Efficient Outdoor 3D Point Cloud Semantic Segmentation for Critical Road Objects and Distributed Contexts

Chi-Chong Wong[1][0000−0002−5562−967X], Chi-Man Vong[1][0000−0001−7997−8279]

University of Macau[1]
amilton.wong@connect.um.edu.mo, cmvong@um.edu.mo

Abstract. Large-scale point cloud semantic understanding is an important problem in self-driving cars and autonomous robotics navigation. However, such problem involves many challenges, such as i) critical road objects (e.g., pedestrians, barriers) with diverse and varying input shapes; ii) distributed contextual information across large spatial range; iii) efficient inference time. Failing to deal with such challenges may weaken the mission-critical performance of self-driving car, e.g, LiDAR road objects perception. In this work, we propose a novel neural network model called Attention-based Dynamic Convolution Network with Self-Attention Global Contexts(ADConvnet-SAGC), which i) applies attention mechanism to adaptively focus on the most task-related neighboring points for learning the point features of 3D objects, especially for small objects with diverse shapes; ii) applies self-attention module for efficiently capturing long-range distributed contexts from the input; iii) a more reasonable and compact architecture for efficient inference. Extensive experiments on point cloud semantic segmentation validate the effectiveness of the proposed ADConvnet-SAGC model and show significant improvements over state-of-the-art methods.

Keywords: 3D semantic segmentation, attention, point convolution, point clouds,

1 Introduction

Point cloud based semantic segmentation is a task to classify the labeling of each 3D point of input point cloud. It is an essential task for many applications, such as service-robots autonomous navigation in indoor scenario, self-driving vehicles in outdoor environment. Specifically, large-scale outdoor LiDAR dataset NPM3D[17] covers kilometers scale in range across multiple cities. Moreover, road objects such as pedestrians, barriers, bollards in NPM3D dataset always exhibit as small-sized 3D objects and diverse shapes. Failing to accurately understanding those critical road objects will cause unexpected damages and even deaths. Thus, tackling large-scale outdoor point cloud semantic segmentation involves many challenges, such as i) critical road objects (e.g., pedestrians, barriers and bollards), as they are always exhibited as small objects with diverse and varying shapes; ii) the contextual information of large-scale scenes are always distributed across long spatial range; iii) the demand of efficient inference operation. Failing to deal with such challenges will significantly lower the performance of 3D scene understanding, e.g. LiDAR-based semantic segmentation on road objects from LiDAR input, which is vital for self-driving cars.

Deep convolutional neural network (CNN) [12, 11] based approach has shown its superior performance in many image-based vision tasks such as 2D semantic segmentation [4, 21] and thus becomes a mainstream method. In retrospect to the recent studies on deep learning based methods for point cloud analysis [3, 16, 9], it is found that such methods cannot provide segmentation results with sufficient accuracy, especially for small objects with diverse shapes or large-scale scenes with distributed contexts. PointNet[3] extracts features individually for each point, but without considering neighboring point information. Its variant, PointNet++[16] applies PointNet model to hierarchically extract local information at small group of each sampled point, while similar work PointCNN[9] extracts local features by convolution operation with an additional learnable module for canonical transformation. However, all of such methods have not considered how to adapt the feature extraction operation for diverse input shapes and how to provide an effective way to capture the distributed contexts across large spatial range.

There are few studies on dynamic adaptions on the diverse shape of irregular point clouds for convolution operation. Similar work such as DeformConv [4] is only applicable to 2D image domain. In order to perform accurate segmentation on small critical 3D objects on the road (e.g., pedestrians, barriers, bollards) which exhibits diverse and varying shapes, we argue that it is essential to adapt convolution operation on the shapes of 3D objects to better extract the shape features, such a way is more natural and avoids the loss of information which comes from inaccurate transformation in PointNet[3] and PointCNN[9] methods.

A recent work, GACNet[22] tries to mitigate these issues by adding the attention mechanism to graph convolution for processing the point cloud. However, we argue that such approach does not provide an effective and adaptive manner in aggregating features, since the potentially valuable information of neighboring points will be lost if they are not selected as the vertices of the graph in initial scale. Also, GACNet[22] lacks of a mechanism in capturing global contextual information for large-scale point cloud.

Taking such motivation, a novel convolution operation called *attention-based dynamic point convolution* (ADConv) is proposed to specifically deal with 3D point clouds with diverse and varying shapes. Instead of learning a transformation to align the input point clouds to have canonical pose, attentional weights are injected into neighboring points within the local area of each input point. Through the learned attentional weights representing the similarities of spatial positions and semantic features between neighboring points and input point, the convolution operations can dynamically focus on the most task-related portions of the input, and ignore the unrelated parts. Such attentional weighting scheme is equivalent to dynamically deform the receptive field of convolution to the shapes of 3D objects. Unlike GACNet[22], ADConv applies sampling and grouping operations in each stage to ensure all neighboring points to have the opportunities to be assigned attention weights, which acts as a more effective approach for neighboring points features extraction.

