SCANET: SPATIAL-CHANNEL ATTENTION NETWORK FOR 3D OBJECT DETECTION

*Haihua Lu, Xuesong Chen, Guiying Zhang, Qiuhao Zhou, Yanbo Ma, Yong Zhao**

School of Electronic and Computer Engineering, Shenzhen Graduate School of Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055

ABSTRACT

This paper aims to achieve high-accuracy 3D object detection, in which we propose a novel *Spatial-Channel Attention Network (SCANet)*, a two-stage detector that takes both LIDAR point clouds and RGB images as input to generate 3D object estimates. The first stage is a 3D region proposal network (RPN) in which we put forward a new Spatial-Channel Attention (SCA) module and an Extension Spatial Upsample (ESU) module. Using the pyramid pooling structure and global average pooling, the SCA module can not only effectively incorporate multi-scale and global context information, but also produce spatial and channel-wise attention to select discriminative features. The ESU module in the decoder can recover the lost spatial information caused by consecutive pooling operators to generate reliable 3D region proposals. In the second stage, we design a new multi-level fusion scheme for accurate classification and 3D bounding box regression. Experimental results demonstrate that SCANet achieves state-of-the-art performance on the challenging KITTI 3D object detection benchmark.

***Index Terms*—** 3D object detection, attention mechanism, spatial upsample, fusion scheme

1. INTRODUCTION

In recent years, plenty of research works have promoted the remarkable progress of 2D object detection which can indicate the position of each object in the image coordinate system and their category [1, 2, 3, 4, 5]. However, in numerous applications like autonomous driving and robot navigation, using 2D detection results only is insufficient to describe objects in 3D real-world scenarios. Therefore, 3D object detection is indispensable in these applications, which can provide additional depth and orientation information [6, 7, 8]. Also, due to the addition of a third dimension, how to improve the accuracy of 3D object detection is still a difficult problem.

Based on the fact that LIDAR point clouds are able to provide more accurate depth information while camera images have more detailed semantic information, we use both point clouds and RGB images as input to build a robust 3D object detection framework and take three main problems of existing 3D object detectors into consideration.

The first issue is that most existing 3D object detection methods fail to extract multi-scale and global context features to encode local and global information. Meanwhile, they are short of precise spatial and channel-wise attention to recalibrate pixel-level and channel-level features. Most methods still use a single feature map for detection directly [9]. Some other methods improve the detection accuracy through feature concatenation [8] and feature pyramid [10, 11]. However, these methods fail to take global information and attention mechanism into account. Inspired by SENet [12] and PAN [13], we design a new Spatial-Channel Attention (SCA) module to simultaneously generate multi-scale and global context features as well as spatial and channel-wise attention to detect object more accurately.

The second issue is that many existing methods are unable to adequately recover the reduced spatial information caused by continuous downsampling in feature extractors. Due to the sparsity of point clouds, objects occupy a small number of pixels in the output feature map, resulting in the loss of spatial information, which is unfavorable for the detection. Most 3D object detection networks apply bilinear interpolation to upsample feature maps directly [7] or just combine the features of corresponding stages [10], which leads to the high-level features receiving only limited spatial information from the corresponding low-level features in the encoder. To solve this issue, we propose an Extension Spatial Upsample (ESU) module in the decoder to provide finer spatial information by combining multi-scale shallow layer features.

The last issue is that we need a better fusion scheme to merge features from the 2D projection of point clouds and RGB images. Early fusion [14] and late fusion [15] only fuse features in the input stage and the prediction stage, respectively, which lacks enough interactions among features. Inspired by deep fusion [7], which alternately performs feature transformation and feature fusion, we propose multi-level fusion to enable more interactions among different view features.

In summary, there are four contributions in this paper:

- We propose a Spatial-Channel Attention (SCA) module to capture multi-scale and global context information while obtaining both spatial attention and channel-wise attention to recalibrate features. To the best of our knowledge, this

*Corresponding author

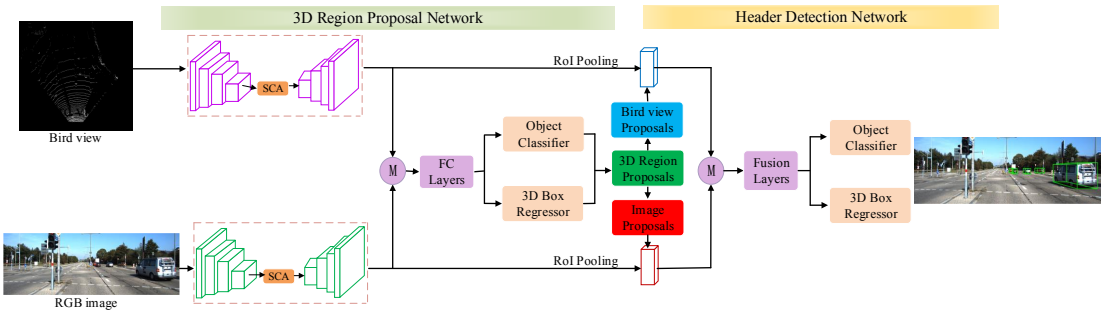


Fig. 1. Overview of the Spatial-Channel Attention Network (SCANet).

is the first work of applying the attention mechanism in 3D object detection task.

