Ground-Aware Point Cloud Semantic Segmentation for Autonomous Driving

Jian Wu
wujianq@mail.ustc.edu.cn
University of Science and Technology of China

Jianbo Jiao
jianbo@robots.ox.ac.uk
Department of Engineering Science, University of Oxford

Qingxiong Yang
Yang@moonx.ai
MoonX.AI

Zheng-Jun Zha
zhazj@ustc.edu.cn
NEL-BITA, University of Science and Technology of China

Xuejin Chen*
xjchen99@ustc.edu.cn
NEL-BITA, University of Science and Technology of China

ABSTRACT

Semantic understanding of 3D scenes is essential for autonomous driving. Although a number of efforts have been devoted to semantic segmentation of dense point clouds, the great sparsity of 3D LiDAR data poses significant challenges in autonomous driving. In this paper, we work on the semantic segmentation problem of extremely sparse LiDAR point clouds with specific consideration of the ground as reference. In particular, we propose a ground-aware framework that well solves the ambiguity caused by data sparsity. We employ a multi-section plane fitting approach to roughly extract ground points to assist segmentation of objects on the ground. Based on the roughly extracted ground points, our approach implicitly integrates the ground information in a weakly-supervised manner and utilizes ground-aware features with a new ground-aware attention module. The proposed ground-aware attention module captures long-range dependence between ground and objects, which significantly facilitates the segmentation of small objects that only consist of a few points in extremely sparse point clouds. Extensive experiments on two large-scale LiDAR point cloud datasets for autonomous driving demonstrate that the proposed method achieves state-of-the-art performance both quantitatively and qualitatively. The project and dataset are available at www.moonx.ai/#/open.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**;

KEYWORDS

Autonomous driving, semantic segmentation, point clouds, sparse LiDAR

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351076

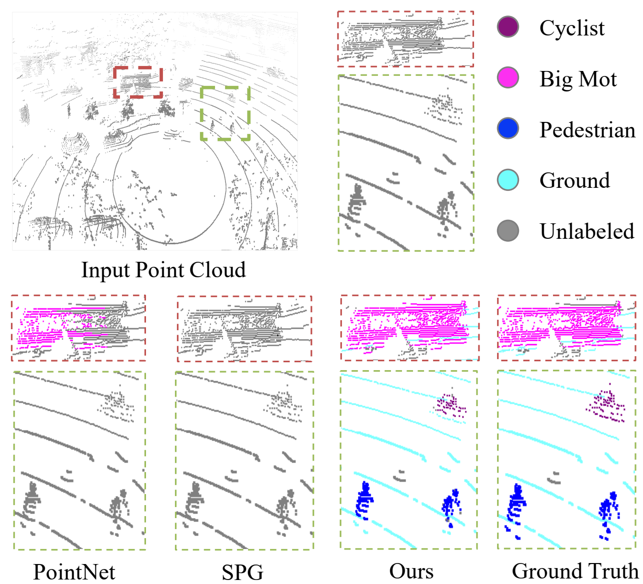


Figure 1: Our ground-aware approach accurately segments small objects (such as pedestrians and cyclists) in sparse LiDAR point clouds and outperforms state-of-the art methods (PointNet [27] and SPG [14]).

ACM Reference Format:
Jian Wu, Jianbo Jiao, Qingxiong Yang, Zheng-Jun Zha, and Xuejin Chen. 2019. Ground-Aware Point Cloud Semantic Segmentation for Autonomous Driving. In *Proceedings of Proceedings of the 27th ACM International Conference on Multimedia (MM '19).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351076

1 INTRODUCTION

With the popularity of autonomous driving and artificial intelligence, scene understanding becomes crucial for the safety and efficiency of machine perception in dynamic, complex scenes. Autonomous driving vehicles are usually equipped with various sensors, among which LiDAR plays an important role in capturing the surrounding environment. LiDAR scanning equipment is insensitive to lighting change and accurate in distance measurement.

From the 3D point cloud data collected by a LiDAR system, the 3D environment can be reconstructed to help an autonomous system make decisions intelligently.

In recent years, great progress has been made using deep learning techniques in semantic segmentation of point clouds [1, 10, 14, 16, 17, 26, 27, 29]. However, the point clouds, captured by LiDAR devices with fewer channels, are extremely sparse. Figure 1 shows an example of the sparse point cloud captured at one single position. This sparsity of LiDAR point clouds poses two challenges when applying previous methods in autonomous driving scenarios. First, from the example in Figure 1, it can be seen that there is an obvious difference between the distribution of ground points and object points. The LiDAR points of the ground are ring-shaped and the distance between the rings increases gradually from the origin of the LiDAR device to distant regions. This heterogeneous anisotropic distribution makes it significantly difficult to apply existing methods that are designed for isotropic point clouds. Secondly, existing methods classify each individual point by extracting features from its local neighborhood. However, for sparse LiDAR point clouds, it is challenging to perceive reliable features in a small local neighborhood due to the heterogeneous anisotropic distribution.

To exploit more information beyond a local neighborhood and extract more reliable features, we propose a ground-aware approach to address the above-mentioned challenges in autonomous driving. Specifically, we introduce a strategy to automatically separate ground points and objects, which supports the subsequent feature extraction for different parts. Considering the complexity of landscapes in urban scenes, we utilize a multi-section plane extraction method to represent the ground surface in a sparse LiDAR point clouds. Although only a rough segmentation for ground points is obtained at the beginning, the extracted ground points provide valuable context to eliminate the ambiguity in categorizing points of the object on the ground.

