2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

Real-time Hierarchical Soft Attention-based 3D Object Detection in Point Clouds

Qiuxiao Chen
Utah State University
Logan,Utah 84322-1400
Email: anny.chen@usu.edu

Xiaojun Qi
Utah State University
Logan,Utah 84322-1400
Email: Xiaojun.Qi@usu.edu

Ziqi Song
Utah State University
Logan,Utah 84322-1400
Email: ziqi.song@usu.edu

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956325

***Abstract*—Deep neural network-based 3D object detection in LiDAR point clouds has achieved excellent performance in various applications including autonomous driving and robot vision. However, achieving high accuracy in real-time is paramount in time-critical applications. We propose a real-time Hierarchical Soft Attention Network (HSAN) to employ soft attention in the backbone of the original network to increase the detection accuracy without slowing down its inference speed. The proposed HSAN applies a hierarchical structure on the baseline network to combine features at different scales to obtain rich and fine-grained information and utilizes the characteristic of a layered attention structure to give more attention to the correct regions of target objects. Our proposed system improves the baseline network and achieves comparable detection results in terms of detection accuracy and inference speed when compared with peer state-of-the-art systems on the KITTI validation 3D detection benchmark.**

I. Introduction

Point cloud object detection has recently gained increasing attention in the computer vision community with the popularity of LiDAR sensors and their widespread applications in autonomous driving, robotic vision, and other fields. Deep Neural Network (DNN) based 3D object detection methods have made tremendous progress towards point cloud object detection compared to traditional machine learning-based methods. Two kinds of pioneer work include voxel-based and point-based 3D object detectors. Due to the importance of high efficiency for real-time systems and the space limitation, inefficient point-based methods will not be discussed here.

Voxel-based 3D object detectors, which process voxels transformed from the original point cloud, have been widely used to detect objects due to their fast inference and training speed. They can be broadly classified into two categories: single-shot methods and region proposal-based methods. Single-shot methods [1], [2], [3] incorporate only one stage to directly detect objects, while region proposal-based methods [4] contain two stages, namely, a pre-processing stage and a refinement stage, to detect accurate objects. Although single-shot methods are less accurate than region proposal-based methods, they are more efficient and are widely applied in real-time systems.

Although the aforementioned voxel-based 3D object detectors achieve superior detection performance, a few main issues including the rapid growth of the point cloud size [5] and the loss of geometric information [6] remain unsolved. In order to alleviate these problems, the attention mechanism is introduced in the DNN to focus on important parts of the input data and simplify the point cloud to capture sufficient feature representation [7]. Attention can be roughly divided into two categories: hard attention and soft attention. Due to random sampling around local feature regions, hard attention is non-differentiable and cannot be embedded into the network for convergence learning. On the contrary, soft attention assigns weights to global features or whole image regions. Thus, it is differentiable and can be broadly used in back-propagation methods to detect 3D objects. Classical soft attention, which has three inputs including query, keys, and values, is applied in natural language processing [8]. SA-Det3D [9] then utilizes its mechanism to improve the performance of 3D object detection. Other soft attentions are also widely applied in the computer vision community. For instance, Li et al. [10] employ MaxPool, AvgPool, and tanh activation functions in a soft attention module CBAM to extract edges and detect small objects. Paigwar et al. [11] extend the soft attention mechanism to enable the network to crop smaller regions containing objects of interest, engender the number of points to be processed, and reduce the inference time. However, cropping local areas might lead to error accumulation. Liu et al. [12] introduce TANet consisting of channel, point, and voxel-level soft attentions to capture fine-grained features and improve the capability and robustness of fine-grained pedestrian detection.

In this paper, we propose a Hierarchical Soft Attention Network (HSAN) to address common drawbacks of single-shot 3D object detection methods, namely, lack of geometric information and combined features at different layers. HSAN consists of a Hierarchical Soft Attention (HSA) module and a network skeleton, which can be the network of any voxel-based single-shot methods. In this research, we use two versions of SECOND, namely, SECOND with a small network and SECOND with a large network, as our backbone to build the HSAN, respectively. To improve the inference speed, we also remove 50% of the parameters of the last 3D convolutional layer [13] in both versions of SECOND to simplify the DNN structure. The HSA module aims to help DNN to make clearer and more robust judgments on object recognition by combining features of different scales and paying more attention to correct areas of interest. To this end, the HSA module transforms the input into a weighted feature map that

