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

Releated Work

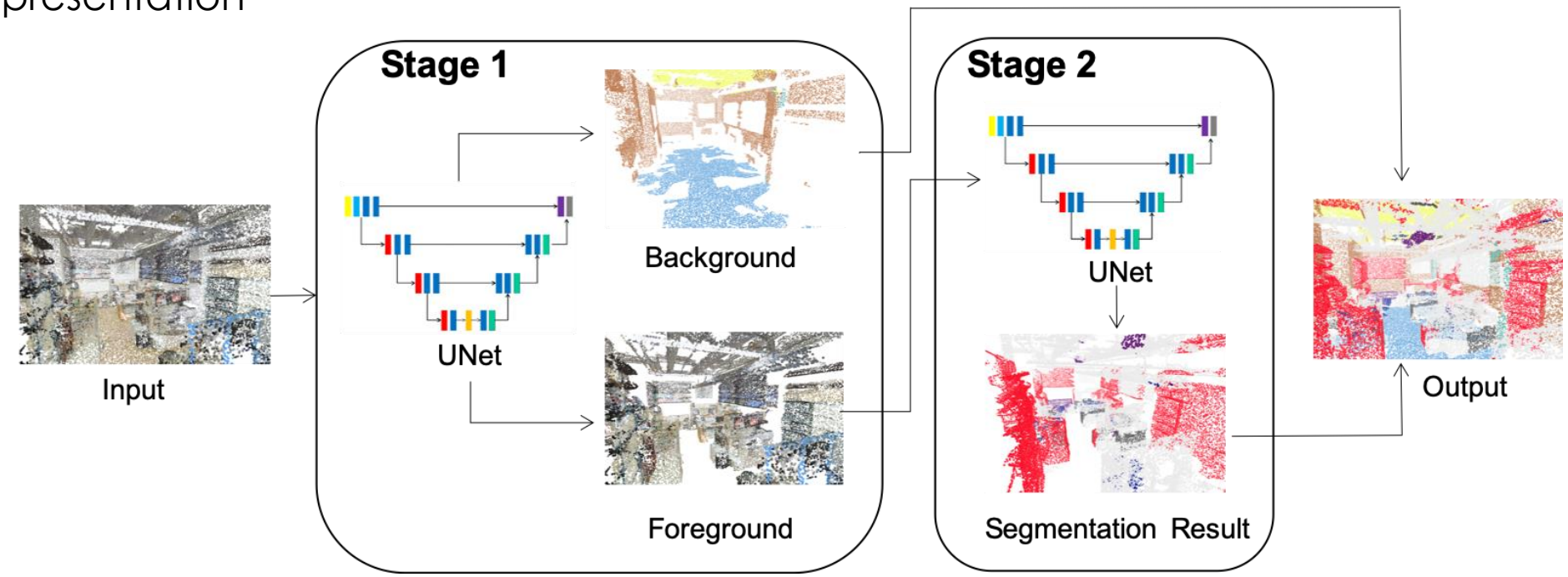
- PointNet first proposed a framework that learns point features directly from unordered point sets.
- PointNet++ extract local features capturing fine geometric structures from small neighborhoods, which outperform the PointNet.
- SparseConv Network adopt sparse tensors and propose the generalized sparse convolution which encompasses all discrete convolutions.

Motivation

- Various approaches based upon point clouds ignore the mathematical distribution of the points and treat the point equally.
- Most of the methods 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.