After a rough segmentation of ground points, we further leverage local shapes as well as the relationship between the ground and objects to extract more reliable features for semantic segmentation. Due to the sparsity and anisotropic distribution of LiDAR points, long-range dependence between points is crucial for feature extraction. While the attention mechanism has been successfully used in many tasks such as language translation [19, 20] and image captioning [9, 42], we introduce a ground-aware attention module to capture the long-range dependence between the ground and objects. We investigate two attention strategies, by considering the point-to-ground distance for each point as features, or learning interaction of the ground points and non-ground points automatically. Extensive experiments on a recently published large-scale point cloud dataset and a newly collected dataset specifically designed for autonomous driving show that our method outperforms other state-of-the-art methods, both quantitatively and qualitatively. The example shown in Figure 1 demonstrates the effectiveness of the proposed method, especially on segmenting small objects.

To sum up, the main contributions of our work are as follows:

- We propose a ground-aware attention network for semantic segmentation of sparse LiDAR point clouds in autonomous driving scenarios.
- We propose a ground-aware attention module that effectively models long-range dependencies between ground and objects in sparse LiDAR point clouds.
- Extensive experiments on two large-scale urban datasets show that our method achieves state-of-the-art performance and outperforms existing methods by a large margin.

2 RELATED WORK

Scene Understanding in Autonomous Driving. Autonomous driving has gained increasing attention in recent years. Accurate scene understanding of such outdoor environments is vital for autonomous driving. The main tasks of scene understanding can be categorized as object detection and semantic segmentation. Earlier works [21, 25, 30, 31, 39] have achieved great progress of object detection in autonomous driving. However, the bounding box representation only provides rough localization without sufficient semantic details. Semantic segmentation presents more detailed point-wise segmentation which is important for visual perception in autonomous driving. Such semantic segmentation not only provides information for decision making but also provides strong support for accurate localization that is vitally crucial in many applications. 2D semantic segmentation [33, 36, 43] for autonomous driving has been studied recently. These techniques well exploit the texture information in 2D images, however, can not be applied to 3D point clouds that are not regularly organized in 3D space.

Semantic Segmentation of Dense Point Clouds. Traditional approaches [2, 32] that deal with point clouds data process each point separately by extracting hand-crafted features in a local neighborhood. Recent years, deep learning has been widely applied to 3D point cloud segmentation and made great progress via learning more comprehensive and discriminative features. PointNet [27] learns point-wise features with multilayer perceptrons (MLPs), and extracts global features with max-pooling. However, it does not capture local structures, which limits its generalizability to complex scenes. The limitations were later addressed by PointNet++ [26], which designs a hierarchical structure to capture local region features by exploiting increasing contextual scales in metric spaces. Notwithstanding promising results have been achieved in indoor scenes, either of these methods cannot be well generalized to large-scale point clouds of outdoor scenes. PointCNN [17] is proposed to learn a transformation of the input points for the feature weighting and point reordering and then apply typical CNN architecture to process irregular and unordered 3D points. PointSIFT [10] employs a module to encode information from different orientations for indoor scene segmentation. Other methods, such as SEGCloud [16] and OctNet [29], use voxel or octree to represent features of point clouds. Unfortunately, these methods require drastically increasing memory for large-scale LiDAR point clouds.

Semantic Segmentation of Large-Scale Point Clouds. To deal with large-scale outdoor point clouds, SPG [14] coarsely segments a point cloud into superpoints and constructs a graph to represent contextual relationships between those parts. Promising improvement has been made by SPG compared to previous methods on dense point cloud data. However, with regard to sparse LiDAR point clouds, large-scale LiDAR data for supervised 3D semantic segmentation is usually very scarce, because of the heavy workload for

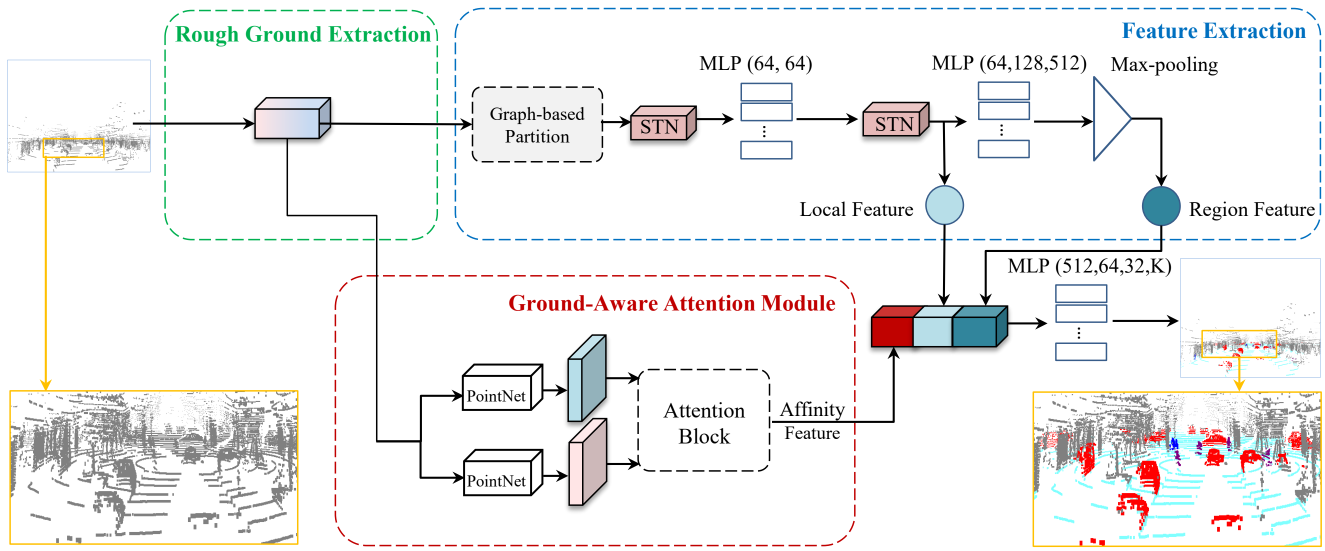


Figure 2: An overview of the proposed ground-aware network for point cloud semantic segmentation. We first roughly extract ground points by multi-section plane-fitting from the input point cloud. Then the point cloud is partitioned into multiple regions for local feature extraction. For each region, point features and region features are extracted using MLPs. The roughly segmented ground and object points are fed into our ground-aware attention module to capture long-range dependencies between points and result in ground affinity features. By concatenating the affinity feature, point feature, and region feature, each point is classified into K categories for the final prediction of semantic labels.

human-involved data annotation. Some recent approaches attempt to investigate semantic segmentation problem for autonomous driving scenes by incorporating 2D image views [38, 40] or using synthetic data [3, 34]. SqueezeSeg [40] transforms 3D point clouds to dense 2D grid representation using spherical projection and utilizes 2D CNN and CRF for semantic segmentation. Following that, PointSeg [38] improves the CRF process of SqueezeSeg to give more consideration to local information. SqueezeSegv2 [41] improves the model of SqueezeSeg with a Context Aggregation Module to increase its robustness to dropout noises. However, there exists a large gap between real sparse 3D point clouds and 2D representation or synthetic data.