contains semantic relationships between features from multiple scales to highlight critical features of objects for detection and suppress spatially irrelevant features from the background. Unlike the classical attention mechanism [8], [9] where three inputs and the scaled dot-product operation are involved, the soft attention in the proposed HSA module involves one input and the addition operation to combine low and high-level convolved image feature maps. Furthermore, it differs from other types of soft attention from multiple perspectives. For example, the proposed HSA employs convolution and sigmoid operations as its basic structure while CBAM [10] employs MaxPool, AvgPool, and tanh activation functions as its basic structure. The proposed HSA keeps the whole point cloud to prevent the omission of key information while Attentional-PointNet [11] crops the small regions of interest. The proposed HSA utilizes voxel-level attention on different scales of features to capture fine-grained features while TANet [12] involves a triple attention to capture fine-grained features.

Our contributions are summarized as follows:

- Proposing a HSAN with a HSA module to incorporate features at different scales to improve detection accuracy of the SECOND network.
- Utilizing the features generated from the HSA module to learn and find the most important locations to focus on and filter out the irrelevant parts of the input point cloud.
- Improving the baseline network and achieving similar accuracy and inference speed comparing with one-stage state-of-the-art systems on the KITTI validation dataset.
- Deploying the HSA model in other voxel-based networks to improve 3D object detection performance.

II. HIERARCHICAL SOFT ATTENTION NETWORK

In this section, we present the overall structure of the proposed HSAN and the details of the HSA module.

A. Overview

Fig. 1 shows the overall architecture of the proposed HSAN, which uses the widely used small SECOND network as its backbone to maintain the detection accuracy with a faster speed. SECOND is an effective 3D object detection system achieving a high accuracy with a fast speed. It first divides the input point cloud data into voxels of the same size for pre-processing. It then converts a certain number of points in each voxel into a vector of voxel features and coordinates to maintain geometric and spatial information. These vectors are next sent to 3D convolution blocks to expand their receptive fields. The 3D voxel features are reshaped into a Bird's-Eye-View (BEV) shape and sent to a 2D convolution block to obtain 2D features. Finally, the 2D features are put into box regression and classification branches to localize and classify detected objects, respectively.

To improve the efficiency and accuracy of SECOND, we use its small network as backbone and modify this simple network structure from two perspectives. First, we cut 50% of the parameters of the last layer of 3D sparse convolutional layers to simplify the 3D convolutional block, speed up the training, and maintain the efficiency. Second, we include the HSA module in 2D convolutional blocks to improve the accuracy of SECOND. Specifically, the HSA module learns the most important positions in the point cloud data, filters out the irrelevant parts, and combines feature maps of different scales to more accurately represent objects.

The section below the block diagram of Fig. 1 presents the details of employing the proposed HSA module at two places in the 2D backbone network, which consists of two layers of encoding and decoding blocks. The encoding block at each layer contains six convolutional layers and the decoding block at each layer contains one deconvolutional layer. We use X to denote SECOND's 2D BEV features, which is the input of the 2D backbone network. We first employ the soft-attention module on X to calculate semantic relationships among voxels and obtain its weighted feature map Y_1 at the first layer. This weighted feature map Y_1 then goes through the encoding block of the first layer to obtain EY_1, whose channel number is reduced by half from Ch to $Ch/2$. EY_1 goes through two branches: one branch is to go through the decoding block of the first layer to obtain DY_1, which has the same dimension as X; the other branch is to employ the soft-attention module on EY_1 to calculate semantic relationships among voxels and obtain its weighted feature map Y_2 at the second layer. This weighted feature map Y_2 goes through the encoding block of the second layer to obtain EY_2, whose channel is doubled from $Ch/2$ to Ch and whose height and width are reduced from H to $H/2$ and from W to $W/2$, respectively. EY_2 finally goes through the decoding block of the second layer to obtain DY_2, which has the same dimension as X. Lastly, DY_1 and DY_2 are concatenated together as the output of the 2D backbone network. This hierarchical structure combines features at different scales, enhances semantic information, and broadens the receptive field.