Overall Architecture

We split the whole scene into 'foreground' and 'background' using the same architecture, which designed as an end to end trainable framework. In addition, the designed attention blocks would be inserted into the bottom of the U-Net to enhance the ability of feature representation

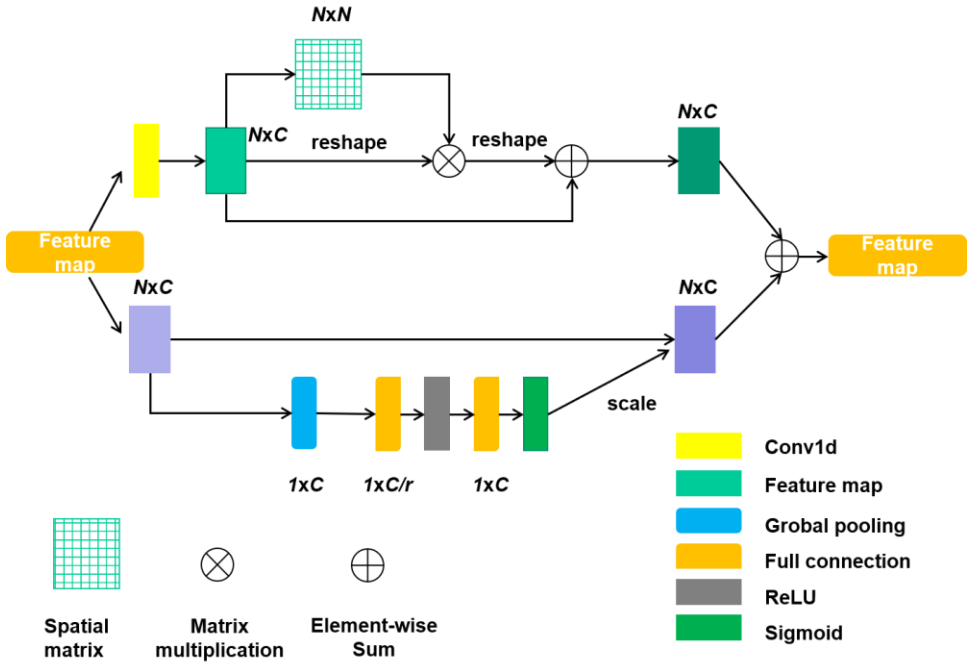


Method Details

- Two-stage submanifold convolution network
  - 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
  - Combining the results of two stage segmentation, finally we get the semantic segmentation result of the whole scene

