SIP (2020), vol. 9, e24, page 1 of 9 © The Author(s), 2020. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2020.21

INDUSTRIAL TECHNOLOGY ADVANCES

Ground-distance segmentation of 3D LiDAR point cloud toward autonomous driving

JIAN WU[1] AND QINGXIONG YANG[2]

In this paper, we study the semantic segmentation of 3D LiDAR point cloud data in urban environments for autonomous driving, and a method utilizing the surface information of the ground plane was proposed. In practice, the resolution of a LiDAR sensor installed in a self-driving vehicle is relatively low and thus the acquired point cloud is indeed quite sparse. While recent work on dense point cloud segmentation has achieved promising results, the performance is relatively low when directly applied to sparse point clouds. This paper is focusing on semantic segmentation of the sparse point clouds obtained from 32-channel LiDAR sensor with deep neural networks. The main contribution is the integration of the ground information which is used to group ground points far away from each other. Qualitative and quantitative experiments on two large-scale point cloud datasets show that the proposed method outperforms the current state-of-the-art.

Keywords: Computer vision, Image processing, Machine learning

Received 17 March 2019; Revised 9 October 2020

I. INTRODUCTION

In recent years, scene understanding has become increasingly important for the safe navigation of autonomous vehicles in complex environments. Autonomous driving vehicles are always equipped with different types of sensors. One of the most important one is the LiDAR (Light Detection And Ranging) which can provide significantly reliable distance measurements robust to the illumination and surface materials. The LiDAR can directly capture 3D point cloud of the complex environment with wide field of view. Therefore a complete environment information can be collected for the computer vision applications.

Semantic segmentation is the key to scene understanding in autonomous driving, mainly by assigning a category label to each 3D point in the point cloud data captured by the LiDAR. 3D point cloud semantic segmentation has been a crucial topic in recent years [1–8]. Traditional methods of processing 3D point cloud are mainly focusing on extracting hand-crafted features which contain complex post-processing operations, and the performance of these approaches are low due to error accumulation. With great progress in deep learning, recent methods have shown promising results in 2D semantic segmentation [9–11]. However, there is limited study on semantic segmentation of sparse LiDAR point cloud, probably due to the lack of public large-scale semantic segmentation datasets for autonomous driving. Additionally, LiDAR point clouds are relatively sparse and contain irregular, i.e. unstructured, points, which is different from the images. The density of points also varies drastically due to non-uniform sampling of the environment. For LiDAR point cloud data, there is a clear difference between the ground and the objects on the ground. The distribution of ground points is anisotropic, but the points in the obstacle area are mostly linearly distributed. This non-uniform, anisotropic distribution makes it difficult to directly adopt existing methods developed for isotropically distributed point clouds.

To exploit effective strategy for the challenge, this paper proposes an effective approach by integrating ground information with cascaded deep learning module in LiDAR point cloud captured in complex urban scenes. In this paper, we firstly propose a method for automatically segmenting ground points, and then use the ground point estimates to process two different distributions separately, which supports following feature extraction modules. In a sparse point cloud, objects near each other tend to be classified into the same category due to the sparsity. As a result, global information will be effective in the structure, while the dominant component of a practical road environment will be the ground. We assume that the ground information will be very useful for the semantic segmentation of 3D LiDAR point cloud due to its global structure. As the road surfaces cannot be simply represented by a single 3D plane, our approach utilized multi-section plane fitting approach to extract the potential ground planes in a sparse point cloud. Although the ground points are only roughly segmented

[1] University of Science and Technology of China, Hefei, China
[2] Moon X. AI, Shenzhen, China

Corresponding author:
Qingxiong Yang
Email: liiton.research@gmail.com

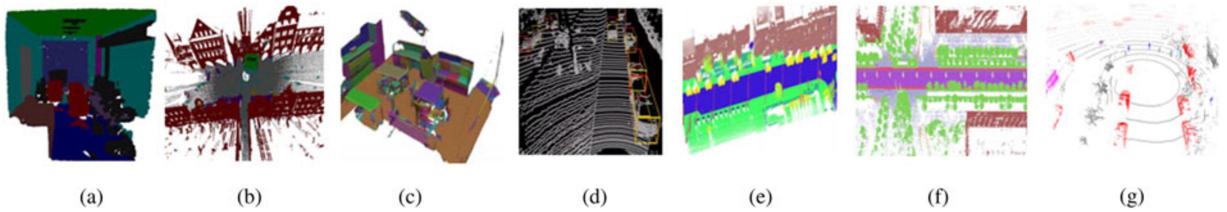

Fig. 1. Different 3D point cloud dataset for semantic segmentation. (a) S3DIS. (b) Semantic3D. (c) ScanNet. (d) KITTI. (e) Oakland3D. (f) Apolloscape. (g) DF-3D.

Table 1. Compared with different 3D point cloud dataset for semantic segmentation

	S3DIS [12]	Semantic3D [13]	ScanNet [14]	KITTI [15]	Oakland3D [16]	Apolloscape [17]	DF-3D [18]	Ours
Scene	Indoor	Outdoor	Indoor	Outdoor	Outdoor	Outdoor	Outdoor	Outdoor
Sparse/Dense	Dense	Very dense	Dense	Sparse	Dense	Dense	Sparse	Sparse
Ground Annot.	Y	Y	Y	N	N	N	N	N
LiDAR	–	–	–	64E	–	–	32E	32E

in the initial stage, the extracted ground points provide rich information for the subsequent feature extraction module, which is conducive to further eliminating the ambiguity between the ground and objects points. After rough segmentation of the ground points, we use deep learning methods for further feature extraction, and establish the connection between the ground points and the object points to enhance the performance of the segmentation Fig. 1.