B. Hierarchical Soft Attention (HSA) Module

The HSA module is composed of two soft attention blocks. Each block has the same soft attention mechanism, which forms two control gate mask branches: one going through two convolutional layers and one going through one convolutional layer. Fig. 2 demonstrates the proposed soft attention block. Here, we use W to represent the 2D convolutional layer weights and the subscript of W to represent specific convolutional layers. Specifically, W_X, W_{XY}, and W_Y represent convolutional layer weights of the convolutional layer connecting feature X and its convolved feature (i.e., X_{conv}), features X and Y, and feature Y and its convolved feature (i.e., Y_{conv}), respectively. For the input of the 2D convolutional networks (e.g., a given BEV feature map X), we use two branches of 1×1 convolutions W_X and W_{XY} to transform X into two new feature maps X_{conv} and Y with the same dimension, respectively. We then employ another 1×1 convolution W_Y to transform Y into Y_{conv} with the same dimensions. We combine X_{conv} and Y_{conv} via the elementwise addition and apply another 1×1 convolution W_S, which represents convolutional layer weights of the convolutional layer after the

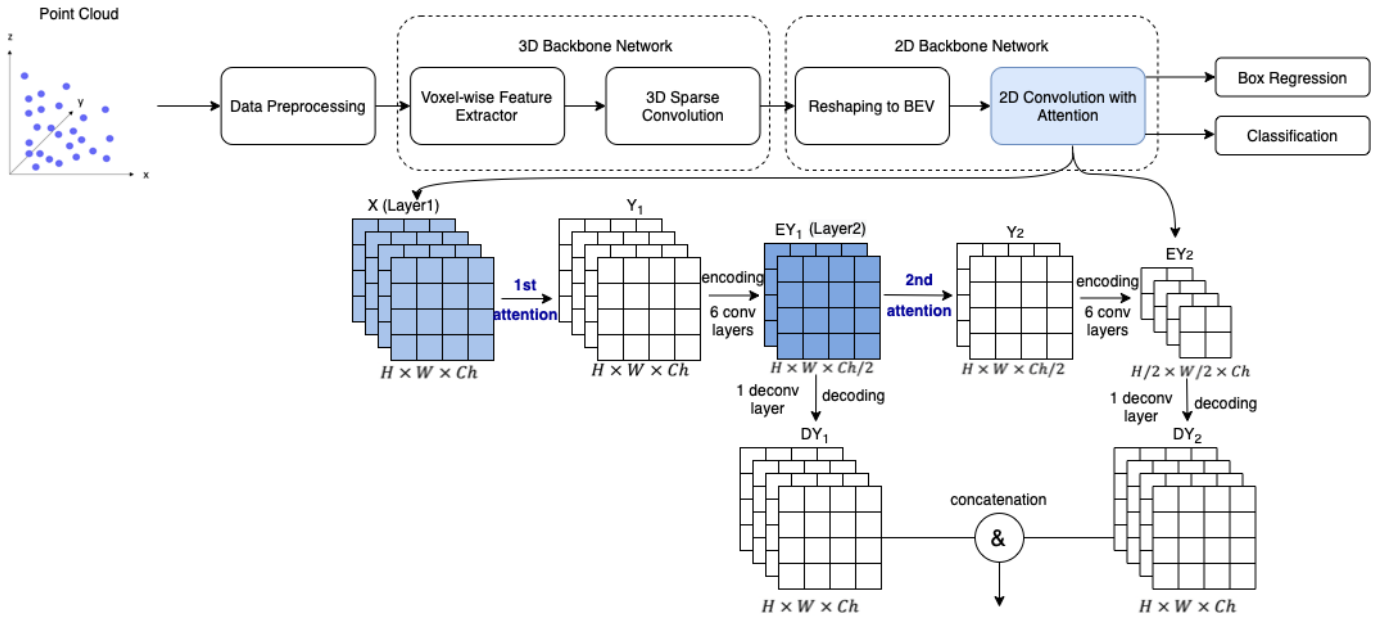

Fig. 1. The overall architecture of the proposed HSAN. The upper part demonstrates its block diagram, where HSA modules reside in the 2D backbone network shown in blue shade. The lower part presents the flowchart of employing the HSA module in the 2D backbone network.

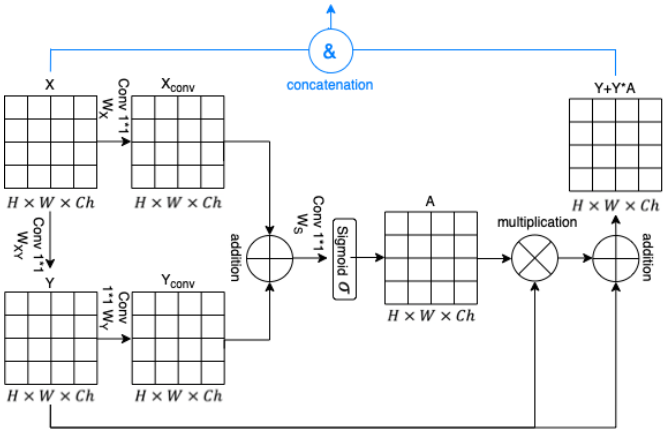

Fig. 2. Illustration of the proposed soft attention block.

addition operation, to transform the combined feature map into a new feature map. The sigmoid operation is then employed on the new feature map to normalize it into a new weighted feature map A. The feature map Y is elementwisely multiplied with A, added with itself and concatenated with input X to obtain the final weighted feature map.

