IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 18, 2021, accepted May 2, 2021, date of publication May 7, 2021, date of current version May 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3078371

AMSASeg: An Attention-Based Multi-Scale Atrous Convolutional Neural Network for Real-Time Object Segmentation From 3D Point Cloud

MOOGAB KIM, NAVEED ILYAS, (Student Member, IEEE), AND KISEON KIM, (Senior Member, IEEE)

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea

Corresponding author: Moogab Kim (anrkq1024@gist.ac.kr)

This work was supported in part by the Ministry of Oceans and Fisheries, South Korea through the Project Development of Automatic Identification Monitoring System for Fishing Gears.

ABSTRACT Extracting meaningful information on objects varying scale and shape is a challenging task while obtaining distinctive features on small to large size objects to enhance overall object segmentation accuracy from 3D point cloud. To handle this challenge, we propose an attention-based multi-scale atrous convolutional neural network (AMSASeg) for object segmentation from 3D point cloud. Specifically, a backbone network consists of three modules: distinctive atrous spatial pyramid pooling (DASPP), FireModule, and FireDeconv. The DASPP utilizes average pooling operations and atrous convolutions with different sizes to aggregate distinctive information on objects at multiple scales. The FireModule and FireDeconv are responsible to efficiently extract general features. Meanwhile, a spatial attention module (SAM) and channel attention module (CAM) aggregate spatial and semantic information on smaller objects from low-level and high-level layers, respectively. Our network enables to encode multi-scale information and extract distinct feature on overall objects to enhance segmentation performance. We evaluate our method on KITTI dataset. Experimental results demonstrate that the proposed network is effective to improve segmentation performance on small to large objects at real-time speed.

INDEX TERMS Deep learning, convolutional neural network, object segmentation, 3D point cloud, autonomous vehicles.

I. INTRODUCTION

Object segmentation from 3D point cloud is an important task for autonomous vehicles to understand driving environment around the vehicles. Nonetheless, segmenting the objects is a challenging task to achieve competitive segmentation performance in real-time due to sparsity and hugeness of the point cloud. To perform 3D recognition such as segmentation and detection, previous approaches directly processed the point cloud [1] or voxelized the point cloud as pixels in an image [2]. However, these methods suffer from expensive computational cost, as they process all sparse and huge points, thus not applicable for real-time application.

Recently, authors in [3] and [4] conducted object segmentation from 3D point cloud by spherically transforming the point cloud into 2D range-images to apply CNN-based approach. They employed efficient convolutional and deconvolutional operations to extract general features on road-objects inspired from [5]. They achieved significant performance on a large size object (car), however, the performance degraded on small size objects such as pedestrian and cyclist since their approaches were not able to simultaneously extract discriminate features on the small to large objects. To deal with these problems, we adopt a spatial and channel attention mechanism for extracting spatial to semantic-aware features on pedestrian and cyclist. Moreover, we leverage multi-scale atrous convolution to capture contextual information on overall objects at multiple scales. We therefore propose an attention-based multi-scale convolutional neural network (AMSASeg) to improve segmentation performance on all objects such as car, pedestrian, and cyclist.

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

The proposed approach consists of three components: (i) DASPP, (ii) SAM, and (iii) CAM, by leveraging CNN architecture of [3]. The DASPP obtains multi-scale aware features on overall objects. SAM is employed to provide high-level features with spatial details on the smaller objects from low-level features via skip-connections. Whereas, CAM is used to enhance semantic features by modeling channel interdependency between pedestrian and cyclist in high-level features. Further, we reduce the number of down-sampling to maintain spatial information on the smaller size objects and aggregate multi-scale information on small to large objects by replacing the first max pooling layer with the DASPP. In summary, the contribution of our research are as follows:

- We design an attention-based multi-scale atrous CNN to obtain distinctive features on small to large objects for enhanced segmentation accuracy.
- The proposed network with DASPP, SAM (skip-connection) and CAM increases the ability of the network to segment all objects such as car, pedestrian and cyclist by obtaining multi-scale information and enhancing spatial details and the final semantic features.
- The proposed approach based on multi-scale atrous convolution and attention techniques satisfies real-time speed, thus applicable for autonomous driving application.

II. RELATED WORK

A. SEMANTIC SEGMENTATION FROM 3D POINT CLOUD

Before deep learning-based approaches are applied to segmenting 3D point cloud, traditional methods conduct ground removal, clustering of points into objects and classifying the objects using handcrafted features for the point cloud segmentation [6], [7]. Since these approaches depend on handcrafted features and clustering algorithms such as random sample consensus (RANSAC) and agglomerative clustering, the methods based on the handcrafted features require expensive computational cost and rely on random initialization of parameters, therefore unable to be applied to autonomous driving scenario.