We also prepare a large-scale dataset for autonomous driving with full semantic segmentation of LiDAR scans. Table 1 shows that it has more information than the publicly-available datasets. For instance, KITTI dataset only provides bounding boxes for the task of detection but is lack of semantic annotation.

Qualitative and quantitative experiments demonstrate that the 3D LiDAR segmentation results can be significantly improved by utilizing the ground information in a weakly-supervised manner.

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

- We build connection between ground points and object points by a ground-distance architecture for semantic segmentation of sparse LiDAR point clouds in autonomous driving scenarios.
- We build a new large-scale dataset for autonomous driving with full semantic segmentation of LiDAR scans.
- Extensive experiments on two large-scale urban datasets show that our method set achieves great performance and outperforms existing methods by a large margin.

II. RELATED WORK

Numerous studies on scene understanding have been conducted. This section briefly discusses some recent works in dynamic outdoor scenes as follows:

- Scene understanding in autonomous driving
- Semantic segmentation of point clouds
- Semantic segmentation of large-scale LiDAR point clouds

A) Scene understanding in autonomous driving

Autonomous driving has received widespread concern and development recently. Scene understanding is one of the key building blocks of autonomous driving in dynamic, real-world environments. Scene understanding tasks are mainly divided into object detection and semantic segmentation. Earlier works [19–23] have made great progress in the detection task of autonomous driving. Unfortunately, the bounding box representation can just acquire rough localization information, but lacks semantic details to distinguish different objects. Semantic segmentation provides annotation for each 3D point, which is essential for the visual perception task, because a self-driving vehicle needs to give different attention to different objects during driving, therefore semantic segmentation can provide not only the important information for vehicle decision-making, but also a powerful auxiliary role for precise positioning, which is essential in many autonomous driving applications. Previous approaches have achieved promising results in 2D semantic segmentation [24–26]. They do make good use of the texture information in 2D images. However, it is a great challenge to apply them to the unstructured 3D points.

B) Semantic segmentation of point clouds

Traditional 3D semantic segmentation approaches [27,28] mainly focus on extracting hand-crafted features which contain complex post-processing operations as presented in [29,30]. Meanwhile, these approaches usually need a lot of parameters and are difficult to tune, and thus the performance is relatively low compared to the recent deep neural networks-based approaches. Deep learning has been widely explored in 2D image segmentation. But until recently, significant progress was achieved by learning comprehensive and differentiated characteristics of 3D point cloud. [4,20,31] propose to represent the point cloud as a high-dimension volumetric form which can be applied with 3D convolutional neural networks. This representation is constrained by its resolution due to data sparsity and

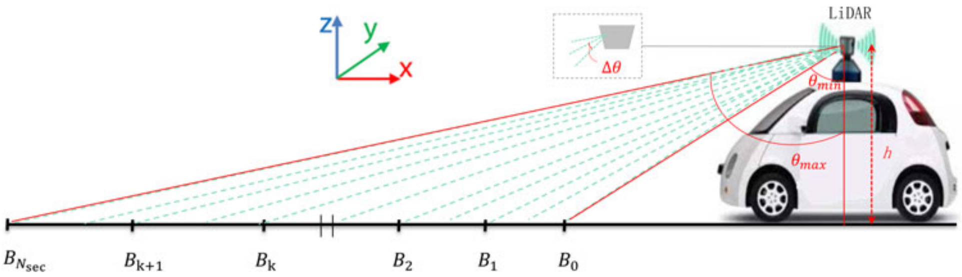

Fig. 2. Multi-section ground plane segmentation. The ground plane was divided by the scanning beams. Each green dashed line represents a beam from the Velodyne HDL-32E Lidar used in this paper.

the redundant computation of 3D convolution operations which is not suitable for LiDAR point clouds. PointNet [7] introduces a new architecture for the unordered and irregular raw 3D point cloud. By learning high-dimensional features with several MLPs and extracting local and global features with max-pooling, it is the first attempt to use the un-ordered point clouds as the input and apply symmetrical operators that are able to deal with ordering problem of the 3D point cloud. Finally, max pooling is used to extract global and permutation-invariant features. However, it is difficult to extract local information, which limits its generalizability to more complex environment. This problem was tackled by designing a hierarchical structure to capture more local structures in PointNet++ [6]. By applying individual PointNet architectures in a local area, PointNet++ captures local information and then applies this hierarchical architecture to exploit increasingly contextual architectures in a high-dimensional space. Although high performance has been demonstrated in the indoor scenes, neither PointNet nor PointNet++ can be well generalized to the sparse outdoor point clouds captured from a 32-beam LiDAR sensor equipped on a self-driving vehicle. PointCNN [5] introduces a new transfer module to learn weighting information from the original data which can benefit from the irregular and unorder 3D data, then typical 3D architecture is applied to obtain final segmentation results. Inspired by the SIFT feature extractor [32], PointSIFT [2] employs local octant directional vectors as feature extraction layers which show good results in indoor scenes. OctNet [8] uses octree architecture to express input point clouds. The memory cost of this method is too high for the autonomous driving application.