The proposed HSA module allows the control gate to perform pixel-to-pixel modeling (i.e., voxel-wise addition or element-wise addition of local features) to make the focused and related resources be assigned to the most intrinsic and informative areas. In other words, the control gate branches function as a masking mechanism to recalibrate local features from multiple scales and selectively strengthen valuable and informative areas and suppress useless and non-informative features such as noise and background. As a result, the values in the attention mask represent the weights of corresponding pixels on the original feature maps of point clouds, which make the attention mask more suitable for pixel-wise classification than global pooling. In summary, the proposed HSA module not only selects the most intrinsic and discriminative features toward the classification objective in the feed-forward process but also prevents the updating of parameters with incorrect gradients during backpropagation [14]. It further makes our network more expressive, robust, and informative.

We also design two variant soft attentions for comparison. The variant 1 intuitively applies the soft attention on the original input feature map X to enhance target objects and filter out irrelevant areas in X. Its final weighted feature map is obtained by adding X and its multiplication with the weighted feature map A. The variant 2 applies the soft attention on the high-level feature map Y to enhance target objects, filter out irrelevant areas in Y, and learn more deformations of target objects. Its final weighted feature map is obtained by adding Y and its multiplication with A. The proposed HSAN takes the advantage of both variants to simultaneously consider both low and high-level feature maps, which contain rich and fine context information, by concatenating X with the final feature map obtained from the variant 2. These three soft attentions are employed at the same two places in the 2D backbone network to construct their counterpart HSAs.

C. Design and Mathematical Formulation of HSA

We treat X and Y_{conv} shown in Fig. 2 as low-level features of layer L and high-level features of layer $L+2$ in the encoding stage, respectively. X_{conv} and Y are considered as high-level features of layer $L+1$. To improve the network sensitivity, we design our soft attention block based on additive

attention since it experimentally has higher accuracy than multiplicative attention [15]. To this end, we calculate the soft attention mask A at layer L by performing an elementwise addition operation between X_{conv} and Y_{conv} to learn critical features of objects. This attention mask A integrates the relationship between features from multiple scales or layers at different regions, focuses on useful regions, and indicates the significance of different regions. We then perform an elementwise multiplication operation between A and Y to identify relevant regions containing objects. Finally, we employ the addition [14] to retain original features, so the final output of the soft attention block is defined as $Y + Y \circ A$, where $\circ$ represents the elementwise multiplication. The following equation summarizes the steps to compute the soft attention mask A at layer L, where concatenation is not involved:

$$A^L = Sigmoid(W_s \star (W_X \star X^L + W_Y \star W_{XY} \star X^L))$$

Here, $X^L \in R^{H^L * W^L * Ch^L}$ are features at layer L with H, W, and Ch being its respective height, width, and channel number, $\star$ represents the conventional convolution operation, and W_X, W_Y, W_{XY}, and $W_S \in R^{Ch^L * Ch^L * k * k}$ are convolutional filters, whose kernel size is $k \times k$ (i.e., $k = 1$), used at different layers L to generate features at the next layers. The proposed soft attention block functions as a feature selector to automatically augment useful structure features during the forward process by replacing X^L with the weighted attention features $\widehat{Y}^L$ concatenating with X^L. In summary, the final output of the variant 1, the variant 2, and the proposed soft attention block is $X + X \circ A$, $Y + Y \circ A$, and concatenation of X and $Y + Y \circ A$, respectively.

The proposed soft attention block and its two variants can be applied to multiple layers to form an HSA module. In our implementation, we apply it to two layers, as shown in Fig. 1, to build an HSA module to integrate the relationship between features from multiple scales. As a result, the HSA enables the DNN to focus on useful areas and salient features and be more robust against big objects such as cars and cyclists.