Attention Mechanism. Traditional CNNs rely on convolutions to extract features of local regions but ignore long-range dependencies. Recently, the attention mechanism has attracted great interests in different areas [4, 11, 22, 28, 44], showing its great ability in modeling long-range dependencies. [35] applies self-attention to capture global dependencies in sequential data for machine translation and demonstrated its effectiveness. [23] combines the self-attention mechanism with autoregressive models and proposes an image transformer model in image generation. [37] utilizes the self-attention mechanism as a non-local operation to model long-range spatial-temporal dependencies for video processing. Inspired by the success of attention mechanism in various tasks, we propose a new ground-aware framework to exploit long-range dependencies between objects and ground points with attention mechanism for semantic segmentation of sparse LiDAR point clouds.

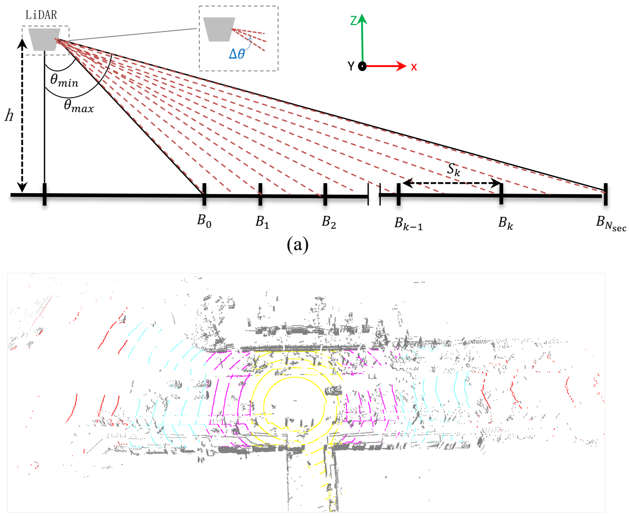
3 METHOD

As shown in Figure 2, our framework of 3D point cloud semantic segmentation mainly consists of three parts, including a rough ground extraction module which roughly segments the input point cloud into ground points and object points, a feature extraction module to extract local and region features, and a ground-aware attention module to exploit long-range dependences between points. In the following subsections, we will introduce each of the above modules in terms of the functionality and the specific architectures.

3.1 Rough Ground Extraction

Due to the different point distribution of ground and objects, we first roughly segment the input LiDAR point cloud $\mathcal{P} = \{p_1, \ldots, p_N\}$ into two subsets $\mathcal{P}_{ground}$ and $\mathcal{P}_{objects}$ by simply fitting the ground planes. In urban scenes, the ground is usually not an ideal plane. Meanwhile, LiDAR devices introduce signal noises when the scanning distance is long. Therefore, a single plane may not be sufficient and robust enough to represent the ground surface in practice. We employ a multi-section plane fitting approach to fit the ground surface and extract ground points from the input point cloud.

Firstly, we divide the input point cloud into multiple sections along the driving direction of the vehicle. Generally, the scanning rays are evenly distributed in angle with an interval of $\triangle\theta$, thus the point density varies greatly at different scanning distance. We divide the input point cloud according to the same angle intervals, as illustrated in Figure 3(a). Taking the part of the point cloud in front of the LiDAR device as an example, we split the points into N_{sec}



(a)

(b)

Figure 3: Multi-section ground plane fitting. (a) Partition by scanning lines. Each red dashed line represents a Velodyne-32 LiDAR scan line. (b) Extracted ground points in multiple sections. Different colors indicate different sections.

sections by computing a set of region boundaries $\{B_k\}_{k=0,\ldots,N_{sec}}$ along the driving direction as

$$B_k = h\tan(\theta_{min} + k\mu\triangle\theta),\ k = 0, ..., N_{sec}, \quad (1)$$

where h is the height of the LiDAR device on the ground. θ_{min} and θ_{max} denote the range of the scanning angles of the LiDAR device. μ denotes the number of scanning rays in each section. All the 3D points whose x coordinates fall into $(B_{k-1}, B_k]$ are divided into the section S_k.

For each section along the driving direction, we estimate a plane using RANSAC [6]. Since there are points of both the ground and objects, we first pick out possible ground points whose $y \in [y_l, y_r]$ and $z \in [z_u, z_b]$, where y_l, y_r, z_u, z_b are predefined as the range of seed ground points on y-direction and z-direction, respectively. Then a plane P_k is fitted for these selected points using RANSAC. After extracting the ground surface of the entire point cloud, we utilize the fitted planes to distinguish non-ground and ground points based on distance measurement. For a point at position $\mathbf{p} = (x, y, z)$ in a section S_k, if its distance to the plane $d(p, P_k) < \sigma$, it is temporally classified as a ground point. Otherwise, it is classified as an object point. Figure 3(b) shows an example of the extracted ground points in multiple sections in a sparse point cloud.

