Progressive Scene Segmentation Based on Self-Attention Mechanism

Yunyi Pan*, Yuan Gan*, Kun Liu*, Yan Zhang*

*State Key Lab for Novel Software Technology, NanJing University

Abstract: Semantic scene segmentation is vital for a large variety of applications as it enables understanding of 3D data. Nowadays, various approaches based upon point clouds ignore the mathematical distribution of points and treat the points equally. The methods following this direction neglect the imbalance problem of samples that naturally exists in scenes. To avoid these issues, we propose a two-stage semantic scene segmentation framework based on self-attention mechanism and achieved stateof-the-art performance on 3D scene understanding tasks. We split the whole task into two small ones which efficiently relief the sample imbalance issue. In addition, we have designed a new self-attention block which could be inserted into submanifold convolution networks to model the long-range dependencies that exists among points. The proposed network consists of an encoder and a decoder, with the spatial-wise and channel-wise attention modules inserted. The two-stage network shares a UNet architecture and is an end-to-end trainable framework which could predict the semantic label for the scene point clouds fed into it. Experiments on standard benchmarks of 3D scenes implies that our network could perform at par or better than the existing state-of-the-art methods.

1.Two-Stage submanifold convolution network

1.1 Overall Architecture

We split the whole task into two small ones, which makes the segmentation task much easier. The building such as wall and floor exist in most scenes could be regarded as the background, therefore, the other object exist in the scene defined as foreground. In order to improve the performance of the segmentation, we first separate the background from the scene so that the segmentation in next stage would be much more concise. Combining the results of two stage segmentation, finally we get the semantic segmentation result of the whole scene. The framework we designed is a twostage decomposition framework which consumes voxelized point clouds of a 3D scene as input and output the semantic labels corresponding with the input points. Figure 1 shows the overall architecture of our framework.

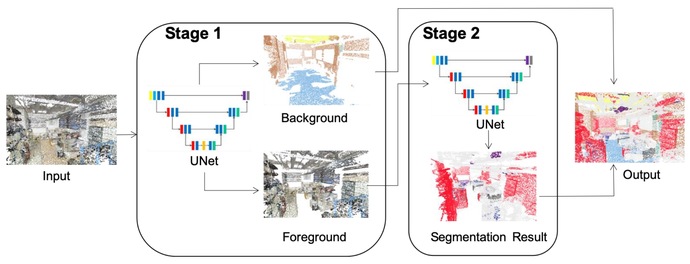

Fig. 1: Overall architecture of the two-stage submanifold sparse convolution network based on self-attention mechanism

1.2 Stage-1

In stage 1, we extracts points that belongs to 'background' from the complete scene. As a result, the original scene point clouds split into two separated ones. The one only consist of 'background' points, while the other consist the rest of the point clouds. We designed a U-Net architecture which consist of an encoder and a decoder. Both of the encoder and the decoder composed of several sparse convolution operators. The encoder encodes the point features by continuous sparse convolution and submanifold sparse convolution blocks which are inserted into the interval of convolution operations that used for keeping the sparsity among the convolutions. The basic building blocks for our framework are pre-activated residual blocks that contain two submanifold sparse convolution SSC(; ; 3) where filter size is 3. Each convolution is preceded by batch normalization and a ReLU non-linearity. We use seven level convolutions accompanied by two pre-activated residual blocks each level in the downsampling process. The attention blocks of both spatial and channel domains are inserted into the last level of convolution block. The decoder just conduct the inverse operations combining with the features which come from the skip connections. In this stage, we use a sparse voxelized input representation similar to [29], and a combination of SSC operations and strided SC convolutions to construct sparse variants of the U-Net networks. we process the original scene point clouds and split it into several class including 'background' and the others

1.3 Stage-2

In stage 2, Obtaining the remaining point clouds from the first stage, we feed them into the same U-Net in stage 1. In this stage, we will get the semantic labels for the remaining point clouds. Summarize the two stage semantic segmentation results, finally we assign semantic labels to the scene point clouds that fed into our framework. Combining the result of two stages, we will get the semantic labels of all point clouds.

2. Self-Attention Block

2.1 Overall Attention Block

Since convolution operations would lead to a limited local receptive field, the features corresponding to the points with same semantic label may have some differences. These differences would bring some noises to the final linear layer, which affects the segmentation accuracy. To address this problem, we find a way to build global associations among features with attention mechanism. The framework we designed could efficiently aggregate long-range contextual information and non-linearity relationship between channels, thus improving the feature representation for 3D scene segmentation. As shown in Figure 2, we design two types of self-attention modules.

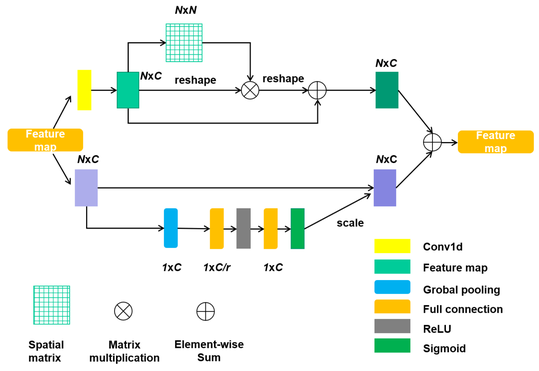

Fig. 2: Spatial-wise and channel-wise self-attention block

1.2 Spatial Attention Module

Discriminant feature representation is the most fundamental things for scene understanding. However, the result of many works imply that there are some limitations in extracting local features only. The feature captured by stacked convolution block may lead to misclassification of the objects. To model long-range dependencies over local features, we proposed a spatial attention module. The spatial attention module enhance the feature representation by re-weighting local features according to its correlations. The details of the module could be found in Figure 3