With the boost of deep learning-based techniques, many authors [1]–[4] performed 3D recognition such as classification, segmentation and detection based on CNN approaches to overcome limitations of the previous approaches. Authors of PointNet [1] employed multi-layer perceptrons to extract local to global features for classification and segmentation by processing all points of the point cloud. Zhou and Tuzel [2] proposed VoxelNet to detect road-objects from 3D point cloud. The VoxelNet voxelizes the point cloud into 3D voxels as pixels in an image to apply conventional convolution. These approaches achieved significant performance in each perception task. However, processing and voxelizing all points are time-consuming, therefore not suitable for real-time applications such as autonomous navigation.

To address the computational problem, the methods in [3], [4] utilized efficient convolutional and transposed convolutional modules of FireModule and FireDeconv, respectively. The modules are capable to extract general features with low computational cost, which enables each method to operate in real-time. However, the approaches could not extract distinctive features on all objects. Thus, their segmentation performance is insufficient for practical uses.

B. SEMANTIC SEGMENTATION FROM IMAGES

Image semantic segmentation have attracted interests of many researchers due to the wide applicability such as biomedical understanding and autonomous vehicles [8]–[15]. U-Net with an encoder-decoder architecture proposed in [8] successfully segmented medical images and presented a standard for segmentation architecture. The work [9] presented 3D Otsu algorithm based on local contrast for multi-level color image segmentation. By combining the thresholded image and input image, the approach preserved fine details and boundaries with reduced execution time for higher quality segmentation.

Recently, several methods proposed attention techniques to enhance feature representation for more accurate segmentation [10]–[14]. The feature pyramid encoding network [10] was presented for real-time segmentation. It also employed spatial and channel attention blocks to provide spatial and contextual information to features. In [11], squeeze-and-excitation block was proposed aiming to enhance feature representation by recalibrating channel-wise features. Squeeze operation aggregates global spatial information through global average pooling to generate a channel descriptor. To model channel-wise dependency, excitation operation regards the descriptor as channel weights using a gating mechanism to generate more informative features.

Similarly, self-attention modules were leveraged to capture long-range contextual and spatial information in [12]. Moreover, Zhang *et al.* [13] proposed self-attention generative adversarial networks, which modeled long-range dependency for image generation tasks. By aggregating informative features at short to long distance across image regions, the discriminator could find cues at every location of images while the generator could accurately draw images with fine details. Further, authors in [14] adopted a hierarchical attention mechanism for a network to learn to combine multi-scale predictions. Using the mechanism, they overcame limitations of combining better prediction with worse prediction to employ multi-scale inference for improved semantic segmentation results.

Meanwhile, authors in [15] considered pooling operation, which highly effected segmentation performance. They proposed an efficient pooling method to extract more distinctive features, namely distinctive atrous spatial pyramid pooling (DASPP). The DASPP uses average pooling layers with different rate to leverage diverse inputs, which are valuable to extract discriminate features inspired by the theory of receptive fields in human visual ability [16], [17]. The DASPP maximizes the effectiveness of atrous convolution

