GA-NET: Global Attention Network for Point Cloud Semantic Segmentation

Shuang Deng, *Member, IEEE*, and Qiulei Dong

***Abstract*—How to learn long-range dependencies from 3D point clouds is a challenging problem in 3D point cloud analysis. Addressing this problem, we propose a global attention network for point cloud semantic segmentation, named as GA-Net, consisting of a point-independent global attention module and a point-dependent global attention module for obtaining contextual information of 3D point clouds in this paper. The point-independent global attention module simply shares a global attention map for all 3D points. In the point-dependent global attention module, for each point, a novel random cross attention block using only two randomly sampled subsets is exploited to learn the contextual information of all the points. Additionally, we design a novel point-adaptive aggregation block to replace linear skip connection for aggregating more discriminate features. Extensive experimental results on three 3D public datasets demonstrate that our method outperforms state-of-the-art methods in most cases.**

***Index Terms*—3D point cloud, semantic segmentation, global attention, convolutional neural networks, deep learning.**

I. Introduction

THREE-DIMENSIONAL point cloud semantic segmentation is an important topic in the field of computer vision. In recent years, a large amount of Deep Neural Networks(DNNs) [4], [5], [7]–[10], [12], [13], [19], [24], [25] for point cloud semantic segmentation have been proposed. Although these methods can capture the geometric structures of local regions well, the relationships between long-range neighborhoods of 3D point clouds are usually ignored.

In fact, the contextual information from long-range neighboring points is essential for 3D point cloud semantic segmentation. Some recent works [20]–[22] showed that the non-local module in NLNet [11] could improve the performances of DNNs on point cloud segmentation significantly. However, due to the computationally expensive nature of the non-local module, these existing methods could not directly handle the input complete 3D scenes, but they have to split each input complete 3D scene into many small cubes in advance, and then use the non-local module to handle each small cube, resulting in an incomplete contextual feature. Although the non-local module is computationally complex, there are some methods [14], [23] to simplify the non-local module in the field of image processing. However, since 3D point cloud is an unordered and irregular structure, these attention mechanisms cannot be applied to 3D point clouds directly.

In addition, the non-local module [11] is point-dependent, where the calculated attention maps are dependent on different points. GCNet [15] indicates that the point-independent attention method could also improve the capabilities of DNNs, where only one attention map is shared by all points. However, this method ignores the differences between local regions.

Addressing these problems, we propose an end-to-end global-attention-based model, named as Global Attention Network (GA-Net) for point cloud semantic segmentation with a moderate computational complexity. GA-Net consists a U-Net-based feature extractor and two modules called the point-independent global attention module and the point-dependent global attention module. In the point-independent global attention module, we compute an attention map and apply it to all points for obtaining point-independent global information. In the point-dependent global attention module, an efficient random cross attention block is designed to replace the non-local module [11], which is of lower computational complexity and could directly handle complete 3D scenes. Specifically, each point only has connections with two randomly sampled subsets, the number of which is much smaller than the entire point cloud. Besides, we design a novel point-adaptive aggregation block to replace linear skip connection for aggregating more discriminate features. The point-independent global attention module aims to learn an attention map for extracting global but relatively coarse contextual features. Then, based on the features outputted from the point-independent global attention module, the point-dependent global attention module aims to extract more deliberate global contextual features for each 3D point respectively.

In sum, the main contributions of this paper include:

- We propose the point-independent global attention module, which could learn global information from entire point clouds in an efficient way.
- We propose the point-dependent global attention module, which has a lower computational complexity than the existing non-local module [11]. Besides, we propose a novel point-adaptive aggregation block.