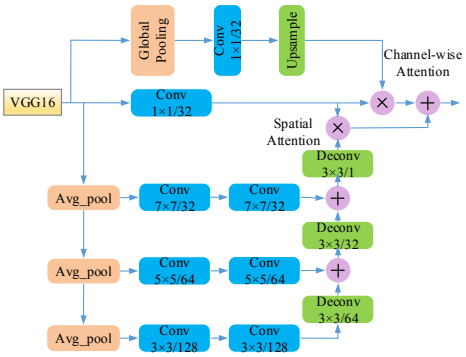
- We present an Extension Spatial Upsample (ESU) module to enrich the spatial information of the high-level features by combining multi-scale low-level features.
- We design a new multi-level fusion scheme to enable more interactions among features from different views so that we can fuse them better.
- Combining the three modules mentioned above, we propose a Spatial-Channel Attention Network(SCANet) for 3D object detection. Experimental results show that our method achieves state-of-the-art results on the KITTI 3D object detection benchmark.

2. SPATIAL-CHANNEL ATTENTION NETWORK

The proposed method is depicted in Fig. 1. Firstly, the feature extractor takes the bird's eye view maps of point clouds and RGB images as input and then passes the extracted features to region proposal network to generate 3D region proposals. Next, the region-based features are fused by the multi-level fusion layers. Finally, the fused features are used to predict object class and regress oriented 3D bounding boxes.

2.1. 3D Point Clouds Representation

3D point clouds are sparse and unstructured. We adopt a compact representation by projecting 3D point clouds into the bird's eye view (BEV) [7]. We form a six-channel BEV map in which the first five channels encode height information and the last one encodes intensity information. We project point clouds into a 2D map and then discretize them with a resolution of 0.1m. For the height channels, we divide the point cloud into five equal slices, each of which produces a height map representing the maximum height of the points in each cell. For the intensity channel, we compute it as the number of points in each cell for the whole point cloud. To obtain homogeneous and significative values among all cells, we normalize the intensity feature map by the maximum possible number of points as $min(1.0, \frac{log(N+1)}{log16})$, which inspired by Log function conversion normalization method.



(a) Spatial-Channel Attention (SCA)

(b) Spatial and channel-wise attention fusion

Fig. 2. Schematic diagram of Spatial-Channel Attention module. (a) Spatial-Channel Attention module structure. The numbers in the convolution boxes represent the kernel size and the number of output channels. (b) Diagrammatic drawing of the fusion of spatial attention and channel-wise attention.

2.2. Feature Extractor

The proposed network has two identical encoder-decoder feature extractors, one for the bird's eye view of point clouds and the other for the RGB images. We use the VGG-16 as the encoder network but with two modifications: the channels are reduced to half of the original and the network is cut off at the conv-4 layer. Then the Spatial-Channel Attention module is used to extract multi-scale and global context features to encode local and global information. Besides, it can produce both spatial attention and channel-wise attention, which is capable of recalibrating features spatially and channel-wisely, thus we can strengthen the discriminative features and restrain the indiscriminative features. Finally, we design an Extension Spatial Upsample module, which combines adjacent low-resolution feature maps as multi-scale low-level features to help high-level features gain finer spatial information.

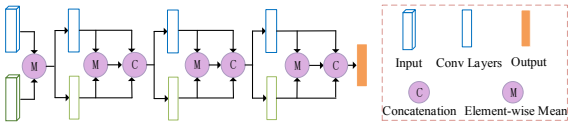


Fig. 3. Schematic diagram of Multi-level Fusion methods.

Spatial-Channel Attention (SCA). The Spatial-Channel Attention module consists of spatial attention block and channel-wise attention block, as shown in Fig. 2(a). The spatial attention block adopts pyramid pooling structure which includes three different pyramid scales and sequentially uses 7×7, 5×5, and 3×3 convolutions. Instead of directly upsampling the low-dimension feature maps to the original size separately, we gradually upsample and merge different scales to obtain more precise multi-scale information. The spatial attention block finally outputs a single channel attention map incorporating multi-scale context information, which is then multiplied by the original features that are compressed by a 1×1 convolution to reweight features spatially. On the other hand, the channel-wise attention block applies global pooling to provide global context information and output channel-wise attention map to select the features channel-wisely. Finally, spatial attention is fused with channel-wise attention, as illustrated in Fig. 2(b).