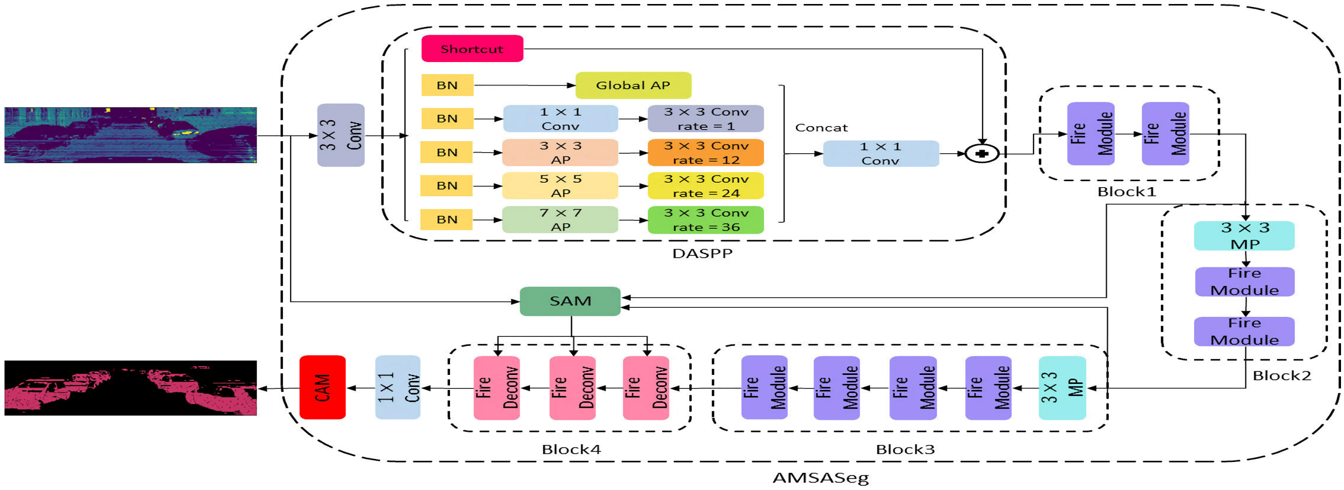


FIGURE 1. An overview of proposed real-time object segmentation approach. DASPP is applied in early part of the proposed network. Blocks 1 to 3 downsample feature maps, whereas block 4 upsamples the features to restore original resolution. Spatial attention module is applied to three skip-connections and channel attention module is applied to a feature map obtained by 1 × 1 convolution. BN: Batch Normalization. AP: Average Pooling. MP: Max Pooling.

with diverse rate to enhance segmentation accuracy based on atrous spatial pyramid pooling (ASPP) [18].

III. PROPOSED APPROACH

An overview of the proposed approach is depicted in Fig. 1. Firstly, we spherically transform 3D point cloud to 2D range-images to efficiently process the point cloud. Secondly, the proposed network employs DASPP, which is used to obtain distinctive features on small to large objects. Thirdly, FireModule and FireDeconv are used to extract general features with low computational cost. Lastly, SAM and CAM are exploited to model spatial information and channel interdependency for discriminant feature representation, respectively. We therefore dedicate following subsections for each component of the AMSASeg to present the proposed approach in detail.

A. SPHERICAL TRANSFORMATION FROM 3D POINT CLOUD TO 2D RANGE-IMAGES

We spherically transform 3D point cloud to 2D range-images in order to efficiently process the point cloud. Each point of the point cloud can be represented as a set of Cartesian coordinate, (x, y, z). Therefore, the formula for spherical transformation can be defined as:

$$\alpha = \arcsin\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right), \hat{\alpha} = \lfloor \frac{\alpha}{\Delta\alpha} \rfloor \tag{1}$$

$$\beta = \arcsin\left(\frac{y}{\sqrt{x^2+y^2}}\right), \hat{\beta} = \lfloor \frac{\beta}{\Delta\beta} \rfloor \tag{2}$$

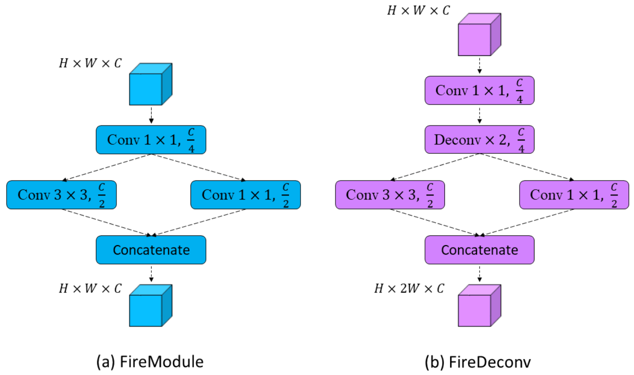
where α and β are azimuth and zenith angles. $\hat{\alpha}$ and $\hat{\beta}$ represent the position of a point on 2D range-images. $\Delta\alpha$ and $\Delta\beta$ are resolutions for discretizing the point cloud. We obtain spherically transformed range-images, which are tensors with a shape of $H \times W \times C$, where H, W and C encode the height, width and channel, respectively. Since the point cloud is generated by a Velodyne HDL-64E LiDAR with 64 vertical channels, the H is 64. Road-objects are annotated based on 3D bounding boxes in a front view area of 90°. We discretize the area into 512 grids, which determine W is 512. In the point cloud, each point contains not only Cartesian coordinates (x, y, z) but also intensity at each point and distance $d = \sqrt{x^2+y^2+z^2}$. Therefore, C of channels of the range-images is 5. We utilize this $64 \times 512 \times 5$ range-images as input data of the proposed network. By using the 2D images based on 2D CNN-based approach, we can efficiently process huge and sparse point cloud to achieve real-time speed.

B. AMSASeg: AN ATTENTION-BASED MULTI-SCALE ATROUS CNN ARCHITECTURE FOR REAL-TIME OBJECT SEGMENTATION FROM 3D POINT CLOUD

An attention-based multi-scale atrous CNN architecture (AMSASeg) has an encoder-decoder architecture as shown in Fig. 1. Our network is able to obtain distinctive features on overall objects such as car, pedestrian, and cyclist by leveraging multi-scale atrous convolution. The proposed network consists of a backbone network [3] and three components: DASPP, SAM, and CAM. The backbone network can extract general features employing FireModule and FireDeconv for efficient convolution and deconvolution. The DASPP aggregates small to large contextual information on all objects. Further, the SAM introduces spatial details from low-level features to high-level features via skip-connections while the CAM enhances semantic representation of the final feature.