Manuscript received March 22, 2021; revised May 10, 2021; accepted May 16, 2021. Date of publication May 24, 2021; date of current version July 2, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants U1805264 and 61991423; in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB32050100, and in part by the Open Research Fund from Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dezhong Peng. *(Corresponding author: Qiulei Dong.)*

Shuang Deng and Qiulei Dong are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shuang.deng@nlpr.ia.ac.cn; qldong@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/LSP.2021.3082851

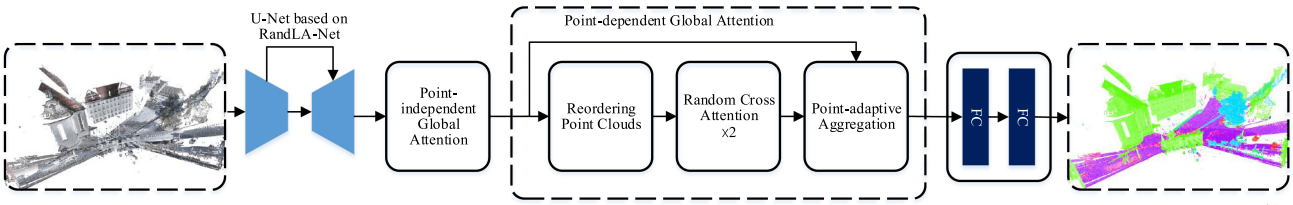

Fig. 1. Architecture of the proposed GA-Net.

- We propose the GA-Net consisting of the point-independent global attention module and the point-dependent global attention module. Extensive experimental results on point cloud semantic segmentation demonstrate that the proposed model surpass state-of-the-art methods in most cases.

II. Global Attention Network

A. Architecture

As shown in Figure 1, our end-to-end GA-Net consists of three modules, including a feature extractor, a point-independent global attention module, and a point-dependent global attention module. The feature extractor is a U-Net based on RandLA-Net [19]. RandLA-Net takes the entire point clouds as input and is able to efficiently infer per-point semantics in a single pass. Besides, the performances of RandLA-Net is somewhat state-of-the-art on several benchmarks.

When a 3D point cloud $\mathbf{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N\} \in \mathbb{R}^{N\times(3+d)}$ is given, where N is the number of points and $3+d$ denotes the xyz-dimension and additional properties (e.g., $d = 3$ for RGB or normal information), we firstly send $\mathbf{P}$ to the feature extractor to construct its high-level representation $\mathbf{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N\} \in \mathbb{R}^{N\times C}$ where C is the dimension of high-level features. Secondly, we send $\mathbf{F}$ to the point-independent global attention module to get feature map $\mathbf{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_N\} \in \mathbb{R}^{N\times C}$. Then, we feed $\mathbf{G}$ into the point-dependent global attention module to generate the context-aware feature map $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \in \mathbb{R}^{N\times C}$. Lastly, the feature map $\mathbf{X}$ is fed into fully-connected(FC) layers for label assignment.

B. Point-Independent Global Attention Module

For aggregating point-independent global information efficiently, we introduce the point-independent global attention module. Firstly, we reduce the number of feature channels to one by sending $\mathbf{F}$ to a shared MLP. Then we perform a normalization using softmax to obtain the global attention map $\boldsymbol{w} \in \mathbb{R}^N$:

$$w_i = \frac{\exp(\mathrm{MLP}(\boldsymbol{f}_i))}{\sum_{j=1}^{N}\exp(\mathrm{MLP}(\boldsymbol{f}_j))} \tag{1}$$

where w_i is the i-th element of $\boldsymbol{w}$. After getting the attention map, we firstly perform attention pooling for all points to obtain the global contextual feature. Then two FC layers are applied to learn channel-wise dependencies. For reducing the difficulty of optimization, we add a layer normalization(LN) between the two FC layers (before ReLU). Finally, the output feature is

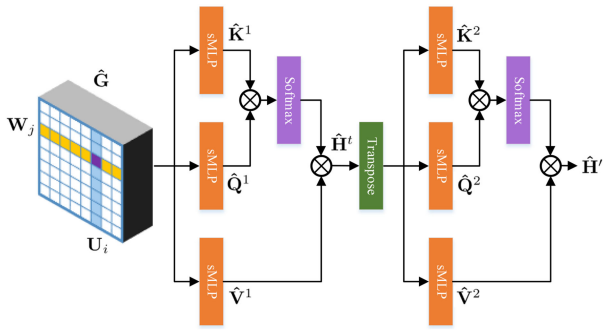

Fig. 2. Architecture of the random cross attention block.

stacked into N copies and connected with $\mathbf{F}$ by the element-wise multiplication to get the feature map $\mathbf{G}$. The whole procedure of the point-independent global attention is formulated as:

$$\boldsymbol{g}_i = \boldsymbol{f}_i \odot \mathrm{FC}\left(\mathrm{ReLU}\left(\mathrm{LN}\left(\mathrm{FC}\left(\sum_{j=1}^{N} w_i \boldsymbol{f}_i\right)\right)\right)\right) \tag{2}$$

where '$\odot$' represents the element-wise multiplication.

