SeSame: Simple, Easy 3D Object Detection with Point-Wise Semantics Supplementary Material

Hayeon O[1][0009−0007−8197−8017], Chanuk Yang[2][0000−0002−2065−4031], and
Kunsoo Huh[2][0000−0002−7179−7841]

[1] Department of Automotive Engineering (Automotive-Computer Convergence)
[2] Department of Automotive Engineering
{gkdus9595,ych901208,khuh2}@hanyang.ac.kr
Hanyang University, Seoul, 04763, Republic of Korea

1 Figure for Ablation Study : Reference and Ours

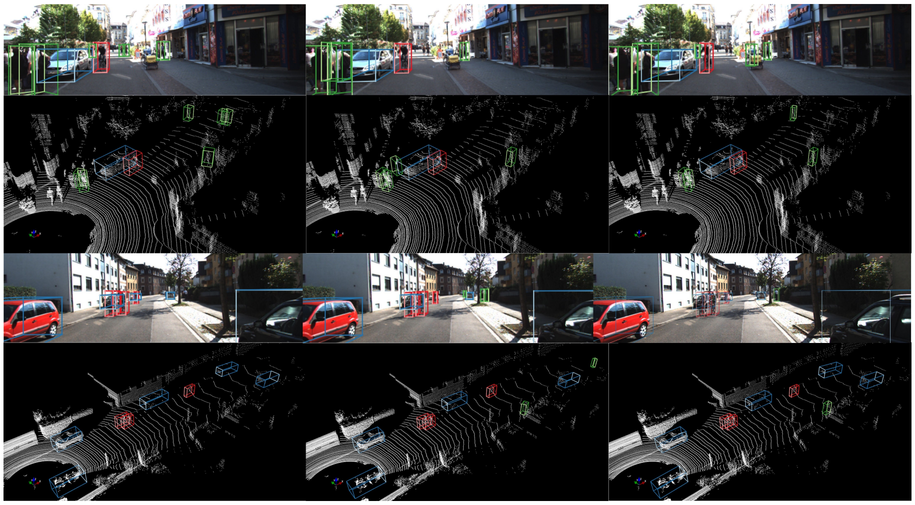

Fig. 1. Qualitative results on the KITTI validation set. There are two scenes. For each scene, the results of ground truth, PointPainting [5], and SeSame+point are shown from leftmost to rightmost.

PointPainting [5], which is our reference model, has only BEV detection result on test split. And thanks to open code, we can get qualitative result of [5] on val set. In Fig. 1, [5] has false positive on pedestrian in upper scene and false negative for car in lower scene. And this supports that pixel-wise semantic features projected onto the LiDAR plane can become misaligned, resulting in false positives. Specifically, these inaccuracies may arise when objects are located behind others, indicating that occlusion contributes to the improper projection of semantic features. In contrast, our approach utilizing point-wise semantic segmentation shows fewer false positives and false negatives when detecting distant, occluded, and densely packed objects.

2 Table for Ablation Study : score vs. label

Method	modality	car (IoU=0.7)			pedestrian (IoU=0.5)			cyclist (IoU=0.5)		
		easy	mod	hard	easy	mod	hard	easy	mod	hard
SeSame+Point w/ score	LiDAR (point)	88.63	**78.55**	**77.66**	**62.84**	**58.53**	**51.53**	86.18	71.88	66.22
SeSame+Point w/ label	LiDAR (point)	**88.76**	78.35	77.43	62.83	55.30	51.35	**87.80**	**72.74**	**67.00**
SeSame+Voxel w/ score	LiDAR (voxel)	**88.47**	78.44	76.96	54.35	50.76	45.70	81.50	**67.06**	**63.81**
SeSame+Voxel w/ label	LiDAR (voxel)	88.18	**78.52**	**77.27**	**56.31**	**51.94**	**46.88**	**82.01**	65.97	61.26
SeSame+Pillar w/ score	LiDAR (pillar)	**86.98**	**77.17**	**75.42**	**54.66**	**49.90**	**45.70**	**80.40**	**63.09**	**60.57**
SeSame+Pillar w/ label	LiDAR (pillar)	85.65	75.94	73.85	53.43	48.43	44.69	77.85	61.45	58.32