To tackle the second issue in point cloud semantic segmentation, such as difficulty in capturing distributed contexts in large-scale scene, a novel *self-attention global context* (SAGC) module is proposed to be integrated with ADConv for capturing distributed contexts, without the need of stacking many layers of convolution operations for enlarging its receptive field. As a result, a more reasonable and compact architecture called ADConvnet-SAGC is proposed for efficient semantic segmentation for large-scale 3D point cloud.

In summary, the main contributions of the work are highlighted as follows:

1. With the novel attention-based dynamic point convolution (ADConv), the convolutional operation can be dynamically adapted to the diverse shapes of 3D objects, especially for critical road objects.
2. A novel self-attention global context (SAGC) module is proposed to efficiently capture the distributed contextual information globally to further improve the accuracy of 3D semantic segmentation.
3. Instead of stacking many layers of convolution operations, a more reasonable and compact architecture is proposed for efficient semantic segmentation.

From these contributions, an accurate and efficient 3D semantic segmentation model for large-scale scene called ADConvnet-SAGC is presented. Extensive experiments are evaluated on challenging benchmarks, which demonstrate that the proposed model achieves state-of-the-arts results.

2 Related Works

In this section, the two main related techniques: deep learning methods on point clouds and attention mechanism are discussed.

2.1 Deep Learning on Point Clouds

With the unprecedented success of convolutional neural network (CNN) in 2D image recognition problem [6, 5, 18], there have been exploratory works to adapt its hierarchical feature learning capability for 3D point cloud input. These works can be mainly categorized as voxel based [12, 23], multiple-views based [19, 15] and point based methods [3, 16, 9, 24]. The point based approach will be mainly discussed.

Point based approach Several developments in feature learning directly from 3D point cloud [3, 16] have been proposed in recent years. PointNet [3] uses a shared multiple layers perceptron (MLP) on each 3D point individually to learn spatial encoding, then max pooling is applied to obtain a global features of the

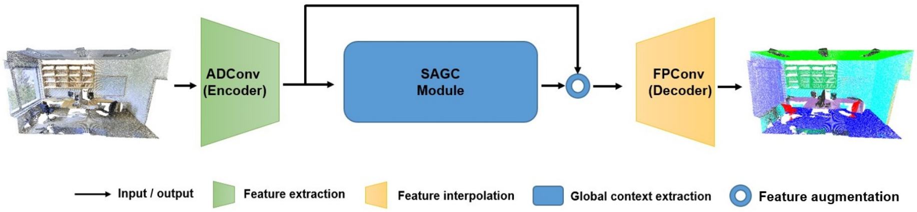

Fig. 1. Illustration on the architecture of ADConvnet-SGAC.

point cloud input. However, such method ignores the local spatial relationships among neighboring points, which leads to limited performance. Further extension work PointNet++[16] is proposed to apply PointNet model on a small neighboring sphere centered at each point for constructing hierarchical features, but its unnecessary high computational complexity limits its performance. PointCNN [9] applies the learned χ-transformation to weight input features. However, the learned χ-transformation does not afford to represent objects with complex 3D shapes and the shape structure of objects cannot be well captured for feature learning. Some recent works [14, 28] also try to improve the feature representations by learning both instance and semantic segmentation tasks simultaneously.

2.2 Attention Mechanism

As attention mechanism can deal with inputs with varying sizes, and focus on the most related parts of input with respect to the task, it has been considered as an effective way in dealing with sequence tasks, such as machine translation[2], image captioning[26], language modeling [21] tasks. Recent works [22, 25] propose to add weights for adapting the convolution operations, but we argue that the weighting schemes in [22, 25] are not convincingly effective. Motivated by these, a novel ADConv is proposed to adopts attention mechanism to extract local features in effective manner.

Global contexts are essential for large-scale point cloud processing. [27] applies RNN to capture contexts but it suffers from computational inefficiency. Self attention[21] has been shown as a powerful approach in capturing long-range dependency of input. In this work, the SAGC module is appended for further capturing the global contexts of point clouds.

3 Methods

In this section, the details of the proposed network model: Attention-based Dynamic Point Convolutional Neural Network with Self-Attention Global Context (ADConvnet-SAGC) for 3D point cloud semantic segmentation is presented. The ADConvnet-SAGC consists of three core modules: 1) Attention-based Dynamic Point Convolution (ADConv) module for dynamically adapting the convolution operation for irregular point cloud input via attention mechanism, 2)

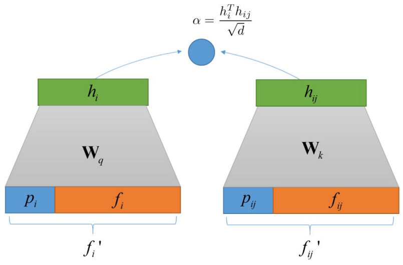

Fig. 2. Illustration on the computation of attention score function α.