Overall, this module can extract not only the point-independent global information efficiently, but also the relationships between different feature channels.

C. Point-Dependent Global Attention Module

To model long-range point-dependent dependencies with a lower computational costs and memory, we introduce a point-dependent global attention module. As shown in Figure 1, we firstly reorder 3D points by random sampling. Then, the feature map $\mathbf{G}$ is sent to two random cross attention blocks to produce $\mathbf{H}' \in \mathbb{R}^{N\times C}$ and $\mathbf{H}'' \in \mathbb{R}^{N\times C}$ respectively. Lastly, the features $\mathbf{H}''$ and $\mathbf{G}$ are sent to a point-adaptive aggregation block for aggregating more discriminate feature map $\mathbf{X}$.

Reordering point clouds. The whole point cloud can be divided into many subsets, where each subset is obtained through uniformly random sampling. Specifically, if the number of points in each subset is set to k_1, the number of a point cloud with N will firstly be expanded to $\hat{N}$, making $\hat{N} = k_1 \times k_2$ ($k_2 \in \mathbb{Z}$), to get the expanded point cloud $\hat{\mathbf{P}}$ where the expanded points are randomly sampled from the original point cloud. Then we sample k_2 times from the point cloud $\hat{\mathbf{P}}$, each of which contains k_1 randomly sampled points from the remaining unsampled points. So the matrix $\hat{\mathbf{P}}$ can be reshaped as $\hat{\mathbf{P}} \in \mathbb{R}^{k_2\times k_1\times(3+d)}$. If transposing the first dimension and the second dimension of this matrix, it can also be explained that the point cloud consists of k_1 subsets with k_2 random-sampled points. As seen in Figure 2,

each point is contained in a subset $\mathbf{U}_i (i = 1, 2, \ldots, k_2)$ of k_1 points and a subset $\mathbf{W}_j (j = 1, 2, \ldots, k_1)$ of k_2 points.

Random cross attention block. After reordering the point cloud, we design a novel random cross attention block to achieve non-local attention in a more efficient way. This block consists of a two-pass procedure which can be seen in Figure 2. According to $\hat{\mathbf{P}}$, the feature map $\mathbf{G}$ can be expanded to $\hat{\mathbf{G}}$, and then reshaped to $\hat{\mathbf{G}} \in \mathbb{R}^{k_2 \times k_1 \times C}$.

In the first step, we feed $\hat{\mathbf{G}}$ into three shared MLPs to obtain the key feature map $\hat{\mathbf{K}}^1 = \{\hat{\mathbf{K}}^1_1, \hat{\mathbf{K}}^1_2, \ldots, \hat{\mathbf{K}}^1_{k_2}\} \in \mathbb{R}^{k_2 \times k_1 \times C}$, the query feature map $\hat{\mathbf{Q}}^1 = \{\hat{\mathbf{Q}}^1_1, \hat{\mathbf{Q}}^1_2, \ldots, \hat{\mathbf{Q}}^1_{k_2}\} \in \mathbb{R}^{k_2 \times k_1 \times C}$ and the value feature map $\hat{\mathbf{V}}^1 = \{\hat{\mathbf{V}}^1_1, \hat{\mathbf{V}}^1_2, \ldots, \hat{\mathbf{V}}^1_{k_2}\} \in \mathbb{R}^{k_2 \times k_1 \times C}$, respectively. The output feature map of the first step is $\hat{\mathbf{H}}^t = \{\hat{\mathbf{H}}^t_1, \hat{\mathbf{H}}^t_2, \ldots, \hat{\mathbf{H}}^t_{k_2}\} \in \mathbb{R}^{k_2 \times k_1 \times C}$. Specifically, for each subset $\mathbf{U}_i$, we use multiplication between $\hat{\mathbf{K}}^1_i$ and the transpose of $\hat{\mathbf{Q}}^1_i$ with a softmax to produce a self-attention map. Then $\hat{\mathbf{H}}^t_i$ is produced by multiplying between the self-attention map and $\hat{\mathbf{V}}^1_i$. The first step can be formulated as:

$$\hat{\mathbf{H}}^t_i = \text{softmax}(\hat{\mathbf{K}}^1_i \hat{\mathbf{Q}}^{1\top}_i)\hat{\mathbf{V}}^1_i \tag{3}$$

In the second step, the feature map $\hat{\mathbf{H}}^t$ is firstly transposed on the first dimension and the second dimension, then similarly sent to three shared MLPs to obtain the key feature map $\hat{\mathbf{K}}^2 = \{\hat{\mathbf{K}}^2_1, \hat{\mathbf{K}}^2_2, \ldots, \hat{\mathbf{K}}^2_{k_1}\} \in \mathbb{R}^{k_1 \times k_2 \times C}$, the query feature map $\hat{\mathbf{Q}}^2 = \{\hat{\mathbf{Q}}^2_1, \hat{\mathbf{Q}}^2_2, \ldots, \hat{\mathbf{Q}}^2_{k_1}\} \in \mathbb{R}^{k_1 \times k_2 \times C}$ and the value feature map $\hat{\mathbf{V}}^2 = \{\hat{\mathbf{V}}^2_1, \hat{\mathbf{V}}^2_2, \ldots, \hat{\mathbf{V}}^2_{k_1}\} \in \mathbb{R}^{k_1 \times k_2 \times C}$, respectively. The output feature map of the second step is $\hat{\mathbf{H}}' = \{\hat{\mathbf{H}}'_1, \hat{\mathbf{H}}'_2, \ldots, \hat{\mathbf{H}}'_{k_1}\} \in \mathbb{R}^{k_1 \times k_2 \times C}$. For each subset $\mathbf{W}_j$, we apply multiplication between $\hat{\mathbf{K}}^2_j$ and the transpose of $\hat{\mathbf{Q}}^2_j$ with a softmax to produce a self-attention map. Then $\hat{\mathbf{H}}'_j$ is calculated by multiplying between the self-attention map and $\hat{\mathbf{V}}^2_j$. The second step can be formulated as:

$$\hat{\mathbf{H}}'_j = \text{softmax}(\hat{\mathbf{K}}^2_j \hat{\mathbf{Q}}^{2\top}_j)\hat{\mathbf{V}}^2_j \tag{4}$$

After obtaining $\hat{\mathbf{H}}'$, we remove the expanded points and reshape it to $\mathbf{H}' \in \mathbb{R}^{N \times C}$. Then the random cross attention block is repeated again to get the enhanced context-aware feature map $\mathbf{H}''$. Theoretically, the attention maps of each point predicted by a random cross attention block only have about $2\sqrt{N}$ weights (if $k_1 = k_2 \approx \sqrt{N}$) which are much less then N in the non-local module [11], leading to reduce the computational complexity from $\mathcal{O}(N^{2\tilde{}}C)$ to $\mathcal{O}(N\sqrt{N}C)$.

Point-adaptive aggregation block. After two random cross attention blocks, how to connect the context-aware feature map $\mathbf{H}''$ and input features $\mathbf{G}$ to get $\mathbf{X}$ needs to be solved. Traditional DNNs aggregate different features using linear aggregations such as skip connection. But linear aggregations are not data-adaptive. To better reflect the characteristics of different points and make the aggregation data-aware, we propose a simple point-adaptive aggregation block. We use shared MLPs with one output channel for the two feature maps to produce two weight maps. Then a weighted summation to the two feature maps is performed to obtain $\mathbf{X}$. The whole procedure can be formulated as:

$$\begin{aligned} \boldsymbol{x}_i = &\frac{\exp(\text{MLP}(\boldsymbol{h}''_i))}{\exp(\text{MLP}(\boldsymbol{h}''_i)) + \exp(\text{MLP}(\boldsymbol{g}_i))}\boldsymbol{h}''_i \\ &+ \frac{\exp(\text{MLP}(\boldsymbol{g}_i))}{\exp(\text{MLP}(\boldsymbol{h}''_i)) + \exp(\text{MLP}(\boldsymbol{g}_i))}\boldsymbol{g}_i \end{aligned} \tag{5}$$

where $\boldsymbol{h}''_i$ is the i-th component of $\mathbf{H}''$.