Extension Spatial Upsample (ESU). Our Extension Spatial Upsample module is designed to provide detailed spatial information to high-level features in the decoder by combining multi-scale low-level features in the encoder. We first use a 3×3 convolution with a stride of 2 to downsample the features from Conv layer n-1 in the encoder and use another 3×3 convolution with a stride of 1 to refine it. Then two 3×3 convolutions with a stride of 1 are performed on the features from Conv layer n in the encoder. After that, we add the features from the two different layers as the multi-scale low-level features. Finally, we concatenate the low-level features with corresponding high-level features and further fuse them by a 3×3 convolution.

2.3. 3D Region Proposal Network

Given the features from the bird's eye view maps and RGB images, RPN first fuses them via an element-wise mean operation and then generates 3D region proposals by regressing the difference between a set of 3D prior boxes and the ground truth boxes. Each 3D prior box is parameterized by (x, y, z, l, w, h), which encodes the center coordinates and dimensions of the anchor. (x, y) is the different position in the bird's eye view with a resolution of 0.5 meters, and z can be computed by the camera height above the ground plane. (l, w, h) is provided by clustering ground truth bounding box sizes in the training set. During training, we use a multi-task loss following Fast R-CNN [2] which includes a cross-entropy loss for binary classification and a smooth L1 loss for regression.

Table 1. Performance comparison on KITTI validation set: Average Precision (AP_{3D}) (in %) of 3D boxes.

Model	Data	3D Detection		
		Easy	Moderate	Hard
Mono3D [9]	Mono	2.53	2.31	2.31
3DOP [17]	Stereo	6.55	5.07	4.10
VeloFCN [18]	LIDAR	15.20	13.66	15.98
MV3D [7]	LIDAR+Mono	71.29	62.68	56.56
VoxelNet [8]	LIDAR	81.97	65.46	62.85
SCANet (ours)	LIDAR+Mono	**83.63**	**74.47**	**67.78**

2.4. Header Detection Network

We first project the 3D proposals into the bird's eye view and the RGB image plane. Then due to the different dimensions of features from them, we apply RoI pooling [2] to obtain feature crops with the same size of 7×7. A new multi-level fusion scheme is proposed to fuse the region-based feature crops. Given the fused features, we regress oriented 3D bounding boxes from the 3D proposals.

Multi-level Fusion. There are three main fusion methods: early fusion, late fusion, and deep fusion. Early fusion [14] fuses multi-view features in the input stage and then uses a single sub-network to predict. Late fusion [15] first uses multiple sub-networks to learn feature transformation separately and then merges features in the prediction stages. Unlike the first two fusion methods, deep fusion [7] alternately performs feature transformation and feature fusion. Inspired by deep fusion, we propose a new multi-level fusion scheme. We first merge the features from the bird's eye view and RGB images by element-wise mean operation and then concatenate them as the input of the next step, which can enable more interactions among different view features, as described in Fig. 3.

3D Bounding Box Regression. The 3D proposals are parameterized by $(x_1, x_2, x_3, x_4, y_1, y_2, y_3, y_4, h_1, h_2)$ that encodes the coordinates of 4 corners and 2 heights above ground plane. Similar to the RPN, we apply a multi-task loss combining a cross-entropy loss for classification and a smooth L1 loss for regression to jointly predict object categories and oriented 3D bounding boxes.

3. EXPERIMENTS

We evaluate our SCANet on the challenging KITTI 3D object detection benchmark [16] which contains 7481 training images and 7518 testing images along with point clouds.

Implementation Details and Metrics. We train the network in an end-to-end fashion and use ADAM optimizer for roundly 130K iterations on a NVIDIA 1080Ti GPU with an initial learning rate of 0.0001, which decays exponentially every 20K iterations with a decay factor of 0.9. In our experiments, we focus on the car category as KITTI provides enough car instances to train deep network. Following the KITTI official setting, we do the evaluation on three difficulty

Table 2. Results of Spatial-Channel Attention module and Extension Spatial Upsample module on KITTI validation set.

Model	3D Detection		
	Easy	Moderate	Hard
VGG16+ESU	77.89	68.28	66.23
VGG16+SE [12]+ESU	77.23	73.15	67.55
VGG16+PPM [19]+ESU	77.86	73.44	67.17
VGG16+FPA [13]+ESU	83.56	68.12	67.40
VGG16+SCA+ESU	**83.63**	**74.47**	**67.78**
VGG16+SCA	67.70	60.68	56.05
VGG16+SCA+BI	71.56	62.10	56.28
VGG16+SCA+ESU	**83.63**	**74.47**	**67.78**

Table 3. Performance comparison of different fusion methods on KITTI validation set.