Self-Attention Global Context (SAGC) module for capturing long range contextual information globally by self-attention mechanism, 3) Feature Propagation Convolution (FPConv) module for feature interpolation. Fig. 1 illustrates the architecture of the entire network. The descriptions of each proposed module are detailed as follows.

3.1 Attention-based Dynamic Point Convolution (ADConv)

An input point cloud is presented as a set of points $P \in \mathbb{R}^{N \times 3}$ and its corresponding set of features $F \in \mathbb{R}^{N \times D}$, where N and D are the number of points and the dimension of input features, respectively. The xyz position and feature of each point P_i is denoted as $p_i \in \mathbb{R}^3$ and $f_i \in \mathbb{R}^D$, respectively. The set of neighboring points of p_i is denoted as $\mathcal{N}_i$. The proposed ADConv operation is designed to learn a mapping function $q : \mathbb{R}^D \rightarrow \mathbb{R}^C$, which maps the features of neighboring points to the new aggregated features of input point P_i, while keeping the advantageous properties of convolutional operation[8], such as i) locality through local receptive field, ii) translation invariance through weights sharing and iii) hierarchical compositionality on extracted features. More importantly, it is required to handle the properties of irregular point cloud, such as the permutation invariance and shape-varying neighborhoods.

For ADConv operation, attention weights injected into neighboring points are used to represent the similarities of spatial positions and semantic features for neighboring points. Through end-to-end training, the attention weights can be adjusted to fit the segmentation task by following general backpropagation. As a result, ADConv can selectively focus on the most task-related parts with sufficient discriminative semantics. Equivalently speaking, convolutional kernel in ADConv can be dynamically adapted to the varying structure of input point cloud.

The pipeline of ADConv is detailed as follows: First, instead of using time-consuming sampling method, such as farthest point sampling (FPS)[16], random sampling (RS) is adopted to sample N input points into M sub-sampled points. The value for M is selected at $N/4$ in general. For each sub-sampled input point $P_i \in \{P_1, P_2, ...P_M\}$, which contains xyz position $p_i \in \mathbb{R}^3$ and features $f_i \in \mathbb{R}^D$,

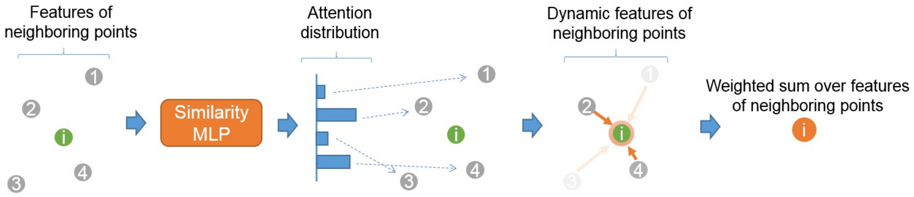
Features of neighboring points
1
2
i
3
4
Similarity MLP
Attention distribution
Dynamic features of neighboring points
Weighted sum over features of neighboring points
i

Fig. 3. The pipeline of attention weighting scheme in the ADConv layer.

its K neighboring points are randomly selected as $P_{i,j}$ $(j = 1, ..., K)$ from its spherical neighborhood.

According to the definition of attention: *Given a set of key-values, and a query, attention is a technique to compute a weighted sum of the values, dependent on the similarity between query and a set of keys.* In our case, query refers to input point P_i: $(p_i,\ f_i)$, and keys refer to neighboring points P_{ij}: $(p_{ij},\ f_{ij})$. To measure the similarity between the query and a set of keys, an attention score function α_{ij} is used. For point cloud, the position proximity and features similarity can naturally reflect such similarity between points. Thus, as illustrated in Fig. 2, position p is first concatenated with feature f as compound features f', which gives f'_i and f'_{ij}. Then, to obtain features with semantics enhanced information for better comparison, a multilayer perceptron (MLP) with learnable weights W_q and W_k are applied onto f'_i and f'_{ij} to get enhanced features $h_i \in \mathbb{R}^C$ and $h_{ij} \in \mathbb{R}^C$, respectively. C is the dimension number of enhanced features. Finally, the attention score function α_{ij} is computed as the inner product between h_i and h_{ij} to obtain similarity measurement between P_i and P_{ij}:

$$\alpha_{ij} = \frac{h_i^T h_{ij}}{\sqrt{C}} \tag{1}$$

$\sqrt{C}$ is used to normalize the effect of feature dimension.