C) Semantic segmentation of large-scale LiDAR point clouds

The primary challenge of LiDAR perception on the whole may be the sparse and large-scale characteristic of 3D point cloud. SPG [3] manages to summarize the local relationships in a similar fashion to PointNet by coarsely segmenting a point cloud into a superpoint graph (SPG). A SPG is designed for the large-scale 3D outdoor point clouds. Superpoints are locally coherent, geometrically homogeneous groups of points that get embedded by feature extractors. SPG outperforms the previous methods designed for dense point clouds. Due to the heavy workload for data annotating, a large-scale LiDAR dataset captured by a low-resolution LiDAR sensor for semantic segmentation was not available. Approaches were proposed by projecting the 3D LiDAR point clouds to a number of 2D images from different viewing directions [33,34] or using synthetic data [35,36]. SqueezeSeg [34] uses a spherical projection of the point cloud enabling the usage of 2D convolutions and conditional random fields (CRF) for semantic segmentation. This representation allows to perform the segmentation by utilizing a light-weighted full convolutional network, and the last step will join the 2D segmentation results to form the 3D segmentation results. PointSeg [33] improves the CRF procedure of SqueezeSeg to provide more local details. SqueezeSegv2 [37] improves the architecture of SqueezeSeg with a new module named Context Aggregation Module and uses batch normalization and focal loss strategies to improve its robustness and effectiveness. However, the performance of these approaches is relatively low for the autonomous driving application, especially on small objects such as pedestrians and cyclists. Additionally, there is a gap between the actual sparse 3D point cloud and their 2D projections or the corresponding synthetic data Fig. 2.

III. APPROACH

Given a 3D LiDAR point cloud which can be acquired by the LiDAR sensor, we aim at assigning a semantic label to each 3D point. Figure 3 gives a brief overview of the proposed architecture which consists of three modules: (1) The ground plane fitting module separates the input point cloud into the ground and object points, and the pseudo ground annotations will be used for weak supervision in subsequent network. (2) The feature extraction module represents the point cloud as superpoint structure. Local and global feature will be extracted in several MLPs followed by the max-pooling operations. (3) The distance feature extraction module connects the ground points and object points in order to make good use of the ground plane in the point cloud.

A) Ground point estimation

There is a visual discrimination in data distribution between the objects on the ground and the ground. The distribution of ground points is anisotropic, but points in the obstacle area are linearly distributed. This non-uniform, anisotropic

https://doi.org/10.1017/ATSIP.2020.21 Published online by Cambridge University Press

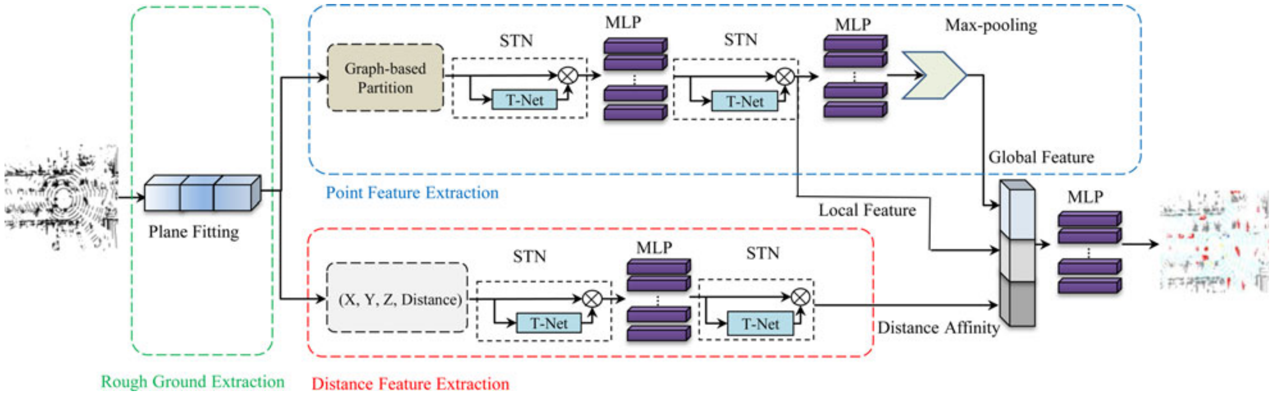

Fig. 3. The proposed framework for point cloud semantic segmentation. We firstly roughly extract ground points from the input point cloud using a multi-section plane-fitting approach. Then the point cloud is separated into multiple parts for feature extraction. Local features and global features will be extracted from each part by utilizing MLPs. The roughly-segmented ground and object points are fed into the proposed distance feature extraction module to capture the relationship between the object and ground points. By concatenating the distance affinity feature, local feature, and global feature, each point is classified into K categories for the final prediction of semantic labels.

distribution makes it difficult to directly adopt the state-of-the-art semantic segmentation approaches which were proposed for isotropically distributed point clouds. The 3D structure of a real-world driving scenario is extremely complex, a ground surface is not a simple 3D plane. As a result, using a single plane model to model the ground surface is not sufficient in the autonomous driving application. In our framework, a multi-segment plane fitting approach was proposed to model the ground points, and ground points are partially divided from a point cloud. The ground point estimates were then used as the pseudo labels for feature extraction. Our approach is motivated by the assumptions that most objects such as vehicles and pedestrians will be above the ground. We suppose that when the ground can be well extracted, then the segmentation accuracy of the other categories can be improved as well. Meanwhile, we empirically find that the ground plays a significant role in the segmentation of sparse point clouds, the objects consisted the number of points, mutual interference among the points leads to segmentation ambiguities.