III. Experiments

We extensively evaluate the proposed one-stage HSAN on the KITTI dataset [16], which is a novel challenging real-world computer vision benchmark captured by driving around the mid-size city Karlsruhe, its rural areas, and its highways. The dataset contains images, videos, 3D point clouds, and their Global Positioning System (GPS) locations. In this research, we focus on the KITTI 3D point cloud dataset, which has 7,481 training and 7,518 testing point clouds in three categories (e.g., cars, pedestrians, and cyclists). Each category has point clouds with three difficulty levels including easy, moderate, and hard based on bounding box height, occlusion, and truncation levels. The height of the bounding box of objects at easy, moderate, and hard difficulty levels is at least 40, 25, and 25 pixels, respectively. The occlusion of the objects at easy, moderate, and hard difficulty levels is fully visible, partly occluded, and hard to see, respectively. The truncated percentage of objects at easy, moderate, and hard difficulty levels is at most 15%, 30%, and 50%, respectively. Objects that do not satisfy the above requirements (e.g., 6,473 cars, 170 pedestrians, and 165 cyclists) are not used for training and validation. In total, the KITTI dataset contains 17,823 easy objects including 13,067 cars, 3,694 pedestrians, and 1,062 cyclists, 9,547 moderate objects including 8,602 cars, 563 pedestrians, and 382 cyclists, and 678 hard objects including 600 cars, 60 pedestrians, and 18 cyclists. We divide the training data into training and validation split with 3,712 and 3,769 point clouds, respectively.

We employ Average Precision (AP) over 11 recall positions as the metric to evaluate the 3D object detection results in the validation split. Different IoU thresholds are empirically determined by other researchers to compute AP. An IoU of 0.7 is commonly used for cars and an IoU of 0.5 is commonly used for cyclists and pedestrians. The leaderboard rank is based on the results of the dataset at the moderate level.

A. Experiment I

Table I lists the AP of the proposed HSAN with a small SECOND network, the proposed HSAN with a large SECOND network, and ten peer one-stage voxel-based 3D object detectors, namely, SECOND with a small network, SECOND with a large network, TANet [12], Voxel-FPN [3], SA-SSD [17], SE-SSD [18], CenterNet3D-SL1 [19], Pointpillars [20], SCNet [21], and AFDet [22], on the KITTI car validation dataset. It shows that HSAN with a large SECOND network achieves better car detection results than HSAN with a small SECOND network at three difficulty levels. It ranks the best in detecting cars at easy and hard levels and the third in detecting cars at the moderate level. Table II lists the AP of HSAN with a small SECOND network, HSAN with a large SECOND network, and three peer 3D object detectors (e.g., VoxelNet [1], TANet [12], and Voxel-FPN [3]) on the KITTI cyclist validation dataset. Since seven peer detectors listed in Table I do not provide the cyclist AP on the KITTI validation dataset, we only compare the detection results of three peer systems in Table II. This table shows that HSAN with a large SECOND network achieves the best performance on cyclists at moderate and hard levels and the second best performance at the easy level. In the following, we will compare the car detection performance of HSAN and several detectors in terms of detection accuracy, detection speed, and ablation studies.

1) Comparison with SE-SSD and SA-SSD, two best car detectors at the moderate level in terms of accuracy: Table I shows that SE-SSD, SA-SSD, HSAN with a large SECOND network, and HSAN with a small SECOND network rank the top four 3D car detectors at the moderate level with detection rates of 85.71%, 79.79%, 78.77%, and 78.31%, respectively. HSAN with a large SECOND network has the highest 3D car detection rate of 88.98% and 77.27% for easy and hard levels, respectively. It improves the second best car detectors SA-SSD at the easy level by 0.26% and HSAN with a small SECOND network at the hard level by 1.06%. However, SE-SSD has a more complex training process than HSAN. Its training process iteratively updates the teacher and student SSDs,

TABLE I
COMPARISON OF AP(%) OF 12 METHODS ON CARS.

3D Detector	Easy	Moderate	Hard	FPS
SECOND (small network)	85.5	75.04	68.78	40
SECOND (large network)	87.43	76.48	69.1	25
TANet	88.21	77.85	75.62	29
Voxel-FPN	88.27	77.86	75.84	***50***
SA-SSD	***88.75***	***79.79***	74.16	25
SE-SSD	N/A	**85.71**	N/A	32
CenterNet3D-SL1	87.92	76.84	75.74	25
Pointpillars	86.13	77.03	72.43	**62**
SCNet	87.83	77.77	75.97	25
AFDet	85.68	75.57	69.31	N/A
HSAN (Proposed, small network)	88.5	78.31	***76.46***	40
HSAN (Proposed, large network)	**88.98**	78.77	**77.27**	25

TABLE II
COMPARISON OF AP(%) OF FIVE METHODS ON CYCLISTS.