To obtain attention distribution a_{ij}, a softmax function is applied to normalize α_{ij} across all neighboring points

$$a_{ij} = \frac{\exp(\alpha_{ij})}{\sum_{j \in \mathcal{N}_i} \exp(\alpha_{ij})} \tag{2}$$

Finally, the attention distribution a_{ij} is applied on the enhanced features of neighboring points h_{ij} to obtain dynamic features h'_i for input point P_i as illustrated in Fig. 3:

$$h'_i = \sum_{j \in \mathcal{N}_i} a_{ij} h_{ij} \in \mathbb{R}^C \tag{3}$$

We can see that the attention output h'_i is the weighted sum of related parts of neighboring points. As a result, the context information around the input

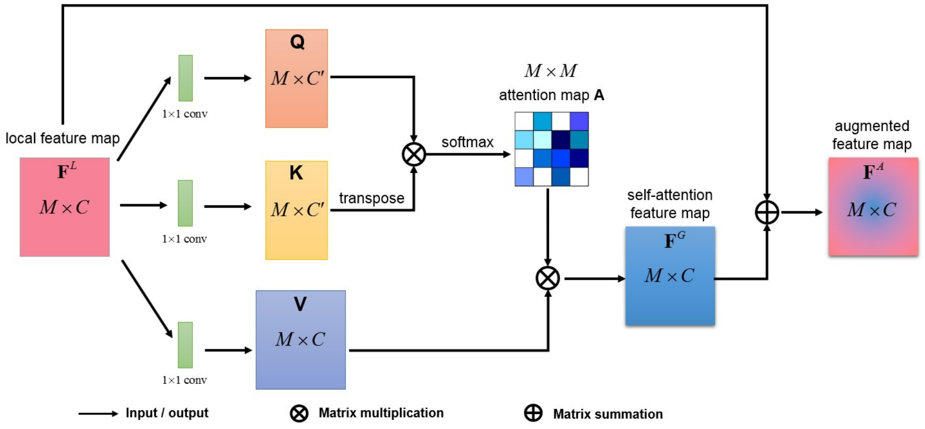

Fig. 4. The pipeline of SAGC module.

point P_i is captured by such attention mechanism. The overall operation on the attention weighting scheme is illustrated in Fig. 3. It is shown that the most related part of neighboring points, P_2 and P_4 are maintained by attention weight, while the other points P_1 and P_3 are eliminated.

Finally, as the same as general convolution operation, h'_i is passed into a non-linear activation such as ReLU [13] to obtain high-level features $f_i^{out} = \text{ReLU}(h'_i) \in \mathbb{R}^C$, as the output of ADConv layer. And all of such points features constitute a feature map $F^{out} \in \mathbb{R}^{M \times C}$, where M is the number of sub-sampled points and can be regarded as the spatial size of feature map.

3.2 Self-Attention Global Context (SAGC) Module

To capture global contextual information from the input, traditional convolution based methods require many convolutional layers to enlarge the receptive field. Such way may consume large volume of memory and longer inference time. In this work, we adopt self-attention mechanism [21] to avoid these issues.

The goal of self-attention is to globally extract the long range contexts from the entire feature map F^{out}, by means of extracting the self-similarity for any pair of elements in feature map F^{out}. As illustrated in Fig. 4, given the feature map $F^L = F^{out} \in \mathbb{R}^{M \times C}$, the SAGC module applies three different 1 × 1 conv filters on F^L to obtain three feature maps: query feature map $Q \in \mathbb{R}^{M \times C'}$, key feature map $K \in \mathbb{R}^{M \times C'}$, value feature map $V \in \mathbb{R}^{M \times C}$. The query feature map Q and key feature map K take the role in measuring the pairwise similarity for each pair of element in F_L. For each spatial position u, a feature vector $Q_u \in \mathbb{R}^{C'}$ of feature map Q is selected for affinity comparison, where $u = [1, ..., M]$. Also, $K_v \in \mathbb{R}^{C'}$ is the feature vector of feature map K for each lookup position $v = [1, ..., M]$. Then, affinity operation is applied between Q_u and K_v to obtain the relevance score $r_{u,v}$,

$$r_{u,v} = Q_u^T K_v \tag{4}$$

where $r_{u,v} \in \mathbf{R}$ represents the degree of relevance between feature vector Q_u and K_v and matrix $\mathbf{R} \in \mathbb{R}^{M \times M}$ is the relevance score map. Then, softmax operation is applied on each element $r_{u,v} \in \mathbf{R}$ to obtain attention weight $a_{u,v}$

$$a_{u,v} = \frac{\exp(r_{u,v})}{\sum_{v=1}^{M} \exp(r_{u,v})} \tag{5}$$

which constitutes the element of attention map $\mathbf{A} \in \mathbb{R}^{M \times M}$ to carry the normalized self-relevance degree existing in the entire feature map F^L.

The value feature map V is a transformation from the feature map F^L, the global contextual information in it will then be extracted. The attention weights $a_{u,v}$ reflect the affinities across all input features. The features with more relatedness, even located non-locally, will be assigned sufficient attention weights to be selected for subsequent feature augmentation. Thus, attention weight $a_{u,v}$ is applied to all feature vectors in value feature map V to obtain the global feature vector F_u^G:

$$F_u^G = \sum_{v=1}^{M} a_{u,v} V_v \tag{6}$$

where $F_u^G \in \mathbb{R}^C$ denotes a feature vector of self-attention feature map $F^G \in \mathbb{R}^{M \times C}$ at spatial position u, which captures the long-range global contexts. Then, it is added to the original local feature map F^L to obtain the augmented feature map $F^A = F^G + F^L$. As a result, long-range contextual information is adaptively aggregated to augment the point-wise representation for better performing the semantic segmentation task.