We first separate the input point cloud into a number of regions along the driving direction (and let it be the x-axis) of a self-driving vehicle.

The variance of the angle interval $\triangle\theta$ between two beams of the LiDAR is normally very small. As a result, the density of point mainly changes on the horizontal distance of the traveling vehicle. As shown in Fig. 2, the input point cloud is mainly divided according to the angular interval of the rays. Take part of the point cloud in front of a moving vehicle for example, we will divide these 3D points into N_{sec} parts by calculating a set of area boundaries $\{B_k\}_{k=0,\ldots,N_{sec}}$ as

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

where h is the vertical distance between the ground and the LiDAR installed on moving vehicle. θ_{min} and θ_{max} represent the range of minimum and maximum scanning angles of the LiDAR. N_{sec} is the default number of sections, and μ denote the number of scanning rays in each block. The 3D points with x coordinates residing in $(B_{k-1}, B_k]$ will be assigned to block S_k. For each divided block, we use RANSAC [38] to estimate a ground plane, and compute the parameters of a plane model using different observations from the data subsets. Points of both the ground and the objects will cause ambiguity for plane fitting. In order to reject outliers, we crop the point cloud by setting a limit to the ranges of the y-axis and z-axis values. When the ground points of the point cloud were extracted, they will be used to distinguish non-ground points from ground points by distance measurement. For a point at position $\mathbf{p} = (x, y, z)$ in a section S_k, if its Euclidean distance to the plane $d(p, P_k)$ is larger then a threshold σ, it will be temporarily considered as an object point. Otherwise, it is classified as a ground point.

B) Feature extraction

Due to long-range scanning characteristic of the point cloud, it is challenging to represent and extract useful information from an extremely sparse point cloud. Previous works usually attempt to utilize the 2D image representation projecting from the 3D points. However, the segmentation result is often a lack of accuracy due to the limited resolution of 2D images. As a result, it is especially not suitable for small objects.

Recent approaches such as [3] have shown the effective progress in the application of graph-based representation of the point clouds in large-scale urban scenes. This method does not classify individual points, but divides them into simple geometric superpoint structures, therefore the scale of the large-scale data can be reduced significantly.

The step of partition is defined as an optimization problem in [3], which could be seen as Lo variant [39] that can quickly find an approximate solution with a few graph-cut iterations. The solution of optimization problem is defined as super point. Then a graph is defined as $G = (S; E; F)$ whose nodes are the set of super points S and edges E represent the adjacency between super points.

The processing method is similar to superpixel methods for image segmentation which divided the point clouds as

https://doi.org/10.1017/ATSIP.2020.21 Published online by Cambridge University Press

several super points. The step is purely unsupervised and makes use of the different dimensionality feature including linearity, planarity, scattering, verticality, and elevation which can maintain the original feature of the 3D point cloud in the whole scene. In the proposed architecture, we also apply a graph-based partition step of the input point cloud. The input large-scale point cloud will be separated into ground parts and object parts. We then apply graph-based step on the ground points and object points individually.

The PointNet [7] is used for point feature extraction due to its simplicity and efficiency. It is applied for each set of points that are clustered as a superpoint. In order to learn spatial information of different parts of point clouds, each superpoint graph is rescaled to the unit sphere before the extracting steps. Figure 3 illustrates the detailed architecture in the point feature extraction module. We employ a Spatial Transform Network (STN) which is combined with T-Net and the transform matrix to align the points in each superpoint to a standard space in the position and feature level.

We also apply Multi-Layer Perceptron (MLP). The values of MLPs have been presented in Fig. 3. The MLPs are used to map the input points independently and identically to a higher-dimensional space which can extract the feature of the point cloud.

MLP returns a feature vector with dimension 64 in local feature and 512 in global feature for each point in the superpoint. For each point in whole point cloud, the local and global feature are aggregated with a 576-D feature vector.

In our architecture, we also build the connection between the ground points and the object points above the ground by utilizing the distance above the ground surface as an extra affinity for feature embedding. The point cloud is computed against the estimated ground model for each point, then we acquire the distances from the point to its projection on the plane model. As shown in the distance feature extraction module in Fig. 3, we use the position and distance as input which is represented as (x, y, z, d), where d represents the distance described above. It allows us to compute a height value with respect to the ground plane. Then the (x, y, z) coordinates and distances are separately analyzed by the partial PointNet with STNs and MLPs. After the embedding module, we acquire the features embedded with distance affinity in the dimension of $N \times 512$. These features extract more information from the point cloud. Finally, different features are concatenated for semantic segmentation.

C) Weighted loss function

To better combine the ground affinity, the entire framework in this paper is trained by cross-entropy loss with the ground estimates computed automatically as described in Section A. It can be seen that there is a severe imbalance in the number of categories in the point cloud dataset, which is biased toward dominant classes rather than the others from the perspective of having higher number of instances. For instance, the number of instances for *vehicles* class is higher than the others. The imbalance condition may lead to lower performance for the class with few instances.