III. Experimental Results

A. Datasets and Implementation Details

The proposed GA-Net is evaluated on three datasets, including Semantic3D [3], S3DIS [1], and ScanNet [2]. In the feature extractor, the U-Net parameters are consistent with the model before the FC layers in RandLA-Net [19]. The output dimensionalities of all the layers in the proposed two modules are 16. The FC has two layers, where the output dimensionalities are 64 and 32 respectively. We train our GA-Net using the Adam optimizer with initial learning rate 0.01 and batchsize 6 for 100 epochs.

B. Results on the Semantic3D Dataset

The Semantic3d dataset contains 30 outdoor scenes, of which 15 are used as training and the remaining are used as online testing. Each point cloud has up to 10^8 points with RGB and intensity values, and is labeled from 8 semantic categories. We conducted experiments on the semantic-8 challenge. To make a fair comparison, we calculated the mean of class-wise intersection over union (mIoU) and the overall point-wise accuracy (OA) following Fast-PCR [16].

We compared our GA-Net to several state-of-the-art methods according to the Semantic3D online evaluation website, as summarized in Table I. Here, we take RandLA-Net [19] as our baseline. In Table I, our GA-Net outperforms its baseline by 2.4% in terms of mIoU and 0.5% in terms of OA. The comparative results also show us that our method achieves best on both metrics, due to its more effective and efficient global contextual-feature learning. Our method outperforms the comparative methods for segmenting the objects (e.g. man-made terrain and natural terrain) which are of a relatively bigger size, but achieves lower performances than several comparative methods (particularly Fast-PCR [16]) for segmenting the objects (e.g. high vegetation and the low vegetation) which are of a relatively smaller size. This is mainly because: global features could generally reflect the characteristics of large-sized objects, while it generally needs local features for discriminating small-sized objects.

C. Results on the S3DIS Dataset

The S3DIS dataset contains 6 areas with 271 rooms in buildings. Each point, with xyz coordinates and RGB features, is annotated with one semantic label from 13 categories. We conducted our experiments in Area-5 validation. To make a fair comparison, the evaluation metrics we used are mIoU and OA following GACNet [12]. The quantitative results are reported in Table II. The proposed GA-Net has a highest mIoU among all the competitive methods on Area-5, which demonstrates that

TABLE I
SEMANTIC SEGMENTATION RESULTS (%) ON THE SEMANTIC3D DATASET (SEMANTIC-8)

Method	mIoU	OA	man-made terrain	natural terrain	high vegetation	low vegetation	buildings	hard scape	scanning artefacts	cars
PointNet++ [5]	63.1	85.7	81.9	78.1	64.3	51.7	75.9	36.4	43.7	72.6
EC-PointNet [17]	64.4	89.6	91.1	69.5	65.0	56.0	89.7	30.0	43.8	69.7
SnapNet [6]	67.4	91.0	89.6	79.5	**74.8**	56.1	90.9	36.5	34.3	77.2
PointGCR [22]	69.5	92.1	93.8	80.0	64.4	66.4	93.2	39.2	34.3	85.3
PointCE [18]	71.0	92.3	92.4	79.6	72.7	62.0	93.7	40.6	44.6	82.5
RandLA-Net [19]	71.9	94.1	95.9	88.3	65.5	61.7	95.9	**50.0**	27.6	90.2
Fast-PCR [16]	72.0	90.6	86.4	70.3	69.5	**68.0**	**96.9**	43.4	**52.3**	89.5
GA-Net(ours)	**74.3**	**94.6**	**96.7**	**91.5**	63.3	61.7	96.1	45.0	49.1	**91.3**