As the feature maps are downsampled after processed by each ADConv layer, the augmented feature map F^A is subsequently interpolated by Feature Propagation (FPConv) module [16] for recovering the point features as the original scale of input point cloud, which is the final point-wise predictions. The whole proposed network model ADConvnet-SAGC can be trained in an end-to-end manner.

3.3 Network Architectures

Following the network design of Pointnet++[16], ADConvnet-SAGC adopts encoder decoder style. The encoder module contains four ADConv layers, which acts as local features extractor. Then, the extracted local features is fed into SAGC module to obtain augmented features with global contextual information. As the semantic segmentation task requires to upsample the feature maps downsampled by the encoder, the decoder module(four FPConv layers) is employed for progressively upsampling the features.

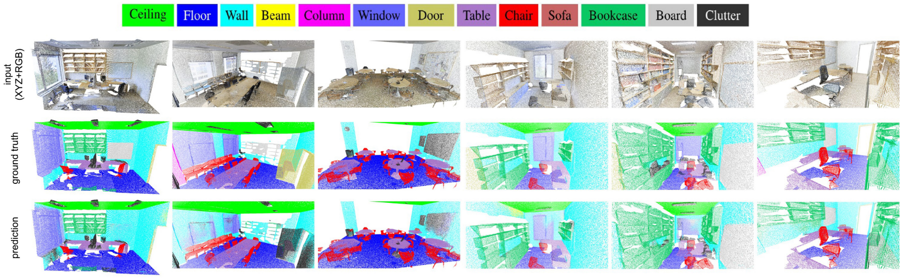

Fig. 5. Visualizations on qualitative results for S3DIS validation set. From left to right are XYZ-RGB input, ground truth labels annotations and prediction of ADConvnet-SGAC model.

4 Experiments

Experiments have been conducted to verify the effectiveness of the proposed ADConvnet-SAGC model in 3D point cloud semantic segmentation task. The evaluations are performed on both indoor and outdoor large-scale benchmarks, such as the Stanford Large-Scale 3D Indoor Spaces (S3DIS) [1] dataset and the Nuage de Points et Modélisation 3D (NPM3D)[17] dataset. For performance measurement, the point-wise overall accuracy (OA), mean intersection over union (mIoU) over all classes and per-class intersection over union (cIoU) are adopted as evaluation metrics. After that, the ablation analysis on each key components of ADConvnet-SAGC model is provided.

4.1 Training Details

The Adam optimizer with default hyper-parameters is applied for optimizing the cross entropy loss as the training loss. The initial learning rate is set to 0.01 and decreased by 10% after each epoch. To train the proposed ADConvnet-SAGC model, $N = 4096$ points are sampled from the input point cloud as the input. The number of neighboring points K is selected as 64. All settings are consistent for both S3DIS[1] and NPM3D[17] dataset. All experiments and runtime analysis were conducted using Nvidia 1080Ti GPU and an i7-5820K CPU clocked at 3.30 GHz. The implementation is built on Pytorch framework.

4.2 Indoor point cloud segmentation on S3DIS dataset

The Stanford 3D semantic parsing dataset (S3DIS) [1] is a large-scale benchmark for indoor 3D semantic perception task in service-robots autonomous navigation or indoor scene understanding. The S3DIS dataset covers six large-scale indoor areas within three different buildings, which contains 272 different rooms in various types (e.g. office, conference room, hall, lobby., etc.) and sums up to 215 million points in total. The indoor scene is captured by a high-quality Matterport scanner which generates dense 3D point clouds with RGB color and

Table 1. Quantitative comparison of *ADConvnet-SAGC* on S3DIS dataset

Method	OA	mIoU	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
PointNet [3]	78.5	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
PointNet++ [16]	80.9	53.2	90.2	91.7	73.1	42.7	21.2	49.7	42.3	62.7	59.0	19.6	45.8	38.2	45.6
DGCNN [24]	83.3	56.3	92.9	93.8	73.1	42.5	25.9	47.6	59.2	60.4	66.7	24.8	57.0	36.7	51.6
GACNet [22]	83.0	56.2	90.8	92.4	75.9	40.3	19.3	47.6	52.8	66.4	70.2	28.9	55.2	40.5	50.6
ADConvnet-SAGC	**87.5**	**60.1**	**93.3**	**95.4**	**78.3**	**43.7**	**27.6**	**50.3**	**68.1**	**69.2**	**71.2**	**30.6**	**57.6**	**41.0**	**54.6**

the corresponding semantic label for each point is densely annotated. Each 3D point is assigned with one of thirteen class labels (e.g. ceiling, floor, wall, table, chair., etc). For quantitative evaluation, we follow the general setting to select Area-1,2,3,4,6 as the training set, and Area-5 as the validation set for accuracy evaluation.