1) BACKBONE NETWORK

The most popular backbone network for a segmentation task [8] has an encoder-decoder architecture. The encoder extracts general features by down-sampling input images,



(a) FireModule (b) FireDeconv

FIGURE 2. Illustration of FireModule and FireDeconv. (a) An efficient convolution module (FireModule). (b) An efficient deconvolution module (FireDeconv).

whereas the decoder upsamples the features to generate specified features for point-wise classification with the same resolution as the input images. However, the backbone network employs general convolution and transposed convolution layers for the encoder and decoder, not applicable for embedded application due to expensive computational cost and parameters.

To reduce the cost and parameters, authors in [3] and [4] exploit efficient convolution and deconvolution modules, namely FireModule and FireDeconv, for the encoder and decoder with reduced computational cost. However, the structure of [3] loses spatial information due to multiple max pooling layers, which disseminate spatial information to aggregate contextual information, thus not suitable for accurate segmentation. Moreover, it is essential to aggregate contextually diverse information for enhanced segmentation performance [16], [17]. Based on these considerations, we hence leverage average pooling layers with different rate and multi-scale atrous convolution for the first pooling operation in a backbone network.

For an encoder, a structure of FireModule is illustrated in Fig. 2 (a). The FireModule applies 1 × 1 convolution (squeeze layer) to input features with C channels to reduce the number of channels to $\frac{C}{4}$. And then, 1 × 1 and 3 × 3 convolutional layers (expand layer) are applied in parallel to obtain two feature maps with channel dimension equal to $\frac{C}{2}$. Finally, the two features are concatenated for generating features with C channels.

For a decoder, we utilize FireDeconv, which is as the same as the FireModule except for a transposed convolutional layer between the squeeze layer and expand layer as illustrated in Fig. 2 (b). Moreover, the number of up-sampling is reduced due to reduction of the number of down-sampling, which alleviates computational cost. After recovering the original resolution, we perform 1 × 1 convolution to generate 4 channels where each channel represents each class (background, car, pedestrian, and cyclist).

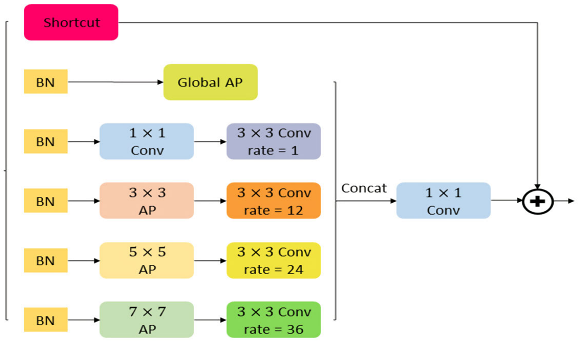


FIGURE 3. A structure of distinctive atrous spatial pyramid pooling (DASPP).

2) DISTINCTIVE ATROUS SPATIAL PYRAMID POOLING

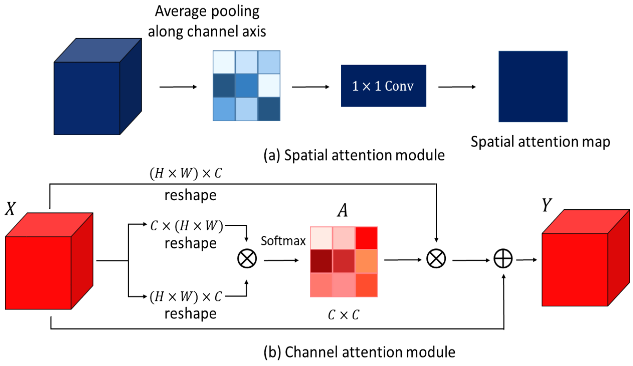
To accurately segment objects regardless of the size, extracting small to large information from input images is essential. Therefore, we employ distinctive atrous spatial pyramid pooling (DASPP) [15] to tackle the problem as shown in Fig. 3. Firstly, we exploit average pooling layers with filter sizes of 3, 5, and 7 in parallel. Next, these pooling layers are connected to atrous convolution layers with dilation rate equal to 12, 24 and 36. In addition, a global average pooling is used to aggregate global contextual information. Shortcut is applied to facilitate information flow. Features from the 5 branches are concatenated and reduced to original numbers by 1 × 1 convolution operation. Finally, features of the shortcut are element-wisely summed with concatenated features. Batch normalization (BN) is used to accelerate the learning of each branch. The DASPP can be formulated as follows:

$$\begin{aligned} Y = I_{pooling}(X) + AT_1(AP_1(X)) + AT_{12}(AP_3(X)) \\ + AT_{24}(AP_5(X)) + AT_{36}(AP_7(X)) \end{aligned} \quad (3)$$

where X and Y encode input and output features of the DASPP. $I_{pooling}(X)$ represents a global average pooling. AT_i denotes atrous convolution with i rate using a 3 × 3 filter. AP_i means average pooling with a $i \times i$ filter. In addition, '+' indicates the concatenation of the features.