In order to solve the imbalanced dataset problem and give more attention to the small objects, we utilize a weighted cross-entropy loss, which can not only classify different objects correctly, but also enhance mean IoU measurement metric of point-wise classification. Let

$$L_{weighted} = -\frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{k=1}^{K} \alpha_k y_{ik} \log p_{ik}, \tag{2}$$

where p_{ik} represents the probability that the ith point belongs to the kth category, and the input point cloud to network has a total of N_t points. We give different weights to different categories according to their frequency in the dataset $\alpha_k = f_{med}/f_k$, where $f_k = \sum_{i=1}^{M} N_{ik}$ represents the total number of points of kth category in the entire training set. f_{med} is the median of f_k in all K categories.

IV. EXPERIMENTAL RESULTS

In this chapter, we evaluate our architecture on a new sparse LiDAR dataset captured using a velodyne HDL-32E LiDAR on a passenger car in the dynamic scene. This LiDAR has 32 beams, and it is a very popular sensor in the autonomous driving society. This section first introduces this dataset and then the evaluation metric in detail. The implementation details will be introduced, followed by the experimental results (Fig. 4).

A) Datasets

A LiDAR sensor is capable of continuously transmitting and receiving scanning lasers over a 360-degree range. LiDAR is now a key component of a self-driving vehicle as it provides an accurate 3D representation of the driving environment. The 64-beam LiDAR were used in the previous dataset. For instance, the KITTI dataset [15] is applied for 3D object detection in autonomous driving. However, 32-beam LiDAR sensors are more popular recently due to its cost and long-range sensing ability. Meanwhile, it has a smaller size and thus is easier for installation. Due to the lack of data and the sparseness of the point cloud, it is challenging to apply deep neural network to point clouds captured from 32-beam LiDAR, which is the target of this paper.

DF-3D dataset.

Alibaba AI Lab® releases a DF-3D dataset [18], which is a point cloud dataset for the semantic segmentation problem. It contains many complex scenes, including city blocks, residential areas, highway landscapes, and rural roads. This dataset contains 80 000 point-wise labeled 3D scans, while the training/testing set are divided as 50 000/30 000. Each 3D scan contains about 50 000 3D points. The labels contain seven categories: *cyclist, pedestrian, tricycle, small mot, big mot, crowds, and others*. The category of “others” belongs to

https://doi.org/10.1017/ATSIP.2020.21 Published online by Cambridge University Press