During the pre-process step, the dataset is first split into 1.0m by 1.0m blocks for each room. Then, each block is uniformly sampled into 4,096 points. The batch size for training is selected as 24.

Quantitative and Qualitative Results: The quantitative results of the experimental results are provided in Table 1. It is shown that the proposed ADConvnet-SAGC provides better results against other competitive methods. Particularly, ADConvnet-SAGC achieves significant gains in objects with varying shapes, such as table, chair, sofa, which is shown as consistent results that the attention mechanism in ADConv takes effects on capturing the irregular shapes of such objects. In terms of mean IoU, ADConvnet-SAGC obtains a 13% relative improvement over Pointnet++ [16] method, and achieves an relative improvement of 6.8% when comparing with DGCNN [24] method and GACNet[22] method. As GACNet[22] didn't open source the code at the period of this work, we directly re-implement it. To give fair comparison, we adopt the same training setting as ADConv-SAGC.

The qualitative segmentation results on validation set of S3DIS dataset are also illustrated in Fig. 5, in which the prediction from ADConvnet model is greatly consistent with the ground truth labels annotations, even though the S3DIS dataset exhibits a lot of intra-class varying shapes.

4.3 Outdoor point cloud segmentation on NPM3D dataset

The proposed ADConvnet-SAGC model is also evaluated on NPM3D dataset [17], which is a large-scale benchmark for outdoor 3D semantic perception task in self-driving vehicles or traffic scene understanding. The NPM3D dataset consists of aggregated scans with up to 160 million point-wise annotated 3D points from mobile laser scanner Velodyne HDL-32e mounted on vehicle. It covers streets in more than 2km range across four different cities, Paris, Lille, Ajaccio and Dijon. Following the official evaluation protocol, 9 classes (ground, buildings, poles, bollards, trash cans, barriers, pedestrians, cars, natural) are selected for evaluation. The dataset gives challenging real-word scenario such as varying point cloud densities and complex scene surroundings.

Table 2. Quantitative comparison of *ADConvnet-SAGC* on NPM3D dataset

Method	mIoU	Ground	Building	Pole	Bollard	Trash can	Barrier	Pedestrian	Car	Natural
Validation set: Lille2 sequence										
PointNet++ [16]	43.0	97.6	90.2	28.8	12.2	10.6	9.5	19.9	59.8	58.2
DGCNN [24]	53.5	97.9	93.2	49.9	37.5	17.2	18.3	19.2	88.9	59.2
PointSIFT [7]	63.4	98.3	95.6	51.5	45.8	54.3	33.9	31.6	88.2	71.5
MS-PCNN [11]	70.5	98.1	95.4	57.6	64.6	63.0	34.1	57.7	95.2	68.3
GACNet [22]	69.3	98.0	94.2	61.3	64.2	59.3	35.2	51.1	89.9	70.9
ADConvnet-SAGC	**80.4**	**98.6**	**96.2**	**67.4**	**75.1**	**68.2**	**54.4**	**80.7**	**95.2**	**88.1**
Test set: Ajaccio and Dijon sequence										
RF-MSSF [20]	56.3	99.3	88.6	47.8	67.3	2.3	27.1	20.6	74.8	78.8
MS3-DVS [17]	66.9	99.0	94.8	52.4	38.1	36.0	49.3	52.6	91.3	88.6
HDGCN [10]	68.3	99.4	93.0	67.7	75.75	25.7	44.7	37.1	81.9	90.0
ADConvnet-SAGC	**80.2**	**99.5**	**96.1**	**69.2**	**76.3**	**53.6**	**56.5**	**83.7**	**94.6**	**92.2**

Quantitative and Qualitative Results: In Table 2, the results of proposed model are shown with comparisons with several state-of-the-art methods on NPM3D dataset. The comparison result shows obvious superiority of the proposed ADConvnet-SAGC model over the state-of-the-art methods such as PointNet++[16], DGCNN[24], PointSIFT[7], MS-PCNN[11], MS3-DVS[17], HDGCN[10]. Following the publicly available benchmark results, the per-class IoU and the mean IoU score for valiation set (Lille2 sequence) and test set (Ajaccio, Dijon sequence) are reported. In terms of the mean IoU, ADConvnet-SAGC outperforms MS-PCNN by more than 14%, and achieves an accuracy improvement of 17.4% when comparing with HDGCN method. It is observed that objects across large spatial range such as ground, building and natural can be segmented in high accuracy. In addition to this result, small critical objects with diverse shapes such as barrier, pedestrian, the proposed model provides significant accuracy gains compared with previous state-of-the-art methods.