Network	**Easy**	**Moderate**	**Hard**
VoxelNet	67.17	47.65	45.11
TANet	**85.98**	64.95	60.40
Voxel-FPN	68.77	61.86	56.40
HSAN (Proposed, small network)	79.58	***67.28***	***63.35***
HSAN (Proposed, large network)	***83.59***	**68.46**	**63.66**

while HSAN's training process is straightforward. SA-SSD has a more complex network structure than HSAN since it maintains an auxiliary network and involves a partial-sensitive deformation operation. Overall, the proposed HSAN with a large SECOND network achieves the best detection accuracy for cars at easy and hard levels and the third best detection accuracy for cars at the moderate level when compared with ten state-of-the-art networks.

2) Comparison with Pointpillars and Voxel-FPN, two best car detectors in terms of speed: Table I shows that Pointpillars, Voxel-FPN, and HSAN with a small SECOND network are three fastest detectors with an inference speed of 62, 50, and 40 FPS, respectively. Pointpillars treats voxels in same (x, y) coordinates as a whole to accelerate speed. Voxel-FPN uses the multi-scale voxel features fusion module to accelerate speed. However, both lead to information loss, which degrades their detection accuracy. HSAN with a small SECOND network improves the detection accuracy of Pointpillars by 2.75%, 1.66%, and 5.56% and Voxel-FPN by 0.26%, 0.58% and 0.82% for cars at easy, moderate, and hard levels, respectively. Overall, HSAN with a small SECOND network is an excellent trade-off between performance and efficiency.

3) Ablation Studies: Comparison with their corresponding baseline methods, SECOND with a small network and SECOND with a large network: HSAN improves the detection accuracy of its corresponding baseline SECOND. Specifically, HSAN with a large SECOND network improves the large SECOND by 1.77%, 2.99%, and 11.82% and HSAN with a small SECOND network improves the small SECOND by 3.51%, 4.35%, and 11.17% to detect cars at easy, moderate, and hard levels, respectively. The improvement is mainly achieved by adding the HSA model to its baseline SECOND. First, HSA considers input as low-level features and processes them with convolutional layers to learn high-level features for abstract semantics. Second, it combines input with high-level convolved features to capture both abstract semantics and detailed information to represent an object. HSAN and SECOND have a similar network structure and the same settings including a learning rate of 0.003, the Adam one cycle optimizer, and the loss function of SmoothL$_1$. However, HSAN's network parameters (e.g., 4.5 and 9.6 millions respectively for the small and large SECOND network) are 18.4% and 24.0% less than their baseline SECOND (e.g., 5.33 and 11.9 millions respectively for the small and large network) due to the removal of 50% of the parameters of the last 3D convolutional layer. As a result, HSAN is a little faster than SECOND.

In summary, our extensive experimental results from three perspectives demonstrate that HSAN has comparable car detection results and the best cyclist detection results when compared with state-of-the-art peer methods. In addition, HSAN significantly improves the detection accuracy of its corresponding baseline due to the addition of its HSA.

B. Experiment II

To verify the effectiveness of the proposed HSAN, we summarize 3D and 2D car detection results of the proposed HSAN, its two variants, and SECOND at three levels in terms of AP in Table III. All four systems are implemented on a small SECOND network. It shows that both HSAN and its two variants achieve better car detection accuracy than SECOND and HSAN outperforms its two variants at three levels.

TABLE III
COMPARISON OF DETECTION RESULTS OF FOUR METHODS.

Network	**3D**			**BEV (2D)**		
	Easy	***Moderate***	***Hard***	***Easy***	***Moderate***	***Hard***
SECOND	85.50	75.04	68.78	*89.79*	86.20	79.55
Variant 1	87.68	77.51	75.19	89.93	87.43	84.99
Variant 2	88.12	77.80	76.10	89.88	87.55	85.49
Proposed	**88.50**	**78.31**	**76.46**	**90.23**	**87.80**	**85.83**