TABLE II
SEMANTIC SEGMENTATION RESULTS (%) ON THE S3DIS DATASET (AREA-5)

Method	mIoU	OA	ceiling	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	clutter
PointNet++ [5]	50.0	-	90.8	96.5	74.1	0.0	5.8	43.6	25.4	69.2	76.9	21.5	55.6	49.3	41.9
PointGCR [22]	54.4	-	90.7	96.1	74.9	**0.1**	16.1	50.2	32.3	69.0	78.1	41.3	60.7	53.8	43.8
PointCNN [8]	57.3	86.0	92.3	98.2	79.4	0.0	17.6	22.8	**62.1**	74.4	80.6	31.7	66.7	62.1	**56.7**
PointWeb [13]	60.3	87.0	92.0	**98.5**	79.4	0.0	21.1	59.7	34.8	76.3	**88.3**	46.9	69.3	64.9	52.5
RandLA-Net [19]	61.6	86.7	91.2	95.6	79.5	0.0	20.6	59.9	43.4	76.5	82.8	60.8	70.4	67.9	52.0
GACNet [12]	62.9	87.8	92.3	98.3	81.9	0.0	20.4	59.1	40.9	**85.8**	78.5	70.8	61.7	**74.7**	52.8
Point2Node [20]	63.0	**88.8**	**93.9**	98.3	**83.3**	0.0	**35.7**	55.3	58.8	79.5	84.7	44.1	**71.1**	58.7	55.2
GA-Net(ours)	**63.7**	87.6	92.9	97.8	81.3	0.0	27.8	**60.3**	41.7	78.3	86.7	**71.4**	69.9	65.8	53.9

TABLE III
SEMANTIC SEGMENTATION RESULTS (%) ON THE SCANNET DATASET

Method	Per-voxel accuracy
PointNet++ [5]	84.5
PointCNN [8]	85.1
PointGCR [22]	85.3
PointWeb [13]	85.9
RandLA-Net [19]	86.1
PointSIFT [9]	86.2
Point2Node [20]	86.3
GA-Net(ours)	**86.6**

TABLE IV
ABLATION STUDY OF THE SUB-MODULES ON THE S3DIS DATASET (AREA-5)

Method	mIoU(%)	FLOPS	params	memory(GB)	time(ms)
baseline	61.6	31333891	4993141	15.4	121
1-RCAB+plus	62.6	31378063	4999861	16.8	142
1-RCAB+PAB	62.9	31378527	4999931	17.1	146
2-RCAB+PAB	63.0	31422699	5006651	18.7	157
GA-Net(ours)	**63.7**	31434393	5009918	18.8	160

our method has a good ability to learn the global information. Comparing with the baseline, our method improves mIoU by 2.1% in terms of mIoU and 0.9% in terms of OA.

D. Results on the ScanNet Dataset

The ScanNet dataset contains 1513 scanned and reconstructed indoor scenes, which provides a 1201/312 scene split for training and testing. 20 categories are provided for evaluation. To make a fair comparison, we reported the per-voxel accuracy following Point2Node [20]. Table III shows the comparisons between our GA-Net and other competitive methods. Our method achieves the state-of-the-art performance, which improves on baseline by 0.5%, due to its more effective and efficient feature learning of long-range dependencies.

E. Ablation Study

For ablation study, we stacked the proposed sub-modules on the baseline step-to-step to prove the effectiveness of our method. Specifically, the comparing experiments are (1) baseline, (2) adding one random cross attention block (RCAB) and aggregating features with a plus operation, denoted as "1-RCAB+plus," (3) adding one random cross attention block and aggregating features by point-adaptive aggregation block (PAB), denoted as "1-RCAB+PAB," (4) adding two random cross attention blocks and the rest is consistent with (3), denoted as "2-RCAB+PAB," and (5) our proposed GA-Net. Our baseline method employs a U-Net based on RandLA-Net [19]. We conducted ablation study on Area-5 of the S3DIS with the evaluation metric mIoU. Besides, we also made statistics on the floating-point operations per second (FLOPs), number of parameters, computing memory, and computing time to prove the efficiency of our method.