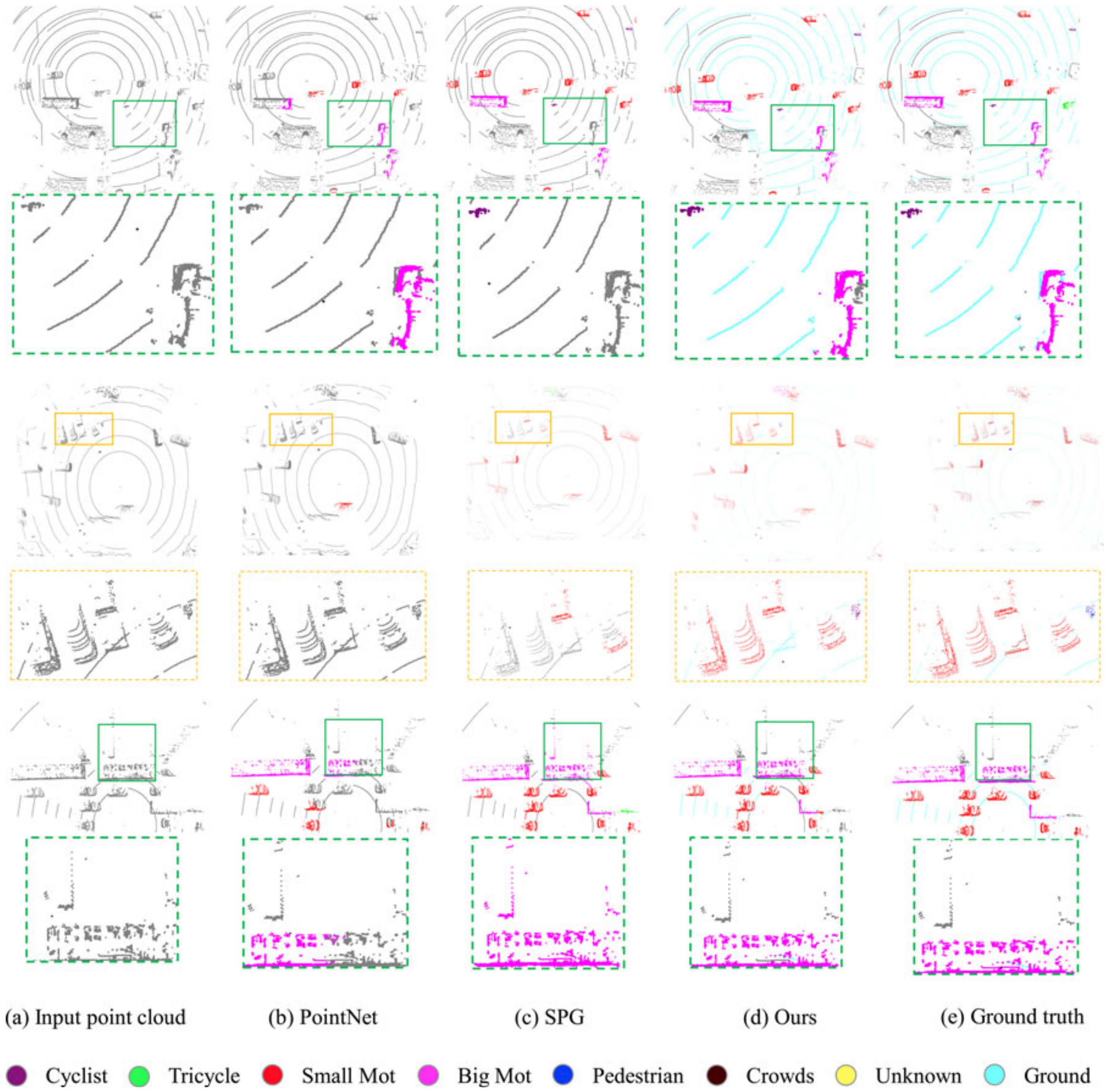

Fig. 4. Visual evaluation of the proposed semantic segmentation approach. From left to right: the input point cloud, the semantic segmentation results of PointNet [7], SPG [3], proposed approach, and the ground truth, respectively. The close-ups are also included to demonstrate the effectiveness for both small and large objects. The cyan ground points in (e) are not manually annotated but segmented using the plane fitting module described in Section A as pseudo labels.

non-moving or moving objects on effective driving area of streets, but is different from the other six classes. The background and ground points are not labeled in this dataset which were regarded as "unlabeled" class. The ground truth of the test set is not available. As a result, we randomly split the original training set into a new training set with 35 000 frames and a new testing set with 15 000 frames as motivated by [40]. Each 3D point has a corresponding position (x, y, z) and the intensity value.

Semantic-LiDAR dataset.

In order to conduct a more comprehensive evaluation of our method, in addition to the DF-3D dataset, we also collect a new dataset. The point cloud data are captured using a similar LiDAR as the DF-3D dataset, with a minor difference for the annotated categories. Different from DF-3D dataset, there are five classes in our own dataset: *cyclist, pedestrian, tricycle, car, and others*, while the training/testing set are divided as 2400/600 with total 3000 frame.

Evaluation metrics.

In order to quantitatively evaluate our method and compare it with the state-of-the-art, a metric that was widely used in previous works [13] in large-scale outdoor scenarios was adopted: IoU(Intersection over Union) over each class, mIoU(average IoU) over all classes, and the OA(overall accuracy).

Implementation details.

The proposed architecture is implemented by utilizing PyTorch [41] framework and trained on 4 GTX 1080 Ti GPUs. The optimization is achieved by the Adam optimizer [42] and the initial learning rate is 0.01. The batch size is set as 20 during training. The network is trained for 300 epochs

https://doi.org/10.1017/ATSIP.2020.21 Published online by Cambridge University Press

Table 2. Evaluation on the DF-3D dataset. We show the quantitative results for each category in DF-3D dataset, our method outperforms the state-of-the-art through out the categories especially the *bigmot* class. 6pt

Methods	Smallmot	Crowds	Pedestrian	Cyclist	Tricycle	Bigmot	Others	m*IoU* (%)	*OA.* (%)
3D FCN [43]	22.7	1.2	0.6	4.7	2.1	21.4	6.2	8.4	10.1
PointNet [7]	45.8	3.1	2.2	8.4	5.3	54.4	13.3	19.0	22.6
PointNet++ [6]	48.3	2.7	3.9	10.5	5.6	50.1	12.9	19.2	23.0
PointCNN [5]	50.4	3.3	6.8	8.2	6.2	46.9	15.2	19.6	23.3
SPG [3]	68.5	9.8	8.4	19.2	7.3	60.1	23.2	26.8	30.2
Ours	**70.2**	**11.3**	**11.1**	**22.3**	**10.1**	**69.1**	**23.7**	**31.1**	**34.7**

Table 3. Evaluation on the Semantic-LiDAR dataset. The quantitative results for each category are shown in the table. Note that the proposed method performs better for both small objects and big objects, resulting in 5% segmentation accuracy improvement on average comparing to the state-of-the-art.

Methods	Vehicle	Cyclist	Pedestrian	Tricycle	Others	m*IoU* (%)	*OA.* (%)
3D FCN [43]	48.5	1.3	1.2	1.2	9.4	12.3	13.8
PointNet [7]	69.5	0.7	5.2	13.8	2.2	13.1	14.8
PointNet++ [6]	73.2	2.6	9.8	**18.6**	5.5	21.9	23.7
PointCNN [5]	70.4	8.3	10.5	14.2	7.7	22.3	24.5
SPG [3]	78.5	3.5	8.5	16.5	4.2	22.2	24.1
Ours	**80.7**	**16.7**	**12.2**	17.2	**10.3**	**27.4**	**29.3**

with the learning rate decay of 0.7 at epochs 150, 200, and 250.

B) Quantitative evaluation

Tables 2 and 3 present the quantitative evaluation results on the 3D-DF dataset and our dataset, which demonstrate that the proposed method outperforms the state-of-the-art for semantic segmentation on sparse point clouds.

The detection or semantic segmentation of small objects is quite challenging in the computer vision community [25,44,45]. Tables 2 and 3 show that the state-of-the-art performs poorly on the small objects (e.g. *crowds, pedestrian, and tricycle*, etc.) on sparse point clouds, while the proposed method has a higher performance around the small objects. For instance, the *pedestrian* category generally has only a few sparse points. It is extremely difficult to identify and perceive the sparse points. However, visual evaluation shows that the better performance can be obtained from the proposed architecture.