Model	3D Detection		
	Easy	Moderate	Hard
Late Fusion [15]	74.94	64.30	63.82
Early Fusion [14]	83.27	68.45	67.29
Deep fusion [7]	83.32	73.61	67.18
Multi-level Fusion	**83.63**	**74.47**	**67.78**

regimes: easy, moderate, and hard, which depend on the object size, occlusion state, and truncation level. We use Average Precision (AP_{3D}) computed at 0.7 IoU in all experiments for full 3D bounding boxes evaluation.

3.1. Evaluation on KITTI Validation Set

Like many methods, we subdivide the KITTI training data into a training set and a validation set with a ratio of about 1:1.

Spatial-Channel Attention Network (SCANet). We first compare our SCANet with other state-of-the-art methods which publicly provide detection on the KITTI validation set. As shown in Table 1, our SCANet exceeds all the competing methods across all the three difficulty regimes, which suggests that SCANet is effective to detect objects of different scales.

Spatial-Channel Attention (SCA). The results are presented in table 2. Take the moderate difficulty level as an example, using the SCA modules, the accuracy can be improved from 68.28% to 74.47%, which shows that the SCA module can significantly improve performance. Besides, we compare our SCA module with other context modules or attention modules. We can see that the performance of those modules is not balanced at three difficulty levels while our SCA module consistently outperforms those modules across all difficulty levels, which demonstrate that our SCA module can extract spatial and channel-wise attention information effectively.

Extension Spatial Upsample (ESU). The results are shown in Table 2. Take the moderate difficulty level as an example, the VGG16 baseline combining with the SCA module achieves an average precision of 60.68%. Using bilinear interpolation to upsample features merely improves performance by 1.42% while using our ESU module to upsample features can significantly improve performance by 13.79%, which demonstrates that the ESU module is helpful to recover spatial information for accurate detection.

Table 4. Performance on KITTI test set.

Benchmark	Easy	Moderate	Hard
Car (3D Detection)	76.09	66.30	58.68

Fig. 4. Qualitative results.

Multi-level Fusion. The results are summarized in Table 3, which attest that our multi-level fusion scheme can fuse features from different views effectively and outperforms other fusion methods.

3.2. Benchmark Results

To evaluate our SCANet on the KITTI test set, we submit the results to the KITTI 3D object detection official server. The results are shown in Table 4.

Finally, the qualitative results are depicted in Fig. 4. It can be seen that in the case of small objects and multiple objects, our method can accurately detect all objects, which indicates that our SCANet is a robust 3D detector.

4. CONCLUSION

In this paper, we have proposed a novel effective framework named Spatial-Channel Attention Network for challenging 3D object detection. Firstly, we propose a new Spatial-Channel Attention module, which is capable of encoding multi-scale and global context information and producing spatial and channel-wise attention to select discriminative features spatially and channel-wisely. Secondly, to generate reliable 3D region proposals, we design an Extension Spatial Upsample module, which uses multi-scale low-level features to guide high-level features to recover spatial information. Finally, a new multi-level fusion scheme is presented to fuse multi-view features for final oriented 3D bounding box regression. Our experimental results show that the proposed method outperforms the state-of-the-art approaches on the KITTI 3D object detection benchmark while running at 11 FPS on an NVIDIA 1080Ti GPU.

5. ACKNOWLEDGEMENT

This work was supported by Science and Technology Planning Project of Shenzhen(No. NJYJ20170306091531561), Science and Technology Planning Project of Shenzhen (No. JCYJ2016050617265 1253), and National Science and Technology Support Plan, China(No. 2015BAKO1B04).

6. REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[2] Ross Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[6] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[7] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia, "Multi-view 3d object detection network for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6526–6534.

[8] Yin Zhou and Oncel Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun, "Monocular 3d object detection for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[10] Bin Yang, Wenjie Luo, and Raquel Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.

[11] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander, "Joint 3d proposal generation and object detection from view aggregation," *arXiv preprint arXiv:1712.02294*, 2017.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[13] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[14] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*, 2016, pp. 354–370.

[15] Shuran Song and Jianxiong Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 808–816.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[17] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun, "3d object proposals for accurate object class detection," in *International Conference on Neural Information Processing Systems*, 2015, pp. 424–432.

[18] Bo Li, Tianlei Zhang, and Tian Xia, "Vehicle detection from 3d lidar using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *International Conference on Neural Information Processing Systems*, 2017, pp. 2881–2890.