3) SPATIAL ATTENTION MODULE

One difficulty in segmenting objects such as pedestrian and cyclist arises due to smallness of appearance. To address this problem, we employ spatial attention module (SAM) [10]. Low-level features contain rich spatial information, however, the information disseminates as passed to high-level layers. To incorporate spatial-aware features, we employ SAM as illustrated in Fig. 4 (a). First, we apply an average pooling to low-level features along the channel axis and apply a 1 × 1 convolutional layer to produce a spatial attention map. The spatial attention map with rich spatial information is transferred via skip-connection to high-level layers for the distinct representation of smaller objects by element-wise multiplication.



(a) Spatial attention module

(b) Channel attention module

FIGURE 4. The structure of attention modules. (a) Spatial attention module (SAM). (b) Channel attention module (CAM).

4) CHANNEL ATTENTION MODULE

Channel attention module (CAM) is employed to enhance semantic feature representation as shown in Fig. 4. (b). In high-level layers, each channel map can be considered as a class-specific response and the semantic response are related with each other [12]. Therefore, the feature representation on specific classes can be enhanced by modeling interdependency among the channels. Firstly, we reshape input features $X \in R^{H\times W\times C}$ to $R^{N\times C}$ by combining width and height dimensions into a dimension N, $(N = H \times W)$. Next, we conduct a matrix multiplication between the reshaped X and the transpose of X. We apply a softmax to get an attention map $\mathrm{A}^{C\times C}$.

$$a_{ji} = \frac{exp\left(X_i X_j\right)}{\sum_{i=1}^{C} exp\left(X_i X_j\right)} \tag{4}$$

where a_{ji} measures the i^{th} channel's impact on the j^{th} channel. Moreover, we multiply the $X^{N\times C}$ by $A^{C\times C}$ and reshape the result to $R^{H\times W\times C}$. Lastly, the result is multiplied by a learnable value (α) and perform an element-wise summation with the original X to obtain the output $Y \in R^{H\times W\times C}$.

$$Y_j = \alpha \sum_{i=1}^{C} \left(a_{ji} X_i\right) + X_j \tag{5}$$

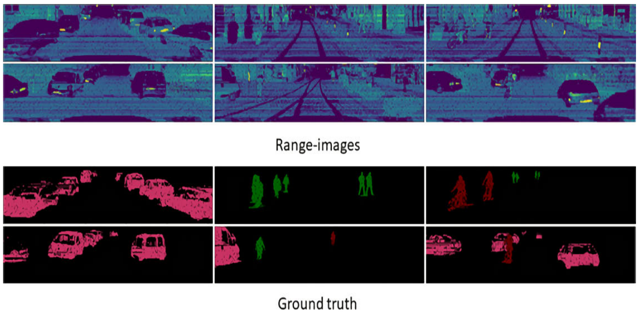
where Y_j represents the final feature map. The equation means that the final feature of each channel is a weighted sum of features of all channels and an original feature, which models semantic interdependency between the channels. Therefore, CAM is applied to the final features generated by 1 × 1 convolutional layer. Finally, we apply softmax activation to the feature maps so as to obtain point-wise prediction.

IV. EXPERIMENTS

In this section, we describe experimental details. This section is divided into four subsections: object scale estimation, implementation details, experimental results, and architecture ablation. In addition, we conduct above-mentioned experiments on KITTI dataset.

A. OBJECT SCALE ESTIMATION

In autonomous driving scenarios, object recognition depends on a scale of road-objects (car, pedestrian, and cyclist). However, the objects can be large or small in various scenes. By approximately estimating the scale of the objects, we can intuitively understand scale-related effectiveness of the proposed network on each object. To estimate the scale of the objects, we select representative scenes, where each object appears large or small, as shown in Fig. 5. In addition, we count points of car, pedestrian, and cyclist to calculate an average scale by dividing object-points with all points in two representative scenes as shown in Table 1. Car occupies 18.5 % of the scenes meanwhile pedestrian and cyclist occupy 3.3 % and 2.8 % of the scenes on average. Therefore, we are able to consider car as a large object and human-related objects (pedestrian and cyclist) as small objects.



Range-images

Ground truth

FIGURE 5. Representative scenes on each object. From left to right, the scenes represent car, pedestrian, and cyclist colored by the pink, green, and brown. The first row shows larger-scale objects meanwhile the second row shows smaller-scale objects.

TABLE 1. Object-points and average scale (%) of each object on representative scenes. All points indicate the number of entire points in two scenes.

