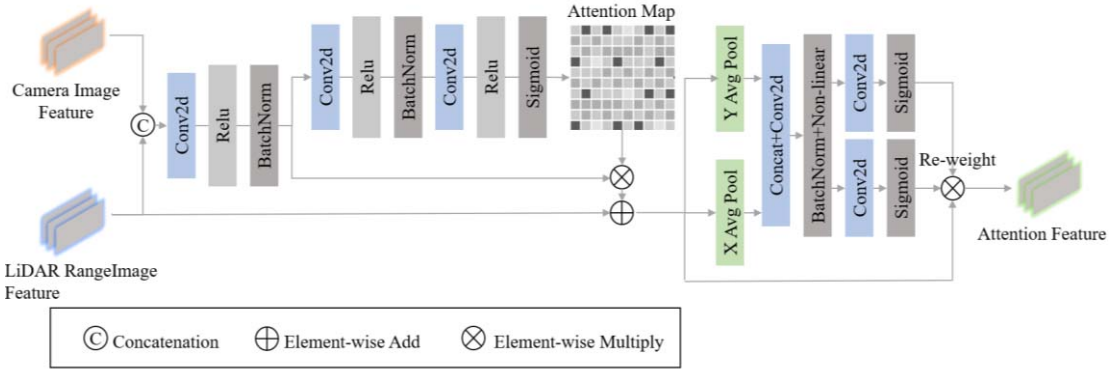


Figure 2. CAF (Coordinate Attention Fusion) module

To achieve the fusion, we first make the concatenation of the features from the two networks along the channel dimension and make the number of channels the same as the original convolutional layers. To better integrate image features into range-view image features, an attention module is introduced. The attention matrix is generated by convolutional and activation layers and then weighted onto the fused features. The fused features are further utilized as a complement to the original point cloud features through residual structures. However, the attention matrix generated by the convolutional layers can only capture local relationships and cannot model the crucial long-range dependencies for visual tasks. Therefore, the feature fusion module can establish the feature dependence relationship in space and channel dimension by introducing coordinate attention.

It has been demonstrated that channel attention is significantly effective in improving model performance by selectively emphasizing interdependent channel feature maps through the integration of correlation features among all channels. However, it does not contain positional information, which plays a crucial role in generating spatial attention maps. Positional attention weights the features from all positions and selectively aggregates features from each position. Similar features are thus correlated regardless of the distance between them.

The CAF module encodes channel relationships and remote dependencies of features respectively, and provides accurate location information. The input to this module is the feature vector $X=[x_1, x_2, \ldots, x_c] \in R^{R\times H\times W}$ from the attention residual structure described earlier. Given the input X, pooling is applied along the parallel coordinates and vertical coordinates separately with ranges of (H, 1) and (1, W), respectively, effectively capturing information along the two dimensions separately. Therefore, the output of channel *c* at height *h* and width *w* can be defined as follows.

$$z_c^h(h) = \frac{1}{W} \sum_{0\le i<W} x_c(h,i) \tag{2}$$

$$z_C^w(w) = \frac{1}{H} \sum_{O\le j<H} x_c(j,w) \tag{3}$$

By combining the above equations, features are aggregated along two spatial directions, leading to the generation of a duo of attention matrices with directional information. These matrices enable the CAF module to establish extensive dependencies along one direction while preserving accurate positional details. This facilitates more accurate localization of the objects of interest and establishes connections between similar features in the CAF module. Consequently, the features of sparse point cloud targets are adaptively aligned with the corresponding object features in camera images, achieving the complementary effect of camera image features on point cloud features.

3. Experiments

3.1 Dataset

Our method is evaluated in the nuScenes-lidar seg dataset [14], which is a multimodal dataset designed for perception tasks in self-driving, such as three-dimensional object detection, tracking, and segmentation. This dataset is the first to include data from a full suite of sensors for fully self-driving vehicles, all providing a complete panoramic view. There are 1, 000 scenarios in this dataset, with each lasting 20 seconds, and each of the 23 categories is labeled. In scenes-lidar seg, each point belonging to keyframes in the nuScenes dataset is annotated with one of 32 possible semantic labels. The scenes-lidar seg dataset comprises 1.5 billion annotated points in 34, 000 point clouds from 850 scenes.

3.2 Evaluation metric

In this paper, the Mean Intersection over Union (mIoU) is used to assess the performance of the model on point clouds.

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c} \tag{4}$$

where $TP_c, FP_c,$ and FN_c represent the true positives, false positives, and false negatives predictions of class *c*, and *C* represents the number of categories.

3.3 Training and inference details

The experimental setup is as follows. In terms of hardware, an Intel Xeon Silver 4210 processor and three NVIDIA RTX3090 24 G graphics cards were used. The software system consisted of Ubuntu 18.04.06, CUDA version 11.3, and PyTorch version 1.10.1. A batch size of 24 and a training epoch of

50 were configured. The hybrid optimization method was employed to train networks, with SGD and Nesterov used for the RGB image segmentation network, and Adam [15] is used for the point cloud segmentation network. SGD [16] was employed for learning rate decay from 0.001 until it reached 0. Inference validation was performed after each training epoch.

