SEED: A Simple and Effective 3D DETR in Point Clouds

Zhe Liu*, Jinghua Hou*, Xiaoqing Ye, Tong Wang, Jingdong Wang, Xiang Bai†

Huazhong University of Science and Technology, Baidu Inc.

* Equal contribution † Corresponding author

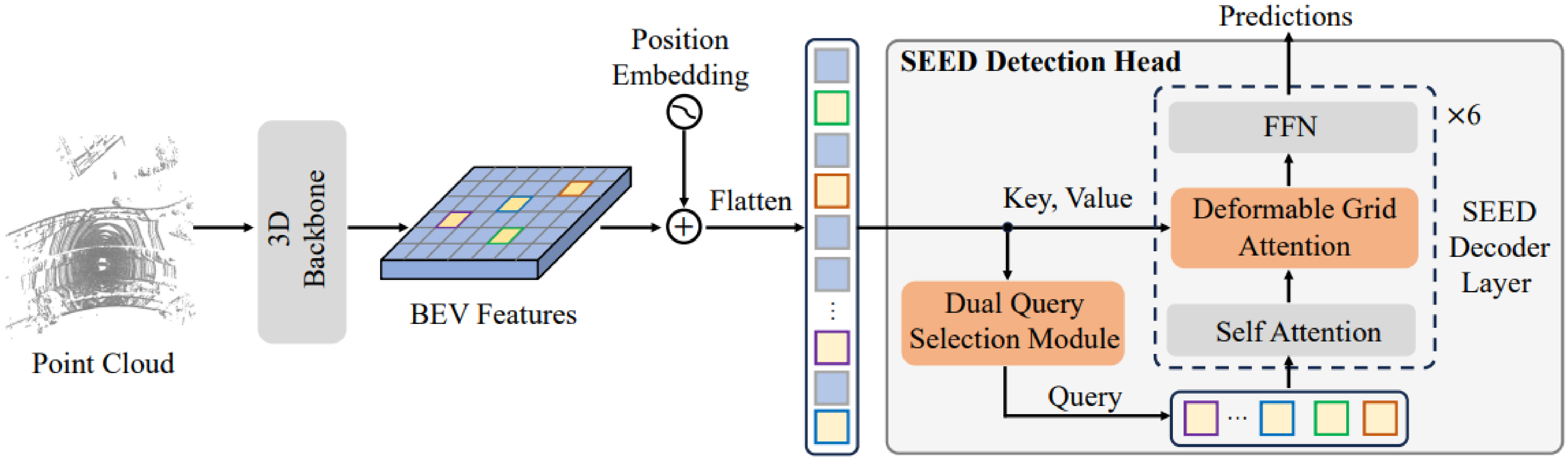
Motivation

- Recently, detection transformers (DETRs) have gradually taken a dominant position in 2D detection thanks to their elegant framework. DETR-based detectors for 3D point clouds are still difficult to achieve satisfactory performance
- How to obtain the appropriate object queries is challenging due to the high sparsity and uneven distribution of point clouds
- How to implement an effective query interaction by exploiting the rich geometric structure of point clouds is not fully explored

Contribution

- We introduce a novel dual query selection module, producing high-quality queries in a coarse-to-fine manner
- We adopt an effective deformable grid attention module, which adaptively aggregates crucial regions and performs informative query interaction by properly leveraging the geometric information of point clouds
- The proposed SEED achieves state-of-the-art performance for 3D object detection on both the large-scale Waymo and nuScenes datasets

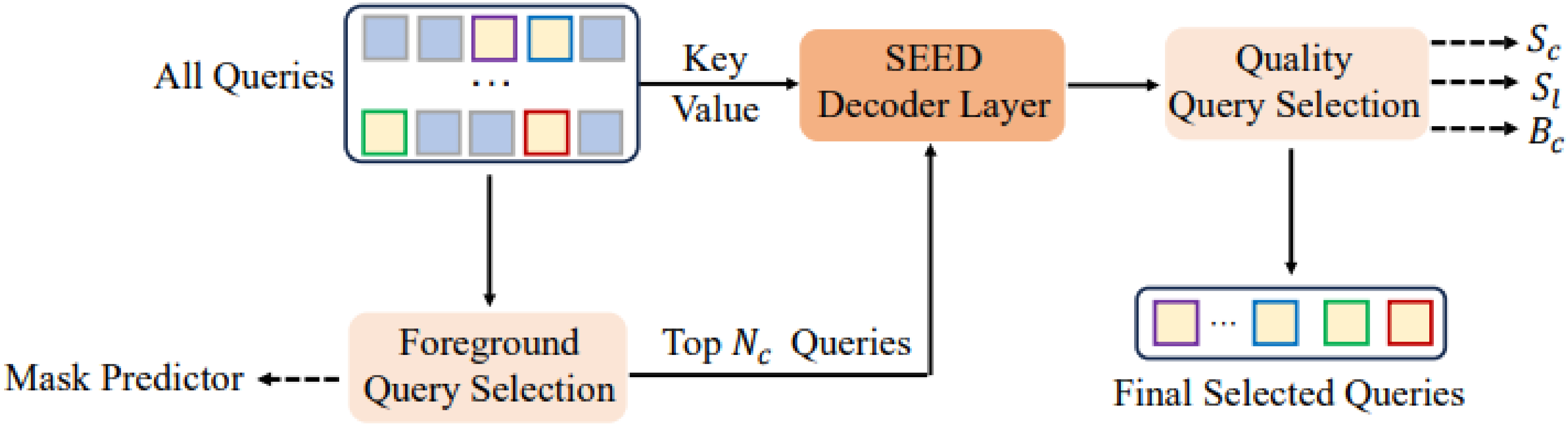
Method



SEED consists of a 3D backbone and a SEED detection head. Specifically, the proposed SEED detection head mainly includes a dual query selection (DQS) module and a transformer decoder.