	Object-points	All points	Average scale
Car	12,142	65,536	18.5
Pedestrian	2,176	65,536	3.3
Cyclist	1,831	65,536	2.8

B. IMPLEMENTATION DETAILS

1) NETWORK CONFIGURATION

The network configuration of AMSASeg is shown in Table 2. From layer 1 to block 3, we efficiently extract feature maps employing FireModule. To downsample features along the dimension of width, we use a convolutional layer and two max pooling (MP) layer with a 3 × 3 filter. Whereas, block 4 contains three FireDeconv to restore original dimension for point-wise prediction. Further, a 1 × 1 convolutional layer is used to generate feature maps equal to the number of object-classes. Finally, CAM is applied to enhance semantic representation.

2) TRAINING DETAILS

We use spherically transformed 10,848 range-images from 3D point cloud in KITTI dataset [19] to train proposed

TABLE 2. Network configuration of proposed AMSASeg.

Layer	Input	Operation	Output
Layer1	64 × 512 × 5	3 × 3 Conv	64 × 256 × 64
Layer2	64 × 256 × 64	DASPP	64 × 256 × 64
Block1	64 × 256 × 64	FireModule	64 × 256 × 128
	64 × 256 × 128	FireModule	64 × 256 × 128
Block2	64 × 256 × 128	3 × 3 MP	64 × 128 × 128
	64 × 128 × 128	FireModule	64 × 128 × 256
	64 × 128 × 256	FireModule	64 × 128 × 256
Block3	64 × 128 × 256	3 × 3 MP	64 × 64 × 256
	64 × 64 × 256	FireModule	64 × 64 × 348
	64 × 64 × 348	FireModule	64 × 64 × 348
	64 × 64 × 348	FireModule	64 × 64 × 512
	64 × 64 × 512	FireModule	64 × 64 × 512
Block4	64 × 64 × 512	FireDeconv	64 × 128 × 256
	64 × 128 × 256	FireDeconv	64 × 256 × 128
	64 × 256 × 128	FireDeconv	64 × 512 × 64
Layer3	64 × 512 × 64	1 × 1 Conv	64 × 512 × 4
Layer4	64 × 512 × 4	CAM	64 × 512 × 4

network. Further, we split the images into 8,057 training set and 2,791 validation set. We implement our network on Pytorch platform and train the network for around 12 hours by using NVIDIA TITAN Xp GPU. Further, we use focal loss [20] to measure the loss of predicted value. The focal loss is given as follows:

$$FL(p_c) = -(1-p_c)^{\gamma} \log(p_c) \tag{6}$$

where p_c is a predicted probability according to the c class and γ is the focusing parameter equal to 2, which is an optimal value for better training [20].

C. EXPERIMENTAL RESULTS

We evaluate AMSASeg on KITTI dataset. For the evaluation, we leverage two metrics: intersection over union (IoU) and average runtime (AR). IoU is given as follows:

$$IoU_c = \frac{|G_c \cap P_c|}{|G_c \cup P_c|} \tag{7}$$

where G_c and P_c are ground-truth and predicted point of the class-c. Further, the AR of processing each image in validation set is measured. The AR can compare complexity of a network with other networks and measure capability of the network for real-time operation. We quantitatively and qualitatively compare proposed AMSASeg with existing real-time segmentation algorithms: SqueezeSeg [3], PointSeg [4], U-Net [8], and SalsaNet [21]. For a comparison of segmentation performance, we train the algorithms with the same network configuration described in [3], [4], [8], [21]. We set batch size as 8 and learning rate as 0.0001. During the training, focal loss with an optimal focusing parameter of 2 and Adam [22] are employed as the loss function and optimizer, respectively. The quantitative results show that AMSASeg improves segmentation performance on all objects such as car, pedestrian, and cyclist compared to backbone network [3] as shown in Table 3. Moreover, we achieve the highest accuracy on smaller objects (pedestrian and cyclist) with competitive performance on car compared to [4], [8], [21]. Further, we achieve comparable AR at the cost of segmenting overall objects accurately. Furthermore, the qualitative results justify enhanced performance as depicted in Fig. 6. By using yellow boxes, we highlight improved segmentation results.

TABLE 3. Comparison of segmentation performance (*IoU*%) and average runtime (AR) (msec).