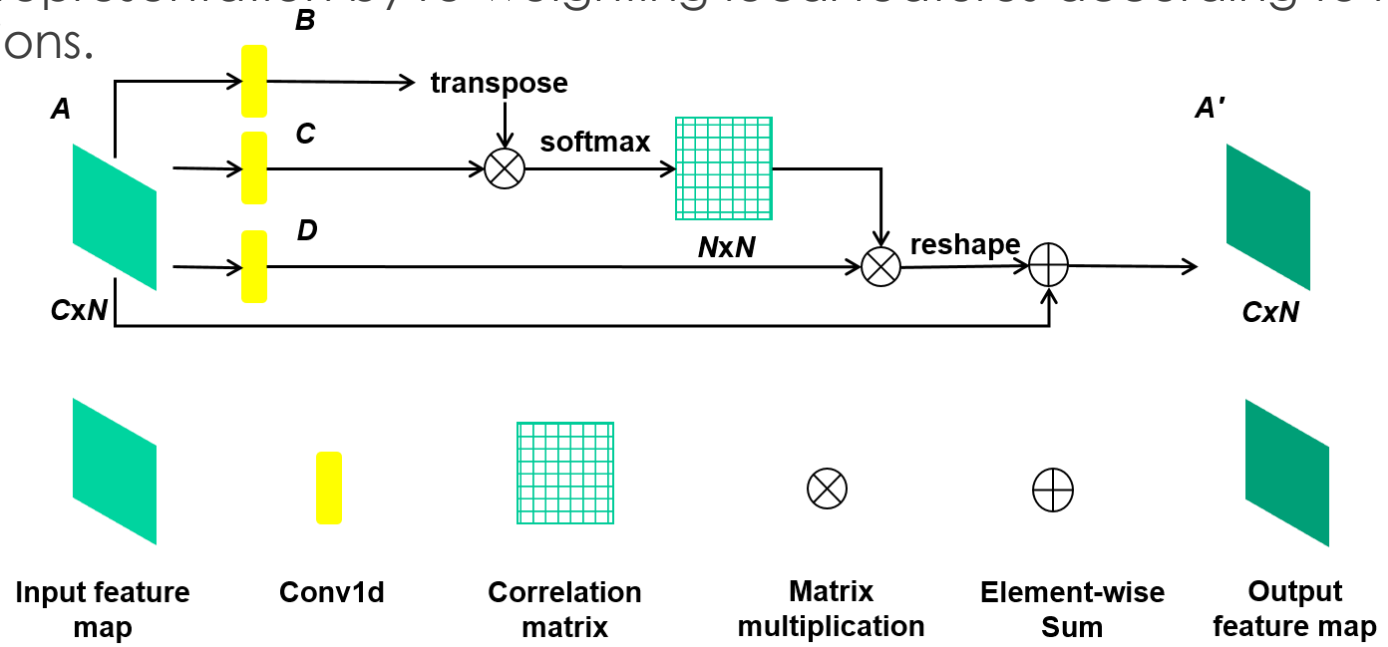
Method Details

- Self-Attention block
  - The framework we designed could efficiently aggregate long-range contextual information and non-linearity relationship between channels
  - Spatial Attention Module
  - Channel Attention Module



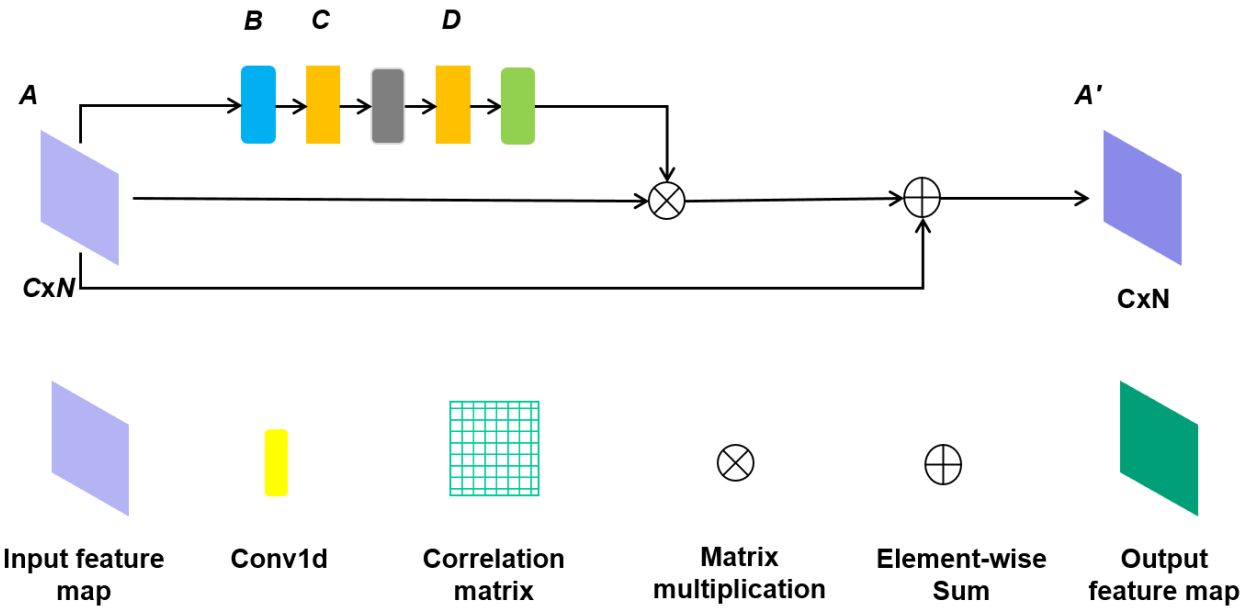
Spatial Attention Module

- 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.



Channel Attention Module

- 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.