3.2 Region Feature Extraction

It is challenging to represent and extract discriminative features from extremely sparse and large-scale point clouds. Recent studies like [14] have shown the successful application of graph-based partition in large urban scenes. Instead of classifying individual points, the point cloud is partitioned into superpoints as geometrically simple primitives to reduce the scale of the whole point cloud. Then a graph is built to model the relationship between adjacent superpoints. In our approach, we also employ a graph-based partition of the input large-scale point cloud. Since the input LiDAR point cloud $\mathcal{P}$ has been divided into ground points $\mathcal{P}_{ground}$ and objects $\mathcal{P}_{objects}$ in the first stage, we run graph-based partition in the two subsets separately.

After graph-based partition, we employ PointNet [27] to extract features for each group of points that are clustered to the same superpoint. The detailed architecture is shown in the feature extraction block in Figure 2. A spatial transform network (STN) is employed to align the points in each superpoint to a canonical space in the position or feature level. After the second STN, we obtain a 64-D feature vector for each point as its local feature and 512-D region feature vector for the superpoint after two MLPs and a max-pooling layer. For each point in the point cloud, we concatenate its local feature and region feature and get a 576-D feature vector. However, these features are relatively local and limited in a small region. Then we propose to use attention mechanism to capture long-range dependencies of different regions in large-scale sparse LiDAR point clouds, essentially with the support of ground planes.

3.3 Ground-Aware Attention Module

Traditional CNNs implicitly model the dependencies across different local regions using convolutions with different kernel size, resulting in difficulty to represent long-range dependencies between regions that are far away from each other. Recently proposed attention architectures have achieved great interests in a wide range of applications [4, 22, 28, 44], showing its superiority in modeling long-range dependencies. We extend the attention scheme to 3D point cloud and propose a new ground-aware attention module to fully exploit the ground support for semantic segmentation. Our ground-aware attention benefits from a cross-attention between the ground and objects and is tailored for point cloud semantic segmentation in autonomous driving scenarios. To the best of our knowledge, this is the first attempt to apply the attention scheme to 3D point cloud semantic segmentation in a cross-attention manner.

The ground-aware attention module is designed to guide the network to focus on ground-aware information. In order to fully exploit the long-range dependencies between objects and the ground, we explore two different attention architectures, including a hard attention module (Figure 4 (a)) and a soft attention module (Figure 4 (b)), to build the long-range dependencies of 3D points.

Hard Attention. An intuitive way of incorporating the ground knowledge is to directly use the distance to the ground surface as an extra channel for feature embedding. Specifically, we employ two feature embedding branches using two PointNets to extract features for the entire point cloud $\mathcal{P}$. One branch takes the position (x, y, z) of each point as input and extracts N position-only features $\boldsymbol{f}$ for the N points. The other branch takes (x, y, z, d_g) as input, where d_g represents the distance of the point to the fitted ground plane and extracts N distance-associated features $\boldsymbol{g}$, as Figure 4 (a) shows. We then employ an attention block to model the long-range dependencies between points according to their embedded features to implicitly represent the support of ground information.

Soft Attention. Although the proposed hard attention scheme is able to simply incorporate the ground information, such an explicit

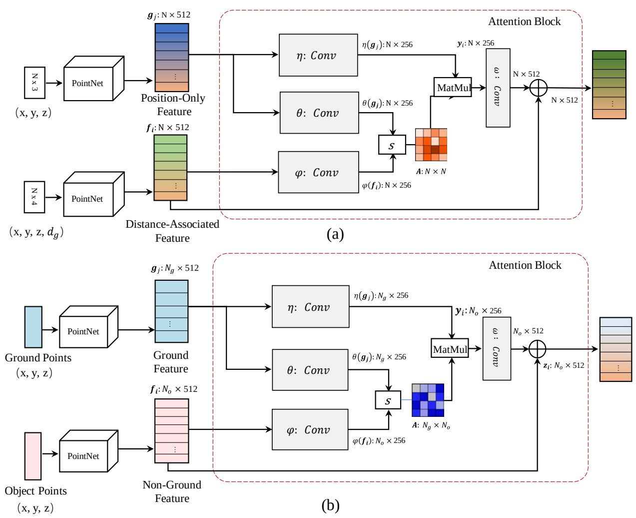


Figure 4: Two ground-aware attention modules: (a) hard attention module and (b) soft attention module. s is the affinity function and $\oplus$ represents the element-wise sum.

distance is not sufficient to capture the underlying relationship between the ground and the points of objects. We come up with a soft attention module to learn the weighting and fusing strategies of the two parts, ground and non-ground point features. After the rough ground extraction, the two point-sets $\mathcal{P}_{ground}$ and $\mathcal{P}_{objects}$ are separately processed by two PointNets for high dimensional feature embedding, as shown in Figure 4 (b). The embedded features $\boldsymbol{f}$ for the N_o object points and $\boldsymbol{g}$ for the N_g ground points are fed into the attention block.

Attention Block. The detailed architecture of the attention module, which is used both in our hard attention module and soft attention module, is shown in Figure 4. Taking two branches of features as input, one branch $\boldsymbol{g}$ carrying ground information while the other branch $\boldsymbol{f}$ carrying point position information only, the designed attention block computes the mutual affinity between points and fuses ground features according to the affinities. Therefore, for the point p_i with non-ground feature $\boldsymbol{f}_i$, the affinity between point p_i and point p_j with a ground feature $\boldsymbol{g}_j$ is given by

$$s(\boldsymbol{f}_i, \boldsymbol{g}_j) = \varphi(\boldsymbol{f}_i) \cdot \theta(\boldsymbol{g}_j), \tag{2}$$

where φ and θ are two projection functions (implemented as a convolutional layer of kernel size 1) to embed feature for the object and ground points respectively. There are other choices for the affinity function s, such as a Gaussian version $s(\boldsymbol{f}_i, \boldsymbol{g}_j) = e^{\varphi(\boldsymbol{f}_i) \cdot \theta(\boldsymbol{g}_j)}$. In our experiments, we found that our approach is insensitive to difference choices, similar to the observation in [37]. An affinity matrix $\boldsymbol{A}$ between the ground features and object features is achieved to fuse the ground features according to the affinities for each point p_i as