	Car	Pedestrian	Cyclist	AR
U-Net [8]	60.9	0.2	2.5	3.0
SqueezeSeg [3]	58.1	1.8	17.8	8.9
PointSeg [4]	66.7	5.1	12.9	10.1
SalsaNet [21]	69.3	4.1	8.7	5.3
AMSASeg (Proposed)	66.7	23.6	22.8	9.6

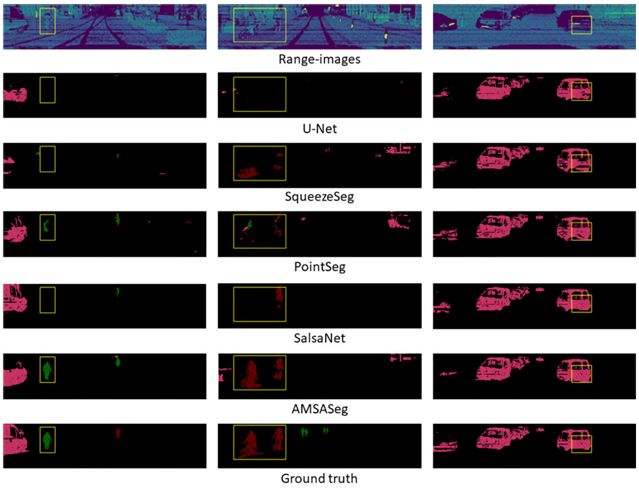


FIGURE 6. Visual comparison of object segmentation. The pink, green, brown, and black indicate car, pedestrian, cyclist, and background.

D. ARCHITECTURE ABLATION

This ablation study is dedicated to verify effectiveness of each component (DASPP, SAM, and CAM) on the backbone network [3] for enhancing object segmentation performance. We apply each component to the network, one by one. Additionally, we compare the effectiveness of DASPP with ASPP [18]. Moreover, we logically consider ablations as shown in Table 4. The ablation study contains six cases.

- SqueezeSeg+SAM: SqueezeSeg with SAM in skip-connections.
- SqueezeSeg+CAM: SqueezeSeg with CAM in the last layer.
- SqueezeSeg(DASPP): SqueezeSeg with DASPP in the first pooling layer.
- SqueezeSeg(DASPP)+CAM: SqueezeSeg with DASPP and CAM.
- SqueezeSeg(ASPP)+CAM+SAM: SqueezeSeg with ASPP, CAM, and SAM.
- SqueezeSeg(DASPP)+CAM+SAM: SqueezeSeg with DASPP, CAM, and SAM. (AMSASeg)

As each component is applied to the backbone network, DASPP and CAM improve segmentation performance by extracting more discriminant features as shown Table 4.

TABLE 4. Comparison of segmentation performance on ablation study (*IoU*%).

	Car	Pedestrian	Cyclist
SqueezeSeg (SS) [3]	58.1	1.8	17.8
SS+SAM	58.5	1.2	11.1
SS+CAM	60.8	2.6	17.0
SS(DASPP)	65.2	10.1	17.9
SS(DASPP)+CAM	68.7	21.6	20.8
SS(ASPP)+CAM+SAM	63.2	8.4	18.7
SS(DASPP)+CAM+SAM	66.7	23.6	22.8

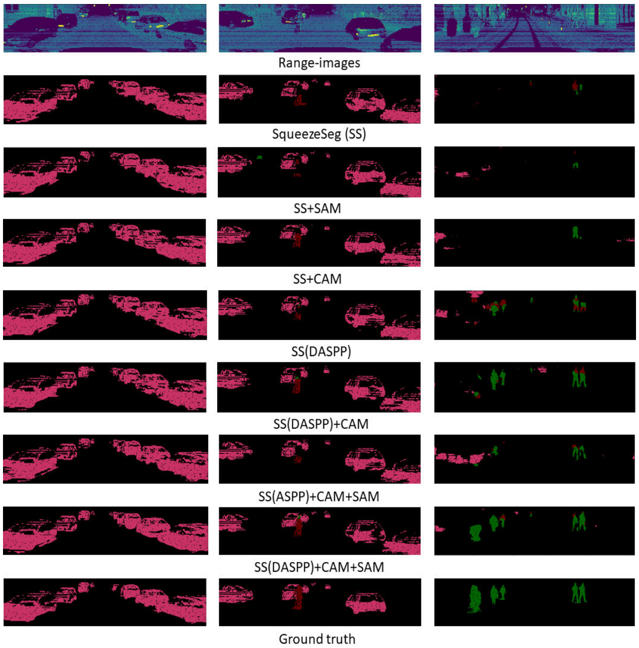


FIGURE 7. Visual comparison of object segmentation on ablation study. The pink, green, brown, and black indicate car, pedestrian, cyclist, and background.