C) Qualitative evaluation

Figure 5 presents the visual evaluation on the segmentation results on the 3D-DF dataset, which contains fine-grained semantic categories (e.g. crowds and pedestrian, big mot and small mot) that can effectively illustrate the model performance. Note that the segmentation results obtained from the proposed method are visually more accurate than the state-of-the-art, especially from the close-ups.

As can be seen from the close-ups (in green boxes) of the first row in Fig. 4, the proposed method correctly separates all small and large objects above the ground. However, the other methods either miss the small object category or cannot handle large vehicles. Meanwhile, it can be seen that our method is robust to points far from the LiDAR due to the integration of the ground information.

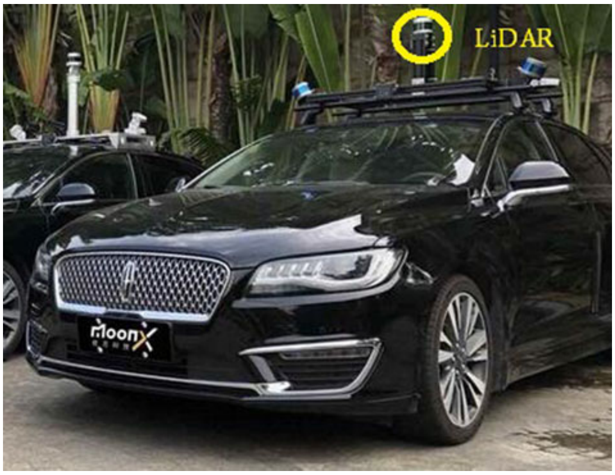

Fig. 5. A picture of the LiDAR and the corresponding passenger car used to captured the proposed dataset.

The close-ups (in orange boxes) of the second row in Fig. 4 show that when there are many vehicles in a small region, the current state-of-the-art methods will fail. The proposed method can better separate this type of crowd scenes. The close-ups (in green boxes) of the third row in Fig. 4 show that our approach is robust to large objects(*big mot*) which is easier to be misclassified in some complex urban scenes. From the visualized results, we can conclude that the proposed approach outperforms the state-of-the-art both quantitatively and qualitatively.

V. CONCLUSION

Autonomous driving is quite challenging and one of the main difficulties is the semantic segmentation of the point clouds acquired from the LiDAR sensor. Due to the requirement of real-time sensing and limited production cost, most of the LiDAR sensors used in self-driving cars can only acquire sparse point clouds. On the other hand, the state-of-the-art semantic segmentation methods were mostly developed for dense 3D point clouds. As a result, this paper proposes an effective architecture for performing semantic segmentation of sparse LiDAR point clouds by implicitly incorporating the ground information. Extensive experiments on two new large-scale point cloud semantic segmentation datasets show that the proposed method performs favorably against the state-of-the-art both quantitatively and qualitatively.

Nevertheless, the proposed framework is not an end-to-end architecture which is likely to be a better solution. The direct use of the distance to the ground surface as an extra channel for feature embedding could be improved as well as it will depend on the accurate extraction of the ground points. A more effective architecture that better utilizes the ground knowledge will be investigated. In this near future, efficient architecture will be also explored to achieve real-time performance, aiming at a better understanding of the 3D environment of a self-driving vehicle.

REFERENCES

[1] Boulch, A.; Saux, B.L.; Audebert, N.: Unstructured point cloud semantic labeling using deep segmentation networks, in *Proc. of the Workshop on 3D Object Retrieval*, 17–24. Eurographics Association, 2017.

[2] Jiang, M.; Wu, Y.; Lu, C.: PointSIFT: a SIFT-like network module for 3D point cloud semantic segmentation. *CoRR*, 2018.

[3] Landrieu, L.; Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 7, 2018, 4558–4567.

[4] Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B.: SEGCloud: semantic segmentation of 3D point clouds, in *Int. Conf. on 3D Vision*, 2017, 537–547.

[5] Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B.: PointCNN: convolution on X-transformed points, in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 2018, 820–830.

[6] Qi, C.R.; Su, H.; Kaichun, M.; Guibas, L.: Pointnet++: deep hierarchical feature learning on point sets in a metric space, in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 9, 2017, 5105–5114.

[7] Qi, C.R.; Su, H.; Kaichun, M.; Guibas, L.: Pointnet: deep learning on point sets for 3D classification and segmentation, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 7, 2017, 77–85.

[8] Riegler, G.; Ulusoy, A.O.; Geiger, A.: OctNet: learning deep 3D representations at high resolutions, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 7, 2017, 3577–3586.

[9] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern. Anal. Mach. Intell.*, **40** (4) (2018), 834–848.

[10] Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P.: Panoptic segmentation, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 7, 2019, 9404–9413.

[11] Long, J.; Shelhamer, E.; Darrell, T.: Fully convolutional networks for semantic segmentation, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 7, 2015, 3431–3440.

[12] Armeni, I., *et al.*: 3D semantic parsing of large-scale indoor spaces, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, 1534–1543.

[13] Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M.: Semantic3d. net: a new large-scale point cloud classification benchmark, in *ISPRS Annals of the Photogrammetry*, Remote Sensing and Spatial Information Sciences, 2017, 91–98.

[14] Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.A.; Nießner, M.: Scannet: richly-annotated 3D reconstructions of indoor scenes, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 2, p. 10, 2017.

[15] Geiger, A.; Lenz, P.; Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 7, 2012, 3354–3361.