1.3 Channel Attention Module

It is not intuitive to clearly explain the relationship between channels, but there are nonlinear relations exist in high dimensions obviously. By reweighting the channel maps, we could strengthen useful channels and ignore the noises from additional channels. Therefore, we have proposed attention module which is a channel filtering module that models the interdependencies between channels. The details of the module could be found in Figure 4

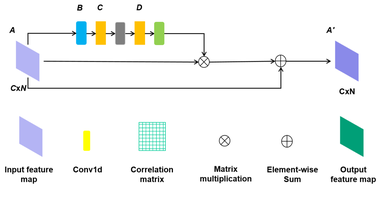

Fig. 3: Spatial-wise self-attention block

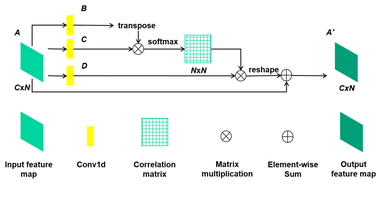

Fig. 4: Channel-wise self-attention block

3. Results

2.1 Experiments on ScanNet & S3DIS

we first introduce the basic experiments settings that we adopt. Then we provide analysis experiments to understand the significance of the progressive segmentation. Finally, we show qualitative results of our method. To validate the proposed framework, we use standard 3D scene benchmarks for 3D semantic segmentation. It makes our methods easier to compare with the others.

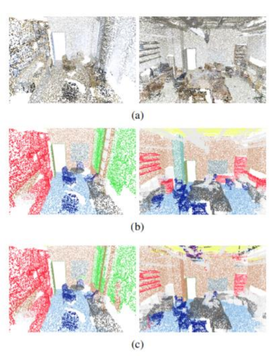

Fig. 5: Visualization of Stanford dataset Area 5 test results. From the top, RGB input (a), ground truth (b), our result (c).

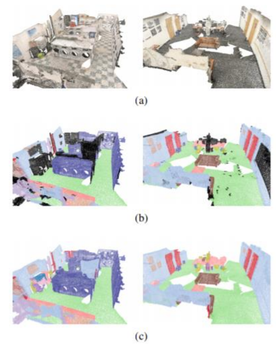

Fig. 6: Visualization of ScanNet validation results. From the top, RGB input (a), ground truth (b), our result (c).

TABLE I: Comparison with State-of-the-art on ScanNet validation set

Methods	mIoU(%)	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	refrigerator	shower curtain	toilet	sink	bathtub	otherfurniture
PointNet++[[2])	33.9	52.3	67.7	25.6	47.8	36.0	34.6	23.2	26.1	25.2	45.8	11.7	25.0	27.8	24.7	21.2	58.4	14.5	54.8	36.4	18.3
SSCNet [6])	70.821	83.6	95.1	65.3	80.7	90.4	82.0	72.2	64.3	60.5	78.0	31.3	62.5	58.7	75.8	49.4	70.8	93.0	63.9	87.4	51.4
Minkowski [7])	70.558	84.5	95.9	63.9	80.8	90.1	81.5	70.9	59.8	60.6	73.4	31.5	66.0	60.5	71.3	55.6	66.5	90.3	65.2	83.5	56.6
Ours	71.553	**85.2**	95.3	**67.0**	80.2	90.2	**84.8**	70.4	61.8	**63.9**	76.1	31.2	65.3	**60.9**	64.2	**56.0**	**83.4**	**93.4**	**66.5**	85.5	49.8

TABLE II: Stanford Area 5 Test (Fold 1) (S3DIS)

Method	mIOU	mAcc
PointNet [1]	41.09	48.98
SparseUNet [30]	41.72	64.62
SegCould [10]	48.92	57.35
TangentConv [31]	52.8	60.7
3D RNN [32]	53.4	71.3
PointCNN [3]	57.26	63.86
SuperpointGraph [33]	58.04	66.5
GACNet [26]	62.85	87.79
MinkowskiNet [7]	65.35	71.71
SparseConvNet [6]	66.7	-
Ours(without attention module)	69.5	-
Ours	**70.9**	-

TABLE III: Result of the network under different parameters on ScanNet val set.

Framework	Scale	Attention-block	Two Stage	mIOU
SSCN	20			0.614
ASSCN	20	✓		0.622
PSSCN	20		✓	0.632
ours	20	✓	✓	0.635
SSCN	50			0.708
ASSCN	50	✓		0.710
PSSCN	50		✓	0.713
ours	50	✓	✓	0.715

References

[1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[2] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in Advances in neural information processing systems, 2017, pp. 5099–5108.

[3] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in Advances in Neural Information Processing Systems, 2018, pp. 820–830.

[4] J. Li, B. M. Chen, and G. Hee Lee, "So-net: Self-organizing network for point cloud analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9397–9406.

[5] M. Jaritz, J. Gu, and H. Su, "Multi-view pointnet for 3d scene understanding," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[6] B. Graham, M. Engelcke, and L. van der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," CVPR, 2018.

[7] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3075–3084.

* Please contact *yanzhangnju@nju.edu.cn* for further information.