Fig. 3 and Fig. 4 demonstrate three sample 3D car detection results and two sample 3D cyclist detection results of the above four systems, respectively, where ground truths are shown in green bounding boxes and detection results are shown in red bounding boxes. In Fig. 3, the first row presents a scenario that SECOND detects three cars with one of them being a false positive. Two variants and HSAN accurately detect two true cars without any false positives. The second row presents a scenario that HSAN successfully detects one true car without false positives and SECOND and two variants detect one true car and some false positives. Specifically, SECOND and two variants detect three and two wrong cars, respectively. The last row presents a scenario that SECOND detects one true car and one false positive car and fails to detect one true car. The variant 1 and HSAN obtain the same detection results as SECOND except that they do not have false positives. The variant 2 correctly detects both true cars without false positives. In Fig. 4, the upper row shows that SECOND detects

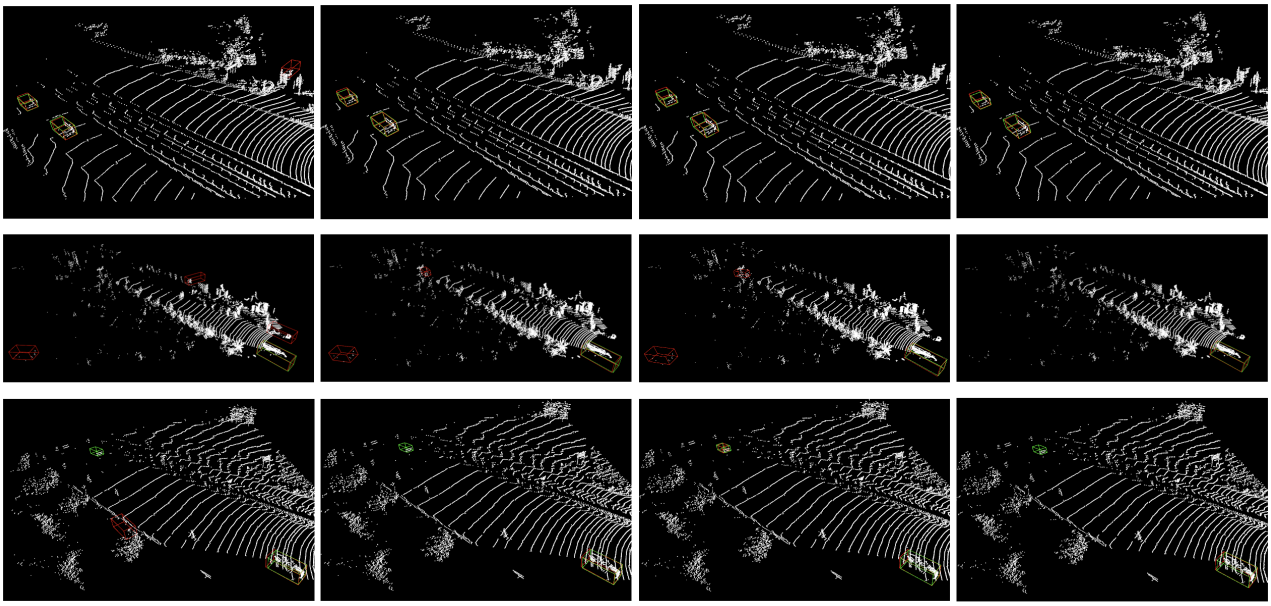

Fig. 3. Three sample car detection results of four methods (from left to right): SECOND, the variant 1, the variant 2, and the proposed HSAN.