As shown in Table IV, "1-RCAB+plus" performing better than baseline demonstrates that the importance of exploring point-dependent global information. More interestingly, the result of "1-RCAB+PAB" achieves better than "1-RCAB+plus," which may be attributed to the data-aware aggregating method. "2-RCAB+PAB" performing better than "1-RCAB+PAB" indicates that stacking more random cross attention blocks benefits context-aware feature learning. Our proposed method achieves best, indicating that combining two kinds of global information can further improve results. Furthermore, the last four columns in Table IV demonstrate that our method does not require much calculation, memory and time. But stacking more random cross attention blocks results in waste of resources, so only two of it is considered.

Remark: It is pointed out that the computational complexity of the non-local module [11] is much greater than the random cross attention block as mentioned in Section II-C. Hence we did not train the network consisting of the baseline and the non-local module since our GPU memory(32 GB) can not meet the storage requirements.

IV. CONCLUSION

For obtaining long-range contextual information of 3D point clouds, we propose a global attention network, called GA-Net, which consists of a point-independent global attention module, and a point-dependent global attention module. These two modules can obtain global information in an efficient way. Extensive experiments on three point cloud benchmarks demonstrate that our method outperforms state-of-the-art methods in most cases.

REFERENCES

[1] I. Armeni *et al.*, "3D semantic parsing of large-scale indoor spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1534–1543.

[2] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2432–2443.

[3] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3d.net: A new large-scale point cloud classification benchmark," in *Proc. ISPRS Annals Photogr. Remote Sens. Spatial Inf. Sci.*, 2017, pp. 91–98.

[4] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[5] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[6] A. Boulch, B. L. Saux, and N. Audebert, "Unstructured point cloud semantic labeling using deep segmentation networks," in *Proc. Workshop 3D Object Retrieval*, 2017, pp. 17–24.

[7] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," in *Proc. ACM Trans. Graph.*, 2019, pp. 1–12.

[8] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on x-transformed points," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 820–830.

[9] M. Jiang, Y. Wu, and C. Lu, "Pointsift: A sift-like network module for 3D point cloud semantic segmentation," 2018, *arXiv: 1807.00652*.

[10] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao, "SpiderCNN: Deep learning on point sets with parameterized convolutional filters," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 87–102.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[12] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, "Graph attention convolution for point cloud semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10296–10305.

[13] H. Zhao, L. Jiang, C.-W. Fu, and J. Jia, "Pointweb: Enhancing local neighborhood features for point cloud processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5565–5573.

[14] Z. Huang *et al.*, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[15] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1971–1980.

[16] G. Truong, S. Z. Gilani, S. M. S. Islam, and D. Suter, "Fast point cloud registration using semantic segmentation," *Digit. Image Comput.: Techn. Appl.*, 2019, pp. 1–8.

[17] J. Contreras and J. Denzler, "Edge-convolution point net for semantic segmentation of large-scale point clouds," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5236–5239.

[18] H. Liu, Y. Guo, Y. Ma, Y. Lei, and G. Wen, "Semantic context encoding for accurate 3D point cloud segmentation," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.3007331.

[19] Q. Hu *et al.*, "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11105–11114.

[20] W. Han, C. Wen, C. Wang, X. Li, and Q. Li, "Point2node: Correlation learning of dynamic-node for point cloud feature modeling," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10925–10932.

[21] J. Wei, G. Lin, K.-H. Yap, T.-Y. Hung, and L. Xie, "Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4383–4392.

[22] Y. Ma, Y. Guo, H. Liu, Y. Lei, and G.-J. Wen, "Global context reasoning for semantic segmentation of 3D point clouds," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2920–2929.

[23] W. Jiang, Z. Xie, Y. Li, C. Liu, and H. Lu, "LRNNET: A light-weighted network with efficient reduced non-local operation for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.

[24] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, and F. Wang, "SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.

[25] S. Deng, B. Liu, Q. Dong, and Z. Hu, "Rotation transformation network: Learning view-invariant point cloud for classification and segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021.