$$\boldsymbol{y}_i = \frac{1}{C_i} \sum_{\forall j} s(\boldsymbol{f}_i, \boldsymbol{g}_j) \cdot \eta(\boldsymbol{g}_j), \tag{3}$$

where C_i is a normalization factor $C_i = \sum_j s(\boldsymbol{f}_i, \boldsymbol{g}_j)$. We then add it to the non-ground feature for each point after a convolutional layer ω with kernel size of 1. The weighted feature can be seen as a residual to the originally extracted feature for complementing it with the ground-fused feature as

$$\boldsymbol{z}_i = \omega(\boldsymbol{y}_i) + \boldsymbol{f}_i. \tag{4}$$

Taking the soft attention module as an example, our attention block computes the affinities between object points and ground points, fuses the features of ground points, and adds the fused information to the original non-ground feature as complementary.

After the attention module, we obtain the features embedded with ground affinity in the dimension of $N \times 512$ which carries long-range contexts of LiDAR point cloud. We concatenate the affinity feature with the local and region features which are acquired in the stage of feature extraction with $N \times 576$. Finally, a few MLPs are utilized to output per-point scores for the K categories.

3.4 Ground-Aware Loss Function

Our whole model is optimized in a weakly-supervised manner, within which the ground information is acquired automatically as described in Sec. 3.1. Although the initial segmentation for ground points may not be perfectly right, these pesudo labels provide weak supervision for the network training, without any human-involved annotation. We empirically found that the LiDAR data for autonomous driving presents a severe imbalance distribution among different categories. For instance, *pedestrians* and *bicycles* have much fewer samples compared to the other categories like *vehicles*, while the background points occupy most of the scenes. In order to increase the segmentation accuracy of small objects, we use a class-balanced cross-entropy loss, given as

$$L_{ground-aware} = -\frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{k=1}^{K} \alpha_k \boldsymbol{y}_{ik} \log \boldsymbol{p}_{ik}, \tag{5}$$

where $\boldsymbol{p}_{ik}$ represents the probability of the ith point belonging to the kth category, with a total of N_t points in the whole training set. The weight for each category is computed as $\alpha_k = w_{med}/w_k$, where $w_k = \sum_{i=1}^{M} N_{ik}$ counting for the total number of points of the kth category in the entire training set. w_{med} is the median of w_k over all the K categories.

4 EXPERIMENTS

In this section, we evaluate our framework on two datasets of sparse LiDAR point clouds, the DF-3D semantic dataset [5] and a newly collected dataset in urban road scenes with our own Velodyne HDL-32E LiDAR device which is equipped on an autonomous driving car. A series of quantitative and quantitative comparisons will be shown to demonstrate the effectiveness of our ground-aware approach.

4.1 Datasets and Implementation Details

LiDAR devices that continuously launch and receive multi-beams lasers at 360 degrees become pervasive for environment perception. In real applications for autonomous driving, more and more companies (e.g., Uber, Ford, Baidu, Alibaba, etc.) adopt 32-channel LiDAR in their autonomous cars. A 32-channel LiDAR is much cheaper compared to a 64-channel LiDAR. In addition, 32-channel

Table 1: Quantitative comparison with state-of-the-art methods on the DF-3D dataset [5]. *mIoU* represents the average IoU, while the *OA* indicates the overall accuracy.

Methods	small mot	crowds	pedestrian	cyclist	tricycle	big mot	others	m*IoU* (%)	*OA* (%)
3D FCN [15]	22.7	1.2	0.6	4.7	2.1	21.4	6.2	8.4	10.1
PointNet [27]	45.8	3.1	2.2	8.4	5.3	54.4	13.3	19.0	22.6
PointNet++ [26]	48.3	2.7	3.9	10.5	5.6	50.1	12.9	19.2	23.0
PointCNN [17]	50.4	3.3	6.8	8.2	6.2	46.9	15.2	19.6	23.3
SPG [14]	68.5	9.8	8.4	19.2	7.3	60.1	23.2	26.8	30.2
SqueezeNetv2 [41]	70.5	8.6	9.2	16.6	7.2	55.8	24.1	27.4	30.9
Ours (Vanilla)	70.4	10.3	10.8	21.9	10.2	68.1	23.9	30.8	33.6
Vanilla + CBL	70.2	11.0	11.2	22.2	9.9	68.9	23.8	31.0	34.6
Vanilla + CBL + Hard-Att	70.3	11.8	11.5	23.3	10.5	70.0	23.8	31.6	35.3
Ours	**71.1**	**12.6**	**13.3**	**24.0**	**10.9**	**71.5**	**24.6**	**32.6**	**37.3**

Table 2: Quantitative comparison with state-of-the-art methods on Semantic-LiDAR dataset.

Methods	vehicle	cyclist	pedestrian	tricycle	others	m*IoU* (%)	*OA.* (%)
3D FCN [15]	46.0	1.2	1.5	6.6	33.0	17.7	19.8
PointNet [27]	67.1	1.1	4.9	12.7	32.6	23.7	25.5
PointNet++ [26]	72.2	3.1	9.4	16.5	40.7	28.4	29.8
PointCNN [17]	72.5	8.7	11.3	14.8	44.3	30.3	32.5
SPG [14]	76.3	4.4	9.1	17.9	42.2	30.0	32.1
SqueezeNetv2 [41]	78.2	16.6	14.8	18.2	45.4	34.6	36.6
Ours	**82.3**	**22.3**	**24.0**	**18.5**	**52.1**	**39.8**	**43.2**

LiDARs have smaller volumes that can be easily equipped on vehicles, making it more suitable for large-scale applications. While 64-channel LiDAR sensors have also been utilized in some previous studies (e.g., 3D object detection in the KITTI dataset [7]), the more challenging and much sparser data captured by 32-channel LiDAR has not been well explored, especially for the task of 3D point cloud semantic segmentation. In our experiments, we focus on the sparse data captured using 32-channel LiDAR.