Results

Result on Stanford 3D Indoor Dataset

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**	-

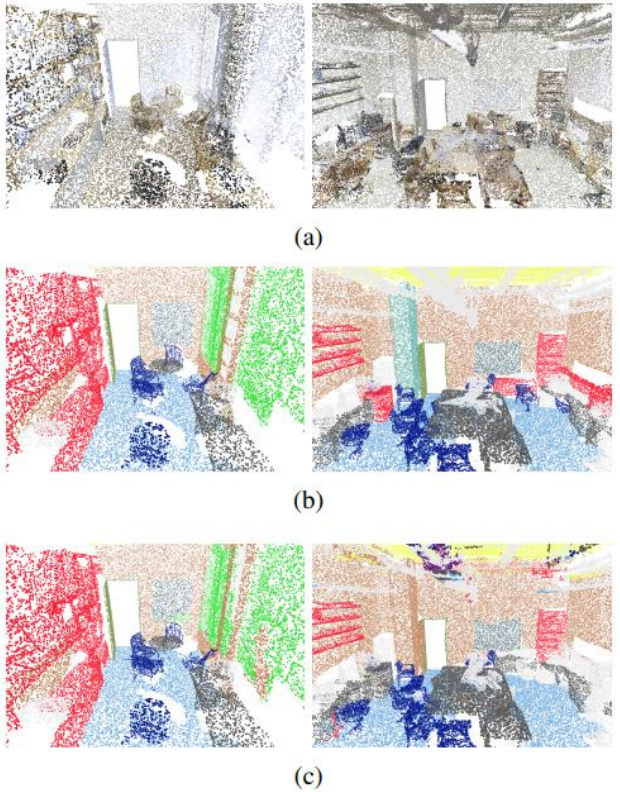


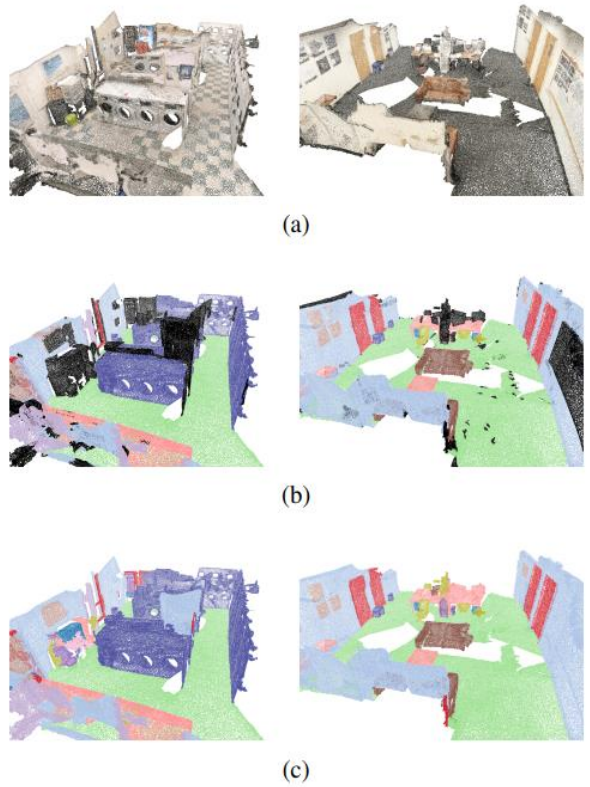
Fig. 5: Visualization of Stanford dataset Area 5 test results. From the top, RGB input (a), ground truth (b), our result (c).

Results

Result on ScanNet Dataset

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
SSCNs([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	75.4	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



(a)

(b)

(c)

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

Conclusion

In this paper, we have presented a progressive scene segmentation framework based upon self-attention mechanism. To some extent, it alleviates the problem of category imbalance. Specifically, we introduce a spatial-wise and channel-wise attention block that suit the submanifold sparse convolution networks.


Thank You