The qualitative segmentation results on test set (Ajaccio, Dijon sequence) of NPM3D dataset are shown in Fig. 6. The left column shows the original XYZ input, the second and the third column are the global view and detailed view on prediction of our model, respectively. The prediction shows fine and coherent segmentation output on scene objects, even on challenging ones, e.g. small objects trash can and bollard, objects with varying shapes within large spatial area like building and natural objects.

4.4 Effectiveness Analysis of Attention Mechanisms

In Fig. 7, an attention map learned from end-to-end training is illustrated. The reference point is highlighted in green color. For better visualization, the neighboring points assigned with enough attentional weights are highlighted in red color. It is observed that those *attended* parts adapt to the shape of the corner of chair and correspond to the most task-related parts with sufficient discriminative semantics. Through self-attention, global contexts distributed distantly are also finely captured, as illustrated in the red regions located at the other three corners (Fig. 7(a)) or three legs (Fig. 7(b)) of the chair.

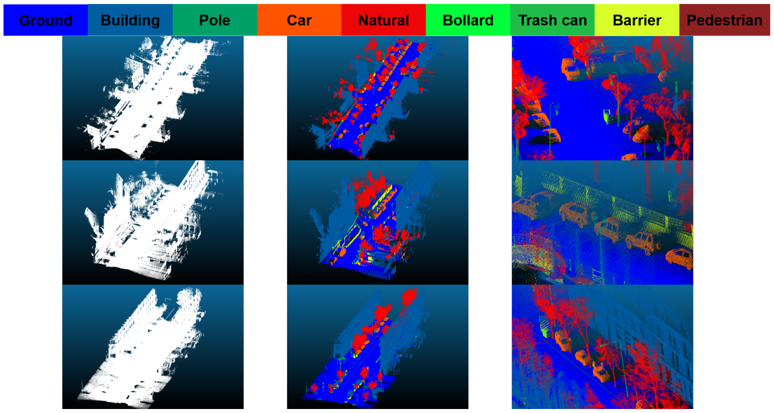

Fig. 6. Qualitative results on NPM3D test set. From left to right are original XYZ input, global view and detailed view on prediction of our model.

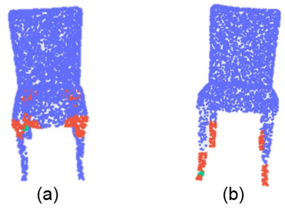

Fig. 7. Visualization of the attention map for example object in ModelNet40 dataset. The green point is the reference point while the regions in red are points assigned with sufficient attentional weights. Two cases of different reference points are illustrated: (a) the corner and (b) the leg of a chair.

Here, we also perform a sensitivity analysis on the value of K. The comparison result for S3DIS dataset is listed in Table 3. It is observed that the option on smaller value ($K = 16$) gives lower score, since too few neighboring points cannot capture sufficient semantic patterns. The options for $K = 64$, $K = 128$ give almost the same accuracy. We argue that the attention approach provide a mechanism to focus on the most task-related parts, the redundant parts in the case of the larger value of K will be ignored since they are assigned with low attention weights. Therefore, our model choose the option $K = 64$ to balance the trade-off between accuracy and memory costs.

4.5 Efficiency of ADConvnet-SAGC

To evaluate the efficiency of ADConvnet-SAGC, the inference time for input point cloud is listed in Table 4. The evaluation is performed on Lille2 sequence of NPM3D dataset and the resulting inference time consumption and mIoU are listed for comparison. In Table 4, the proposed ADConvnet-SAGC con-

Table 3. The effects of different options of K value. The evaluation metric is the mIoU score on Area 5 sequence of S3DIS dataset

K=16	K=64	K=128
59.0	60.5	60.6

sumes 318.0 seconds for inferencing the whole Lille2 sequence (around 30 million points), which consumes up to 2.3 times less inference time than PointNet++ and DGCNN. The percentage of improvement against PointNet++ method is also listed, which shows significant the reduction in inference time of ADConvnet-SAGC.

Table 4. Inference Time Comparison

	NPM3D (Lille2 sequence)	
Method	Inference time (second)	mIoU
PointNet++	756.2 (0%)	43.0
DGCNN	630.2 (120%)	53.5
ADConvnet-SAGC	**318.0 (238%)**	**80.4**

4.6 Ablation analysis

To validate the effectiveness of the proposed ADConvnet-SAGC model, several ablation analyses are conducted on both S3DIS and NPM3D dataset. The ablation result is listed in Table 5.

(1). Replacing ADConv with Set Abstraction layer. The ADConv operation enables the model to adaptively focus on the most related neighboring features for better features aggregation. For comparison, Set Abstraction layer in PointNet++ model tends to hard aggregate the feature in irregular shape. As a result, it provides lower performance.

(2). Removing SAGC module. The SAGC module provides the long range contextual information to the model for augmenting the feature representation. By removing this module, the performance is greatly decreased due to the loss of long range contexts.

(3). Replacing the learned attention weights with constant attention weights. The way in using constant attention weights is equivalent to conventional mean among neighboring points. By doing so, features similarity and shape structure of neighboring points are ignored. Thus, the performance is greatly decreased.