[16] Munoz, D.; Bagnell, J.A.D.; Vandapel, N.; Hebert, M.: Contextual classification with functional max-margin markov networks, in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[17] Huang, X., *et al.*: The apolloscape dataset for autonomous driving, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2018, 954–960.

[18] DataFountain: https://www.datafountain.cn/competitions/314/details/rank?sch=1367&page=1&type=A. 4 September 2018.

[19] Maturana, D.; Scherer, S.: VoxNet: a 3D convolutional neural network for real-time object recognition, in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, 922–928.

[20] Qi, C.R.; Su, H.; Nießner, M.; Dai, A.; Yan, M.; Guibas, L.J.: Volumetric and multi-view cnns for object classification on 3d data, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, 5648–5656.

[21] Shi, S.; Wang, X.; Li, H.: Pointrcnn: 3D object proposal generation and detection from point cloud, 2019, 770–779.

[22] Song, S.; Xiao, J.: Deep sliding shapes for amodal 3D object detection in RGB-D images, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, 808–816.

[23] Wang, Z.; Jia, K.: Frustum convnet: sliding frustums to aggregate local point-wise features for amodal 3D object detection. *arXiv preprint arXiv:1903.01864*, 2019.

[24] Teichmann, M.; Weber, M.; Zoellner, M.; Cipolla, R.; Urtasun, R.: Multinet: real-time joint semantic reasoning for autonomous driving, in *IEEE Intelligent Vehicles Symp. (IV)*, 2018, 1013–1020.

[25] Wang, P., *et al.*: Understanding convolution for semantic segmentation, in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018, 1451–1460.

[26] Yang, G.; Zhao, H.; Shi, J.; Deng, Z.; Jia, J.: SegStereo: exploiting semantic information for disparity estimation, in *Proc. of the European Conf. on Computer Vision*, 2018, 636–651.

[27] Chen, D.-Y.; Tian, X.-P.; Shen, Y.-T.; Ouhyoung, M.: On visual similarity based 3D model retrieval, in *Computer Graphics Forum*, 2003, 223–232.

[28] Sun, J.; Ovsjanikov, M.; Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion, in *Computer Graphics Forum*, vol. 28, 2009. Wiley Online Library, 1383–1392.

[29] Feng, C.; Taguchi, Y.; Kamat, V.R.: Fast plane extraction in organized point clouds using agglomerative hierarchical clustering, in *2014 IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014, 6218–6225.

[30] Himmelsbach, M.; Mueller, A.; Lüttel, T.; Wünsche, H.-J.: Lidar-based 3D object perception, in *Proc. of 1st Int. Workshop on Cognition for Technical Systems*, vol. 1, 2008.

[31] Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J.: 3D shapenets: a deep representation for volumetric shapes, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, 1912–1920.

[32] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, **60** (2) (2004), 91–110.

[33] Wang, Y.; Shi, T.; Yun, P.; Tai, L.; Liu, M.: PointSeg: real-time semantic segmentation based on 3D lidar point cloud, *CoRR*, 2018.

[34] Wu, B.; Wan, A.; Yue, X.; Keutzer, K.: SqueezeSeg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D lidar point cloud, in *IEEE Int. Conf. on Robotics and Automation*, 2018, 1887–1893.

[35] Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T.: Multi-view 3D object detection network for autonomous driving, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, 1907–1915.

[36] Varga, R.; Costea, A.; Florea, H.; Giosan, I.; Nedevschi, S.: Super-sensor for 360-degree environment perception: point cloud segmentation using image features, in *IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, 2017, 1–8.

[37] Wu, B.; Zhou, X.; Zhao, S.; Yue, X.; Keutzer, K.: Squeezesegv2: improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. *arXiv preprint arXiv:1809.08495*, 2018.

[38] Fischler, M.A.; Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of The ACM, 1981, 381–395.

[39] Landrieu, L.; Obozinski, G.: Cut pursuit: fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM. J. Imaging Sci.*, **10** (4) (2017), 1724–1766.

[40] Wu, J., *et al.*: Ground-aware point cloud semantic segmentation for autonomous driving, in *Proc. of the 27th ACM Int. Conf. on Multimedia*, 2019, 971–979.

[41] Gross, S., *et al.*: Automatic differentiation in pytorch, in *Proc. of the Int. Conf. on Neural Information Processing Systems Workshop*, 2017.

[42] Kingma, D.P.; Ba Adam, J.: A method for stochastic optimization, in *Int. Conf. on Learning Representations (ICLR)*, 2014.

[43] Li, B.: 3D fully convolutional network for vehicle detection in point cloud, in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017, 1513–1518.

[44] Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R.: 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)*, 2016, 1–9.

[45] Lin, T.-Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J.: Feature pyramid networks for object detection, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, 2117–2125.

Jian Wu is currently a Master student in the Department of Electronic Engineering and Information Science at University of Science & Technology of China. His research interests reside in computer vision and graphics.

Qingxiong Yang received the BE degree in Electronic Engineering & Information Science from University of Science & Technology of China in 2004 and the PhD degree in Electrical & Computer Engineering from University of Illinois at Urbana-Champaign in 2010. He is an assistant Professor in the Computer Science Department at City University of Hong Kong. His research interests reside in computer vision and computer graphics. He is a recipient of the best student paper award at MMSP 2010 and the best demo award at CVPR 2007.

https://doi.org/10.1017/ATSIP.2020.21 Published online by Cambridge University Press