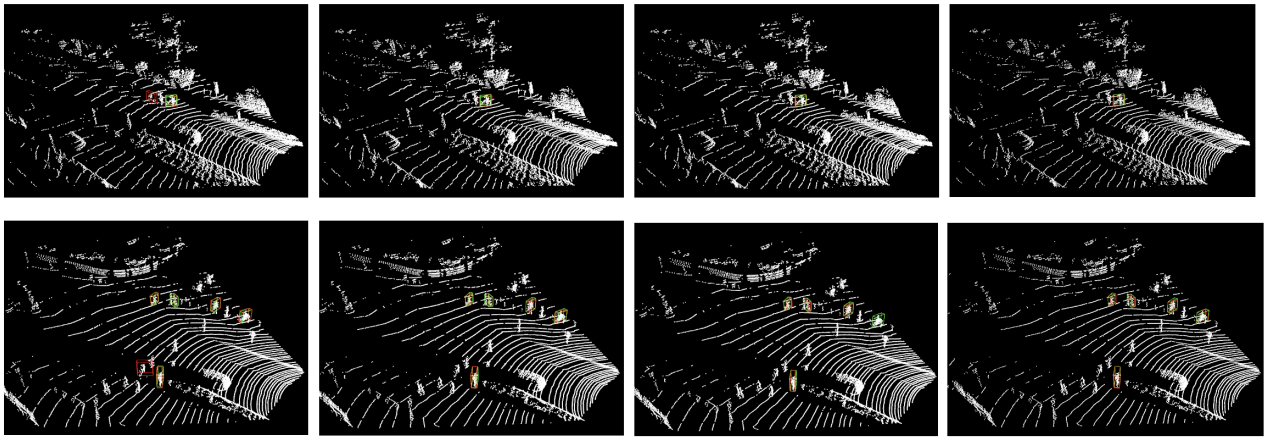

Fig. 4. Two sample cyclist detection results of four methods (from left to right): SECOND, the variant 1, the variant 2, and the proposed HSAN.

two cyclists, while one of them is a false positive. HSAN and its two variants detect one cyclist target object precisely. The lower row shows that SECOND, the variant 1, and HSAN successfully detect all five ground truth cyclists. But SECOND detects one false positive cyclist. The variant 2 successfully detects four truth cyclists and misses one truth cyclist.

These qualitative results show that HSAN and its variants outperform their counterpart SECOND in detecting cars and cyclists. Overall, HSAN detects the fewest false positives and achieves the same true object detection as its variants.

IV. CONCLUSION

In this paper, we propose three soft attentions and employ them in the 2D backbone network of SECOND to build a HSAN and its two variants. The hierarchical structure of HSAN combines features of multiple scales to obtain rich and fine information to capture target object features. It also helps the network focus on real object areas and filter out irrelevant areas in point clouds at the low-level map (variant 1), the high-level map (variant 2), and both low and high-level maps (the proposed). Our extensive experiments on the KITTI validation dataset confirm HSAN and its variants improve their counterpart SECOND by at least 2.55%, 3.29%, and 9.32% for cars at easy, moderate, and hard levels, respectively. HSAN with a large SECOND network achieves the best detection accuracy for cars at easy and hard levels and the third best accuracy for cars at the moderate level and HSAN with a small SECOND network is the third fastest car detector when compared with ten peer networks. HSAN with a large SECOND achieves the best performance on cyclists at moderate and hard levels and the second best performance at the easy level when comparing with three peer networks.

REFERENCES

[1] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

[2] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[3] Hongwu Kuang, Bei Wang, Jianping An, Ming Zhang, and Zehan Zhang. Voxel-fpn: Multi-scale voxel feature aggregation for 3d object detection from lidar point clouds. *Sensors*, 20(3):704, 2020.

[4] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[5] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020.

[6] Can Qin, Haoxuan You, Lichen Wang, C-C Jay Kuo, and Yun Fu. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. *arXiv preprint arXiv:1911.02744*, 2019.

[7] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[9] Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. Sa-det3d: Self-attention based context-aware 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3022–3031, 2021.

[10] Yiran Li, Han Xie, and Hyunchul Shin. 3d object detection using frustums and attention modules for images and point clouds. *Signals*, 2(1):98–107, 2021.

[11] Anshul Paigwar, Ozgur Erkent, Christian Wolf, and Christian Laugier. Attentional pointnet for 3d-object detection in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 1297–1306, 2019.

[12] Zhe Liu, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. Tanet: Robust 3d object detection from point clouds with triple attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11677–11684, 2020.

[13] Yutian Wu and Harutoshi Ogai. Realtime single-shot refinement neural network for 3d obejct detection from lidar point cloud. In *Proceedings of the 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 332–337, 2020.

[14] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[15] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

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

[17] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11873–11882, 2020.

[18] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14494–14503, 2021.

[19] Guojun Wang, Jian Wu, Bin Tian, Siyu Teng, Long Chen, and Dongpu Cao. Centernet3d: An anchor free object detector for point cloud. *arXiv preprint arXiv:2007.07214*, 2020.

[20] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[21] Zhiyu Wang, Hao Fu, Li Wang, Liang Xiao, and Bin Dai. Scnet: Subdivision coding network for object detection based on 3d point cloud. *IEEE Access*, 7:120449–120462, 2019.

[22] Runzhou Ge, Zhuangzhuang Ding, Yihan Hu, Yu Wang, Sijia Chen, Li Huang, and Yuan Li. Afdet: Anchor free one stage 3d object detection. *arXiv preprint arXiv:2006.12671*, 2020.