3.4 Results on nuScenes

Table 1. Comparisons on the nuScenes-lidar seg validation set

Method	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrian	Traffic-cone	Trailer	Truck	Driveable	Other-flat	Sidewalk	Terrain	Manmade	Vegetation	MIoU (%)
RangeNet++ [8]	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8	65.5
PolarNet [7]	74.7	28.2	85.3	90.0	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74.0	87.3	85.7	71.0
Salsanext [12]	74.8	34.1	85.9	88.4	42.2	72.4	72.2	63.1	61.3	76.5	96.0	70.8	71.2	71.5	86.7	84.4	72.2
AMVNet [17]	79.8	32.4	82.2	86.4	62.5	81.9	75.3	72.3	83.5	65.1	97.4	67.0	78.8	74.6	90.8	87.9	76.1
Cylinder3D [6]	76.4	40.3	91.3	93.8	51.3	78.0	78.9	64.9	62.1	84.4	96.8	71.6	76.4	75.4	90.5	87.4	76.1
CAFNet (Ours)	73.6	52.6	88.0	92.2	63.7	53.5	83.0	71.8	65.7	82.1	95.9	74.2	74.5	77.0	91.0	89.9	**76.8**

The 3D semantic segmentation network CAFNet was validated on the nuScenes-lidar seg dataset. The average Intersection over Union (IoU) and per-class IoU for 16 object categories, as provided by the dataset, were compared with existing SOTA 3D semantic segmentation networks, as shown in Table 1. By examining the table, it is evident that the CAFNet attains an average Intersection over Union (IoU) of 76.8% for point cloud semantic segmentation, which demonstrated a 4.6% improvement in average IoU. Compared to the projection-based method Salsanext, it is 0.7% higher than the multi-view fusion method AMVNet and the voxel-based method Cylinder3D.

Table 2. Comparison of IoU on small object segmentation

Method	mIoU (%)	Bicycle	Pedestrian	Traffic-cone
PMF [10]	76.9	46.6	80.9	70.9
RPVNet [18]	77.6	43.4	80.5	66.0
CAFNet (Ours)	76.8	**52.6**	**83.0**	**71.8**

Bicycles, pedestrians, and traffic cones are commonly encountered as small objects in typical driving scenarios. Due to the limitations of laser radar (LiDAR) data acquisition, these object point clouds often exhibit limited quantity and partial occlusion. Moreover, in scenarios involving long distances, the point cloud becomes sparser, which further exacerbates the difficulty of segmenting small objects, thereby posing a challenge to the network.

Table 2 compares the CAFNet with the baseline network PMF, as well as the point-distance-image-voxel fusion-based RPVNet. Table 2 shows that our proposed network achieves a similar average Intersection over Union (IoU) to the baseline network PMF. However, there are notable improvements in bicycle segmentation precision (an increase of 6%), pedestrian segmentation precision (an increase of 2.1%), and traffic cone segmentation precision (an increase of 0.9%). In comparison to RPVNet,

although there is a certain gap in terms of average IoU, the segmentation precision for small objects like bicycles, pedestrians, and traffic cones has significantly improved. This validates the effectiveness of the CAF module in enhancing the segmentation performance for small objects.

4. Conclusion

Taking advantage of the coordinate attention mechanism, this research introduces a multimodal feature fusion and alignment network module, which is applied to point cloud semantic segmentation networks. The proposed module enables the network to locate and align two modalities of interest and establish connections between similar features. The efficacy of the CAF module in improving the segmentation precision of small objects is demonstrated through experiments conducted on the nuScenes-lidar seg public dataset. This validates the efficacy of the proposed network module in enhancing the alignment of small object features in multimodal networks. However, it should be noted that the CAF module does not achieve improvement in the precision for all objects compared to the SOTA methods. Potential future directions involve improving the segmentation precision by enhancing the representation of LiDAR raw data and feature extraction modules based on the proposed method.

References

[1] Long J., Shelhamer E., and Darrell T. Fully convolutional networks for semantic segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3, 431-3, 440.
[2] Wang P., Chen P., Yuan Y., et al. Understanding of convolution for semantic segmentation [C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Ieee, 2018: 1, 451-1, 460.
[3] Qi C. R., Su H., Mo K., et al. Pointnet: Deep learning on point sets for 3d classification and segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 652-660.
[4] Qi C. R., Yi L., Su H., et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space [J]. Advances in neural information processing systems, 2017, 30.
[5] Tang H., Liu Z., Zhao S., et al. The study of efficient 3d architectures with sparse point-voxel convolution [C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 685-702.
[6] Zhou H., Zhu X., Song X., et al. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation [J]. arXiv preprint arXiv: 2008.01550, 2020.
[7] Zhang Y., Zhou Z., David P., et al. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9, 601-9, 610.
[8] Milioto A., Vizzo I., Behley J., et al. Rangenet++: Fast and accurate lidar semantic segmentation [C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 4, 213-4, 220.
[9] Krispel G., Opitz M., Waltner G., et al. Fuseseg: Lidar point cloud segmentation fusing multi-modal data [C]// Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1, 874-1, 883.
[10] Zhuang Z., Li R., Jia K., et al. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16, 280-16, 290.
[11] Yan X., Gao J., Zheng C., et al. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds [C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 677-695.
[12] Cortinhal T., Tzelepis G., and Aksoy E. E. Salsanext: Fast semantic segmentation of lidar point clouds for autonomous driving [J]. arXiv preprint arXiv: 2003.03653, 2020, 3(7).
[13] Hou Q., Zhou D., and Feng J. Coordinate attention for efficient mobile network design [C]//

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13, 713-13, 722.
[14] Caesar H., Bankiti V., Lang A. H., et al. Nuscenes: A multimodal dataset for autonomous driving [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11, 621-11, 631.
[15] Kingma D. P. and Ba J. Adam. A method for stochastic optimization [J]. arXiv preprint arXiv: 1412.6980, 2014.
[16] Loshchilov I. and Hutter F. Sgdr. Stochastic gradient descent with warm restarts [J]. arXiv preprint arXiv: 1608.03983, 2016.
[17] Liong V. E., Nguyen T. N. T., Widjaja S., et al. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation [J]. arXiv preprint arXiv: 2012.04934, 2020.
[18] Xu J., Zhang R., Dou J., et al. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16, 024-16, 033.