Table 1. Ablation Study : score vs. label on KITTI 3D detection val split

Method	modality	car (IoU=0.7)			pedestrian (IoU=0.5)			cyclist (IoU=0.5)		
		easy	mod	hard	easy	mod	hard	easy	mod	hard
SeSame+Point w/ score	LiDAR (point)	**89.97 (89.9716)**	**87.45**	**86.24**	68.26	**61.74**	**54.54**	87.79	73.49	67.99
SeSame+Point w/ label	LiDAR (point)	89.97 (89.9698)	87.17	86.07	**69.22**	61.05	54.39	**88.57**	**74.80**	**68.45**
SeSame+Voxel w/ score	LiDAR (voxel)	**90.06**	**87.90**	**86.46**	59.24	53.69	50.76	86.41	70.47	66.23
SeSame+Voxel w/ label	LiDAR (voxel)	89.86	87.73	86.14	**60.71**	**55.33**	**51.71**	**89.47**	**70.49**	**66.38**
SeSame+Pillar w/ score	LiDAR (pillar)	**89.60**	**86.96**	**84.62**	**60.17**	**54.76**	**51.01**	**82.79**	**67.66**	**62.47**
SeSame+Pillar w/ label	LiDAR (pillar)	88.90	86.44	84.17	58.81	54.21	49.83	81.26	65.46	61.39

Table 2. Ablation Study : score vs. label on KITTI BEV detection val split

Method	modality	car (IoU=0.7)			pedestrian (IoU=0.5)			cyclist (IoU=0.5)		
		easy	mod	hard	easy	mod	hard	easy	mod	hard
SeSame+Point w/ score	LiDAR (point)	74.30	56.92	48.14	31.13	23.33	20.07	9.99	8.31	6.87
SeSame+Point w/ label	LiDAR (point)	**85.25**	**76.83**	**71.60**	**42.29**	**35.34**	**33.02**	**69.55**	**54.56**	**48.34**
SeSame+Voxel w/ score	LiDAR (voxel)	61.57	47.14	41.06	34.14	28.26	26.15	53.37	40.05	35.71
SeSame+Voxel w/ label	LiDAR (voxel)	**81.51**	**75.05**	**70.53**	**46.53**	**37.37**	**33.56**	**70.97**	**54.36**	**48.66**
SeSame+Pillar w/ score	LiDAR (pillar)	82.32	73.15	66.64	33.87	27.23	25.27	11.47	14.29	12.57
SeSame+Pillar w/ label	LiDAR (pillar)	**83.88**	**73.85**	**68.65**	**37.61**	**31.00**	**28.86**	**64.55**	**51.74**	**46.13**

Table 3. Ablation Study : score vs. label on KITTI 3D detection test split

Method	modality	car (IoU=0.7)			pedestrian (IoU=0.5)			cyclist (IoU=0.5)		
		easy	mod	hard	easy	mod	hard	easy	mod	hard
SeSame+Point w/ score	LiDAR (point)	83.44	67.18	57.68	33.98	25.79	22.50	10.65	8.90	7.68
SeSame+Point w/ label	LiDAR (point)	**90.84**	**87.49**	**83.77**	**48.25**	**41.22**	**39.18**	**75.73**	**61.70**	**55.27**
SeSame+Voxel w/ score	LiDAR (voxel)	71.98	63.36	57.52	39.42	33.76	31.31	58.94	45.61	40.68
SeSame+Voxel w/ label	LiDAR (voxel)	**89.86**	**85.62**	**80.95**	**50.12**	**41.59**	**37.79**	**76.95**	**59.36**	**53.14**
SeSame+Pillar w/ score	LiDAR (pillar)	90.43	86.11	81.38	39.11	32.78	30.87	15.92	19.53	17.61
SeSame+Pillar w/ label	LiDAR (pillar)	**90.61**	**86.88**	**81.93**	**44.21**	**37.31**	**35.17**	**72.22**	**60.21**	**53.67**

Table 4. Ablation Study : score vs. label on KITTI BEV detection test split

When preprocessing segmented point cloud in the proposed method, it can be divided into two cases: with softmax applied and without softmax applied, as "w/ score" and "w/ label" respectively. Tab. 1 and 2. compare these two cases on the validation split, showing similar performance. However, for the comparisons on the test split presented in Tab. 3 and 4, the "w/ score" case does not exhibit performance similar to that of the validation split, unlike the "w/ label" case, and even worse.