The mIoU scores of all ablated variants are compared in Table 5. We can conclude that: i) The most impact comes from the removal of the SAGC module, since the long range contextual information is essential in large-scale point cloud and large-sized objects become the majority parts. ii) The role of ADConv shows the next important factor in performance, especially for point cloud with diverse and varying 3D shape. iii) The learned attention weights are essential. From this ablation analysis, it is shown that how each proposed module which constitutes the full ADConvnet-SAGC model obtains the state-of-the-art accuracy.

Table 5. The mIoU scores of all ablated variants based on the full ADConvnet-SAGC. Both the Area-5 sequence of S3DIS and the Lille2 sequence of NPM3D are selected for evaluation

	mIoU	
	(S3DIS)	(NPM3D)
(1). Replace ADConv with Set Abstraction layer	57.4	77.5
(2). Remove SAGC module	56.7	76.3
(3). (1)+(2)	52.8	72.2
(4). Replace learned attention weights with constant attention weights	53.0	71.6
(5). The full ADConvnet-SAGC model	**60.1**	**80.4**

5 Conclusion

In this paper, a novel point cloud based neural network model called ADConvnet-SAGC is proposed which integrates attention mechanism to tackle several challenges in large-scale point cloud semantic segmentation problem. The proposed ADConvnet-SAGC integrates attention mechanism into point convolution to handle input with diverse and varying shapes. It also applies self-attention technique to efficiently capture long-range contextual information for enhancing feature representations. With these two improvements, the segmentation results for road objects (especially critical small objects such as pedestrians, cars, barriers, bollards) have significant gain in accuracy. The performance of ADConvnet-SAGC is validated in terms of accuracy and efficiency over challenging benchmarks. From the experiments, ADConvnet-SAGC outperforms several state-of-the-art point cloud based semantic segmentation methods. The experimental results also validate the contributions of ADConvnet-SAGC: i) an attention-based adaptive module which can be easily integrated to obtain dynamic point convolution operation; ii) a more accurate point cloud based semantic segmentation with global spatial consistency; iii) much lower computational cost (e.g., about 2.3 times faster than PointNet++) compared with other state-of-the-art point cloud based approaches.

Acknowledgement

This project was partially funded by The University of Macau (MYRG2019-00016-FST), and The Science and Technology Development Fund, Macau S.A.R. (File no. 0004/2019/AFJ).

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2016)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translatea. In: Proceedings of the IEEE International Conference on Learning Representations (2015)
3. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 77–85 (July 2017). https://doi.org/10.1109/CVPR.2017.16
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(1), 142–158 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
7. Jiang, M., Wu, Y., Zhao, T., Zhao, Z., Lu, C.: Pointsift: A sift-like network module for 3d point cloud semantic segmentation. arXiv preprint arXiv:1807.00652 (2018)
8. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
9. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. In: Advances in Neural Information Processing Systems. pp. 820–830 (2018)
10. Liang, Z., Yang, M., Deng, L., Wang, C., Wang, B.: Hierarchical depthwise graph convolutional neural network for 3d semantic segmentation of point clouds. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8152–8158 (May 2019). https://doi.org/10.1109/ICRA.2019.8794052
11. Ma, L., Li, Y., Li, J., Tan, W., Yu, Y., Chapman, M.A.: Multi-scale point-wise convolutional neural networks for 3d object segmentation from lidar point clouds in large-scale environments. IEEE Transactions on Intelligent Transportation Systems (2019)
12. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 922–928. IEEE (2015)
13. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
14. Pham, Q.H., Nguyen, T., Hua, B.S., Roig, G., Yeung, S.K.: Jsis3d: joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8827–8836 (2019)
15. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5648–5656 (2016)

16. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems. pp. 5099–5108 (2017)
17. Roynard, X., Deschaud, J.E., Goulette, F.: Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. The International Journal of Robotics Research **37**(6), 545–557 (2018). https://doi.org/10.1177/0278364918767506
18. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Tansactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2017)
19. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 945–953 (2015)
20. Thomas, H., Goulette, F., Deschaud, J.E., Marcotegui, B.: Semantic classification of 3d point clouds with multiscale spherical neighborhoods. In: 2018 International Conference on 3D Vision (3DV). pp. 390–398. IEEE (2018)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
22. Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J.: Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10296–10305 (2019)
23. Wang, P.S., Liu, Y., Guo, Y.X., Sun, C.Y., Tong, X.: O-cnn: Octree-based convolutional neural networks for 3d shape analysis. ACM Transactions on Graphics (TOG) **36**(4), 72 (2017)
24. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (TOG) (2019)
25. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9621–9630 (2019)
26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
27. Ye, X., Li, J., Huang, H., Du, L., Zhang, X.: 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 403–417 (2018)
28. Zhao, L., Tao, W.: Jsnet: Joint instance and semantic segmentation of 3d point clouds. In: AAAI. pp. 12951–12958 (2020)