DF-3D Dataset. The DF-3D dataset [5] is published by Alibaba® for 3D semantic segmentation in autonomous driving scene. It was collected with a Velodyne HDL-32E LiDAR sensor on a moving vehicle on urban streets for the purpose of evaluating perception for autonomous driving. This dataset contains 80,000 frames, in which 50,000 point-wise labeled frames are used for training and 30,000 for testing. Each frame contains approximately 50,000 3D points. The semantic labels of the objects above ground are manually annotated. The annotations contain seven classes (*cyclist, pedestrian, tricycle, small mot, big mot, crowds, and others*) in total. A point of the "others" category is likely to be an obstacle or moving objects on streets, but does not belong to any of the other six categories. The background and roads are not annotated. We treat these points as the "unlabeled" category in our segmentation network. Since the labels for the test set are not available, we randomly split the training set into a new training set and a new testing set, containing 35,000 and 15,000 frames, respectively. Each 3D point is represented by its position (x, y, z) in the global coordinate system.

Semantic-LiDAR Dataset. In addition to the DF-3D dataset, we collected a new dataset with a Velodyne HDL-32E LiDAR sensor for a more thorough evaluation. In contrast to the DF-3D dataset, we define five categories including *cyclist, pedestrian, tricycle, car, and others*. Our dataset contains 3,000 frames in total, of which 2,400 frames act as the training set and 600 frames as the test set.

Implementation Details. We implemented the proposed model using PyTorch [24]and trained it on four GTX 1080 Ti GPUs. The optimization is achieved by the Adam optimizer [13] with the initial learning rate as 0.01. We set the batch size as 20. The model was trained for 300 epochs with the learning rate decay of 0.7 at epochs 150, 200, and 250.

4.2 Quantitative Evaluation

To quantitatively evaluate our method and compare with other approaches, we use three metrics that are commonly applied in prior works [8] for large-scale outdoor scenes: the Intersection over Union(*IoU*) over each category, the average IoU (mIoU) over all the categories, and the overall accuracy (OA).

In Table 1, we provide quantitative results of our approach compared with other state-of-the-art methods on the 3D-DF dataset, while the comparison on our newly collected dataset is shown in Table 2. From the results, we can see that our method achieves superior performance compared to the state-of-the-art approaches for 3D point cloud semantic segmentation on both datasets.

More importantly, semantic segmentation of small objects is a challenging problem [12, 18, 36]. As the results shown in Table 1 and Table 2, previous methods have difficulties to accurately segment points for small objects, such as *crowds, pedestrian, and tricycle* in such sparse point clouds. For instance, the *pedestrian* category in the LiDAR data only has several points in the scene. Our method performs better than the state-of-the-art approaches by a large margin on small objects. The large performance gain owes to the effectiveness of our ground-aware framework that fuses object-ground affinity via the ground-aware attention module. Our