3 Links for KITTI 3D object detection benchmark

You can find the results for the val split and test split of the KITTI 3D object detection dataset in the github link referenced in the paper.

4 In-depth analysis on limitation and failure case

4.1 SeSame for "pedestrian" and "cyclist"

In this section, we analyze why the proposed method is effective for certain classes but not for others. Improvements were observed for the car, but not for pedestrian and cyclist. However, as mentioned in Sec. 4.2 of the paper, this is dependent on the performance of LiDAR semantic segmentation which has lower accuracy than 2D semantic segmentation for pedestrian and cyclist. This can be addressed with SOTA point-wise semantic segmentation which has higher accuracy for those.

Conversely, when compared to the method base on 2D semantic segmentation [5], its performance declined for car while it improved for pedestrian and cyclist as shown in Tab.5 of the paper. Analyzing the results from [5] and ours, image has advantage in extracting semantic information for objects with sparse point cloud, such as pedestrian and cyclist. On the other hand, the point cloud can identify an object with occlusion and truncation. For example, we can see this that SeSame+point successfully captured vehicles obscured by overlapping structures as illustrated in Fig. 1, and this supports that point clouds are more robust to occlusion and truncation than images. Therefore, LiDAR semantic segmentation has the advantage of accurately capuring semantics even in environment with occlusion and truncation.

4.2 The reason why AP is dropped for "easy"

As shown in Tab. 3 of the paper, there was performance drop for "easy" objects, which are easier to identify due to less occlusion and truncation. This decline is due to the data augmentation process discussed in Sec. 4 of the paper, where "easy" was augmented to resemble "moderate" and "hard", leading to reduced generalization performance for "easy". This effect is also observable in val set; while some performance improvement for "easy" objects was noted, it was less significant compared to the improvements seen in the "moderate" and "hard". The consistent results between the val and test set further substantiate this observation.

	AP_{3D}		
	easy	moderate	hard
PointRCNN[4]	86.75	76.05	74.30
SeSame (point)	88.76	78.35	77.43
delta	2.01	2.30	3.13
PointRCNN[4]	86.96	75.64	70.70
SeSame (point)	85.25	76.83	71.60
delta	-1.71	1.19	0.90

Table 5. Performance gain for car with **(up)** val and **(down)** with test set

5 Efficiency problem

Concerns regarding the necessity to train the semantic segmentation model separately stem from the modular nature of the proposed method, as opposed to an end-to-end approach. In other words, this provides the flexibility to individually optimize or replace each module. Additionally, there are concerns about computational cost, given that point-wise semantic segmentation must be performed first. Nevertheless, the proposed method, which incorporates point-wise semantic segmentation and 3D object detection, exhibits lower latency compared to the multi-modal methods it outperforms.

	method	latency (ms)
SeSame (pillar)	L	**126**
F-PointNet[2]	L+C	170
MV3D[1]	L+C	360
F-ConvNet[3]	L+C	470

Table 6. Comparison of latency with KITTI dataset

References

1. Chen, Xiaozhi and Ma, Huimin and Wan, Ji and Li, Bo and Xia, Tian, *Multi-View 3D Object Detection Network for Autonomous Driving*, , vol. , no. , pp. , 2017.
2. Qi, Charles R. and Liu, Wei and Wu, Chenxia and Su, Hao and Guibas, Leonidas J., *Frustum PointNets for 3D Object Detection From RGB-D Data*, , vol. , no. , pp. , 2018.
3. Wang, Zhixin and Jia, Kui, *Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection*, , vol. , no. , pp. 1742-1749, 2019.
4. Shi, Shaoshuai and Wang, Xiaogang and Li, Hongsheng, *PointRCNN: 3D object proposal generation and detection from point cloud*, , vol. 2019-June, no. , pp. 770–779, 2019.
5. Vora, Sourabh and Lang, Alex H. and Helou, Bassam and Beijbom, Oscar, *PointPainting: Sequential Fusion for 3D Object Detection*, , vol. , no. , pp. , 2020.