However, SAM degrades the performance on smaller objects such as pedestrian and cyclist since the backbone network contains spatially poor information. We consider combinations with DASPP and CAM or with DASPP, CAM, and SAM because the SAM may play an important role in segmenting the smaller objects at spatially rich features. The combination of DASPP and CAM achieves enhanced accuracy on car, pedestrian, and cyclist. Further, SAM added to the combination also improves segmentation accuracy on pedestrian and cyclist by 2%, despite degradation on car. In addition, DASPP utilizes pooling operations with different sizes based on ASPP to extract more distinctive features. To verify effectiveness of DASPP on accurately segmenting all objects compared with ASPP, we conduct an ablation of backbone network with ASPP, CAM, and SAM. The segmentation performance demonstrates that DASPP is effective for accurate segmentation than ASPP. Therefore, we adopt a backbone network with DASPP, CAM, and SAM as proposed network (AMSASeg) to enhance segmentation performance on all objects such as car, pedestrian, and cyclist. Furthermore, we illustrate segmentation results on all ablations to visually verify effectiveness of each component as shown in Fig. 7.

V. CONCLUSION AND FUTURE WORK

In this work, we proposed an attention-based multi-scale atrous convolutional neural network for real-time object segmentation from 3D point cloud. Multi-scale atrous convolution extracted distinctive features aggregating small to large information on all objects. Furthermore, spatial and channel attention module enhanced feature representation focusing spatial to semantic information on smaller objects. Experimental results showed that the proposed network is effective to enhance segmentation accuracy on all objects with comparable average runtime. In the future, we will focus on extracting object-level information to improve segmentation performance on all objects by using manifold learning.

REFERENCES

[1] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[2] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[3] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1887–1893.

[4] Y. Wang, T. Shi, P. Yun, L. Tai, and M. Liu, "Pointseg: Real-time semantic segmentation based on 3d lidar point cloud," 2018, *arXiv:1807.06288*. [Online]. Available: https://arxiv.org/abs/1807.06288

[5] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[6] M. Himmelsbach, A. Mueller, T. Läuttel, and H.-J. Wäunsche, "Lidar-based 3D object perception," in *Proc. Int. Workshop Cognition Tech. Syst.*, 2008, pp. 1–7.

[7] C. Feng, Y. Taguchi, and V. R. Kamat, "Fast plane extraction in organized point clouds using agglomerative hierarchical clustering," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 6218–6225.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[9] A. K. Bhandari, A. Ghosh, and I. V. Kumar, "A local contrast fusion based 3D otsu algorithm for multilevel image segmentation," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 1, pp. 200–213, Jan. 2020.

[10] M. Liu and H. Yin, "Feature pyramid encoding network for real-time semantic segmentation," 2019, *arXiv:1909.08599*. [Online]. Available: http://arxiv.org/abs/1909.08599

[11] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," 2017, *arXiv:1709.01507*. [Online]. Available: http://arxiv.org/abs/1709.01507

[12] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[13] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[14] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," 2020, *arXiv:2005.10821*. [Online]. Available: http://arxiv.org/abs/2005.10821

[15] G. Dong, Y. Yan, C. Shen, and H. Wang, "Real-time high-performance semantic image segmentation of urban street scenes," *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 19, 2020, doi: 10.1109/TITS.2020.2980426.

[16] B. A. Wandell and J. Winawer, ''Computational neuroimaging and population receptive fields,'' *Trends Cognit. Sci.*, vol. 19, no. 6, pp. 349–357, Jun. 2015.

[17] S. Liu, D. Huang, and Y. Wang, ''Receptive field block net for accurate and fast object detection,'' in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[19] A. Geiger, P. Lenz, and R. Urtasun, ''Are we ready for autonomous driving? The KITTI vision benchmark suite,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[20] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolláar, ''Focal loss for dense object detection,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Feb. 2017, pp. 2980–2988.

[21] E. E. Aksoy, S. Baci, and S. Cavdar, ''SalsaNet: Fast road and vehicle segmentation in LiDAR point clouds for autonomous driving,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 926–932.

[22] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

MOOGAB KIM received the bachelor's degree in electrical engineering from Pusan National University, Pusan, South Korea, in 2019, and the master's degree in electrical engineering with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), South Korea, in 2021. His research interests include deep learning, object recognition, and computer vision for autonomous driving.

NAVEED ILYAS (Student Member, IEEE) received the bachelor's degree in computer engineering and the master's degree in electrical engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2011 and 2015, respectively, and the Ph.D. degree in electrical engineering with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), South Korea, in 2021. He has authored more than ten research articles in international journals and conferences. His research interests include machine learning, deep learning, computer vision, and crowd analysis.

KISEON KIM (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1978 and 1980, respectively, and the Ph.D. degree in electrical engineering systems from the University of Southern California, Los Angeles, CA, USA, in 1987. From 1988 to 1991, he was with Schlumberger, Houston, TX, USA. From 1991 to 1994, he was with the Superconducting Super Collider Laboratory, TX. In 1994, he joined the Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently a Professor. His current research interests include biomedical applications design, wideband digital communications system design, sensor network design, and analysis and implementation both at the physical layer and at the resource management layer. He is a member of The National Academy of Engineering of Korea and a Fellow of the IET. He is also the Senior Editor of the IEEE Sensors Journal.