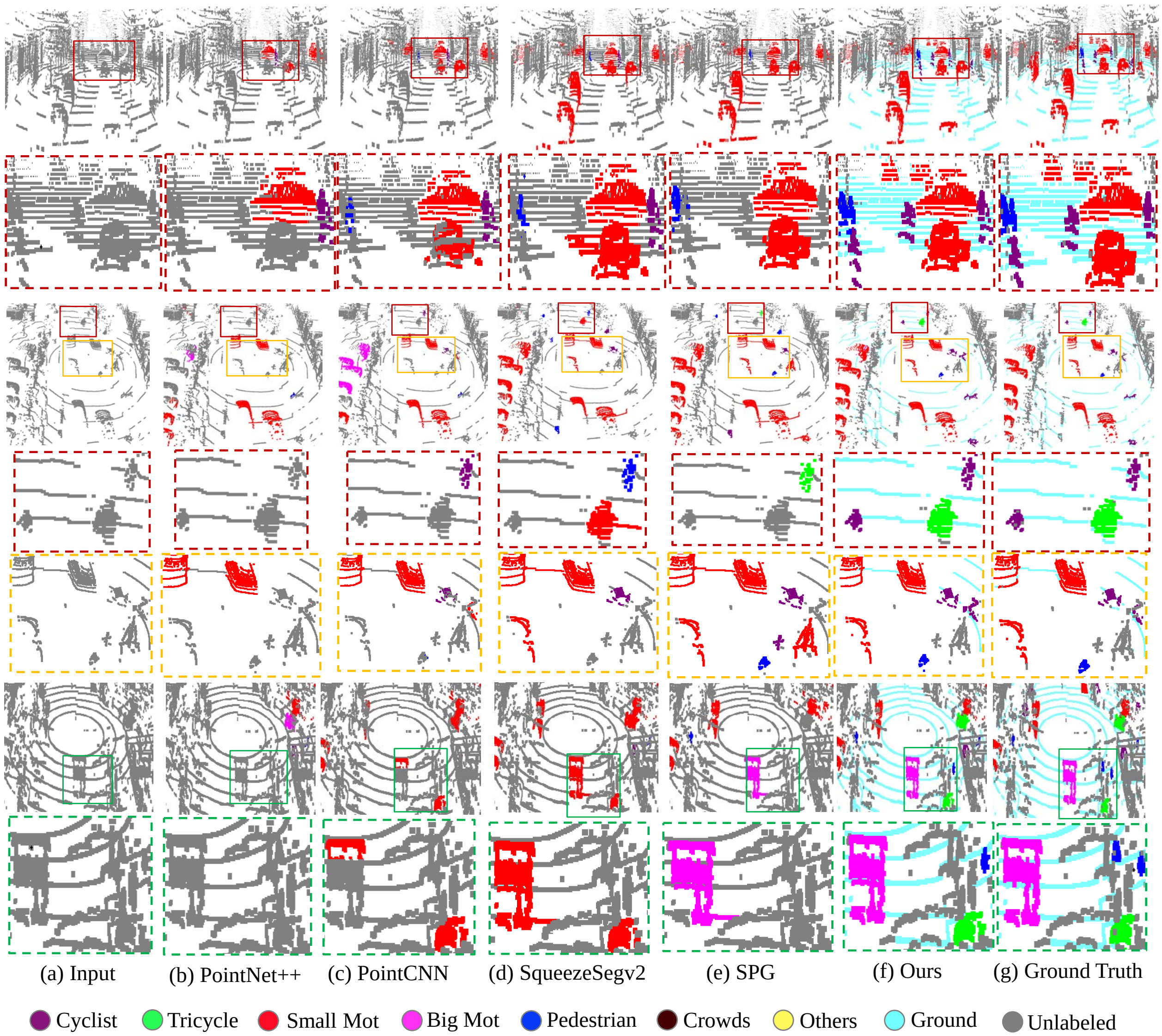


Figure 5: Three groups of semantic segmentation results on the 3D-DF dataset using different methods. From left to right: the input point cloud, the semantic segmentation results of PointNet++ [26], PointCNN [17], SqueezeSegv2 [41], SPG [14], the result from our method, and the corresponding ground truth. For each group, we also show close-ups of one or two local regions to demonstrate the effectiveness of our method for segmenting small objects. The cyan ground points in (g) are not manually annotated but extracted using the ground extraction module described in Sec. 3.1 as pseudo labels.

approach well exploits long-range dependencies between small objects and the ground for accurate segmentation.

4.3 Qualitative Performance

The qualitative results are shown in Figure 5 for the 3D-DF dataset. The results show that our framework outperforms state-of-the-art methods both locally and globally. Specifically, we show the results in details in close-up views in Figure 5. In the first example (the red box), our method correctly segments all the small objects above the ground, while the other methods fail to segment pedestrians and cyclists in this example. Our approach is robust for those *small mot* points which are far away from the LiDAR device, demonstrating the benefit from long-range dependencies between objects and the ground. In the second example, when many vehicles are crowded in a small area, existing methods can hardly distinguish the vehicles and the ground (orange box), or misclassify the vehicles that are close to each other (red box). In comparison, they are correctly classified by our ground-aware approach. In the third example (green box), we can observe that our method is robust for *big mot* which is easy to be confused with the category of *small mot*. These improvements and the robustness of our method mainly come from the proposed ground-aware architecture.

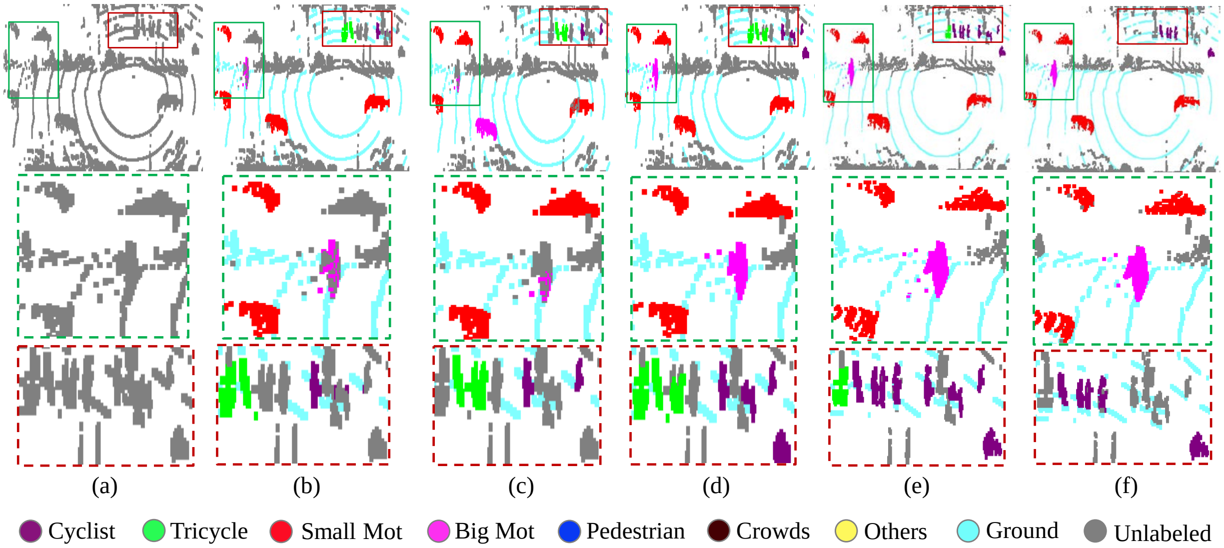


Figure 6: Qualitative comparison of our model under different configurations. (a) Input point cloud. (b) Vanilla model. (c) Vanilla + Class-Balanced Loss (CBL). (d) Vanilla + CBL + Hard Attention. (e) Vanilla + CBL + Soft Attention. (f) Ground Truth.

4.4 Ablation Study

To illustrate the effectiveness of our architecture and understand the influence of each proposed module better, here we present an ablation study. In Table 1 (lower part) we also show the quantitative performance of our method with different architecture configurations on the DF-3D dataset.

Compared to existing approaches, our model takes the ground as an extra category to the annotated categories in the DF-3D dataset and uses the extracted ground points by multi-section plane fitting as pseudo labels to train the model. The first vanilla version of our model only consists of the rough ground extraction module and the feature extraction module, trained using the cross-entropy loss without class balance. We can see that by adding the ground information, even with pseudo labels, the performance is improved by a large margin, especially for the categories of small objects, such as pedestrian and cyclists. In the second version (Vanilla + CBL), we replace the loss function with the class-balance cross-entropy loss that is defined in Eq. (5). Compared with the performance gain from simply adding the ground category, the promotion here is relatively small, which indirectly reveals that the original feature extraction network does not make good use of the ground feature information in sparse point clouds. Nevertheless, by adding the class-balanced loss function, the performance is improved by a certain margin. In the third version (Vanilla + CBL + Hard-Att), we add the ground-aware attention module but with the hard attention scheme. The bottom line is our full model with the proposed soft attention scheme. With the introduction of our final ground-aware soft attention block, we can observe that the performance of each category is improved to a large extent. This demonstrates that the soft attention module learns more effective context between objects and the ground by separately embedding features for the ground and objects and modeling the long-range dependencies between them. In comparison, the hard attention scheme, which directly takes the distance to the ground plane as an extra channel, does not represent the relationship between points as effective as the soft attention scheme.

Figure 6 compares the segmentation results of different versions of our ground-aware model. We can see that our soft attention module effectively learns long-range dependencies between objects and the ground, which significantly improve the segmentation accuracy on small objects in sparse LiDAR point clouds.

5 CONCLUSION

In this paper, we present an effective ground-aware architecture for semantic segmentation on large-scale 3D sparse point clouds for the autonomous driving scenarios. The proposed ground-aware architecture effectively exploits the ground information and captures long-range dependencies between objects and the ground via the proposed soft attention module. Extensive experiments on two new large-scale point cloud semantic segmentation datasets show that our method performs favorably against other state-of-the-art methods both quantitatively and qualitatively, especially on small objects. The proposed ground-aware framework will benefit the 3D point cloud semantic segmentation in outdoor scenes and help promote 3D scene understanding for autonomous driving.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant 2017YFB1300201, the National Natural Science Foundation of China (NSFC) under Grants 61632006, 61622211, and 61620106009, as well as the Fundamental Research Funds for the Central Universities under Grants WK3490000003 and WK2100100030. Jianbo Jiao is supported by the EPSRC Programme Grant Seebibyte EP/M013774/1. This work was partially conducted when Jian Wu was an intern at MoonX.AI.

REFERENCES

[1] A Boulch, B Le Saux, and N Audebert. 2017. Unstructured point cloud semantic labeling using deep segmentation networks. In *Proceedings of the Workshop on 3D Object Retrieval*. Eurographics Association, 17–24.

[2] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. 2003. On visual similarity based 3D model retrieval. In *Computer Graphics Forum*. 223–232.

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3D object detection network for autonomous driving. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 1907–1915.

[4] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 551–561.

[5] DataFountain. [n. d.]. https://www.datafountain.cn/competitions/314/details/rank?sch=1367&page=1&type=A. Sep 4, 2018.

[6] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of The ACM* (1981), 381–395.

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 3354–3361.

[8] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan Dirk Wegner, Konrad Schindler, and Marc Pollefeys. 2017. Semantic3d. net: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 91–98.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 1125–1134.

[10] Mingyang Jiang, Yiran Wu, and Cewu Lu. 2018. PointSIFT: A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation. *CoRR* (2018).

[11] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson W.H. Lau, and Thomas S. Huang. 2019. Geometry-Aware Distillation for Indoor Semantic Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops(CVPRW)*. 1–9.

[13] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

[14] Loic Landrieu and Martin Simonovsky. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 4558–4567.

[15] Bo Li. 2017. 3d fully convolutional network for vehicle detection in point cloud. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1513–1518.

[16] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. 2017. SEGCloud: Semantic segmentation of 3D point clouds. In *International Conference on 3D Vision*. 537–547.

[17] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. 2018. PointCNN: Convolution On X-Transformed Points. In *Proceedings of the International Conference on Neural Information Processing Systems*. 820–830.

[18] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. 2017. Feature Pyramid Networks for Object Detection.. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[19] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2015), 1412–1421.

[21] Daniel Maturana and Sebastian Scherer. 2015. VoxNet: A 3D convolutional neural network for real-time object recognition. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 922–928.

[22] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2016), 2249–2255.

[23] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *International Conference on Learning Representations*. 4052–4061.

[24] Soumith Chintala Gregory Chanan Edward Yang Zachary DeVito Zeming Lin Alban Desmaison Luca Antiga Paszke, Sam Gross and Adam Lerer. 2017. Automatic differentiation in PyTorch.. In *Proceedings of the International Conference on Neural Information Processing Systems Workshop*.

[25] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. 2016. Volumetric and multi-view cnns for object classification on 3D data. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 5648–5656.

[26] R Qi Charles, Hao Su, Mo Kaichun, and Leonidas Guibas. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the International Conference on Neural Information Processing Systems*. 5105–5114.

[27] R Qi Charles, Hao Su, Mo Kaichun, and Leonidas Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 77–85.

[28] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *Proceedings of the International Conference on Neural Information Processing Systems*. 217–225.

[29] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. 2017. OctNet: Learning Deep 3D Representations at High Resolutions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 3577–3586.

[30] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 770–779.

[31] Shuran Song and Jianxiong Xiao. 2016. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 808–816.

[32] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. 2009. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 1383–1392.

[33] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. 2018. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1013–1020.

[34] Robert Varga, Arthur Costea, Horatiu Florea, Ion Giosan, and Sergiu Nedevschi. 2017. Super-sensor for 360-degree environment perception: Point cloud segmentation using image features. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1–8.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems*. 5998–6008.

[36] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 2018. Understanding convolution for semantic segmentation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1451–1460.

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 7794–7803.

[38] Yuan Wang, Tianyue Shi, Peng Yun, Lei Tai, and Ming Liu. 2018. PointSeg: Real-Time Semantic Segmentation Based on 3D LiDAR Point Cloud. *CoRR* (2018).

[39] Zhixin Wang and Kui Jia. 2019. Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection. *arXiv preprint arXiv:1903.01864* (2019).

[40] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. 2018. SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D lidar point cloud. In *IEEE International Conference on Robotics and Automation*. IEEE, 1887–1893.

[41] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. 2018. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. *arXiv preprint arXiv:1809.08495* (2018).

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *IEEE International Conference on Machine Learning*. 2048–2057.

[43] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. 2018. SegStereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision*. 636–651.

[44] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. *IEEE International Conference on Machine Learning* (2019), 7354–7363.