- DQS utilizes a coarse-to-fine query selection strategy to select high-quality queries
- The transformer decoder, including six SEED decoder layers, takes these queries as inputs and then iteratively performs a self-attention operation for inter-query interaction and a proposed deformable grid attention (DGA) for feature interaction between query and BEV features

DQS



DQS adopts a coarse-to-fine manner, which consists of a foreground query selection and a quality query selection. Sc, Sl, and Bc are the predicted classification score, localization score, and regression for proposal boxes through three feed-forward networks (FFN) branches, respectively

DQS

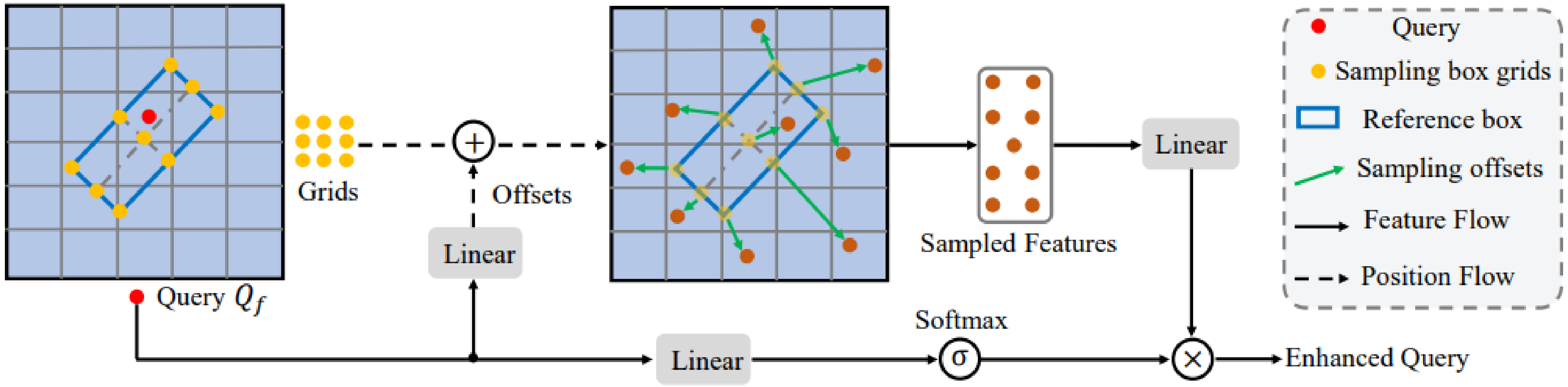
DQS adopts a coarse-to-fine manner, which consists of a foreground query selection and a quality query selection. Sc, Sl, and Bc are the predicted classification score, localization score, and regression for proposal boxes through three feed-forward networks (FFN) branches, respectively

$$\boldsymbol{S}_q^i = \begin{cases} (\boldsymbol{S}_c^i)^{1-\beta} \cdot (\boldsymbol{S}_l^i)^{\beta} \quad , & \text{if } \boldsymbol{S}_c^i > \tau, \\ \boldsymbol{S}_c^i \quad , & \text{otherwise}, \end{cases}$$

$$\boldsymbol{B}_f = \mathrm{Top}_{N_f}(\boldsymbol{B}_c, \boldsymbol{S}_q),$$

$$\boldsymbol{Q}_f = \mathrm{MLP}(\mathrm{Concat}(\boldsymbol{B}_f, \boldsymbol{S}_f)).$$

DGA



DGA first uniformly divides each reference proposal into grids as the reference points and then utilizes the predicted offsets to achieve a flexible receptive field

- Unlike 2D images, a nearby object may occupy most of the whole image, which even requires a global receptive field to detect the object well. However, a 3D object usually only occupies a small local area
- It is important to rationally utilize the geometric information of 3D objects
- A flexible receptive field is needed for some irregular objects or some hard objects

DGA

$\boldsymbol{F}_{bev}(*)$ denotes sampling the corresponding features of the positions * on the BEV features Fbev by a bilinear interpolation operation

ϕ: a linear function $\boldsymbol{g}_k$: reference grids

$$\mathrm{DGA}(\boldsymbol{g}, \boldsymbol{F}_{bev}) = \sum_{j=1}^{K} \boldsymbol{A}_j \cdot \phi(\boldsymbol{F}_{bev}(\boldsymbol{g}_j + \Delta\boldsymbol{g}_j)),$$

Specifically, we first regard the estimated proposal boxes Bf from DQS as the reference boxes and uniformly divide each reference box into k × k grids gk. Then, we feed the corresponding selected queries Qf from DQS into a linear function, producing the predicted offsets Δg. Next, we add the offsets to the grids g to generate the final sampling positions, which can capture the geometric information of 3D objects in a flexible receptive field. Meanwhile, the attention weight A is predicted by feeding Qf into a linear function and a softmax function.

Quality-aware Hungarian Matching

Different from the traditional Hungarian Matching, we introduce a quality-aware Hungarian Matching (QHM) to effectively assign the ground truth. QHM adopts the quality scores Sf instead of the classic classification scores in computing classification cost

$$\boldsymbol{\mathcal{C}}_{pos} = -(1-\alpha) \cdot (\boldsymbol{S}_f)^{\gamma} \cdot \log(1-\boldsymbol{S}_f),$$

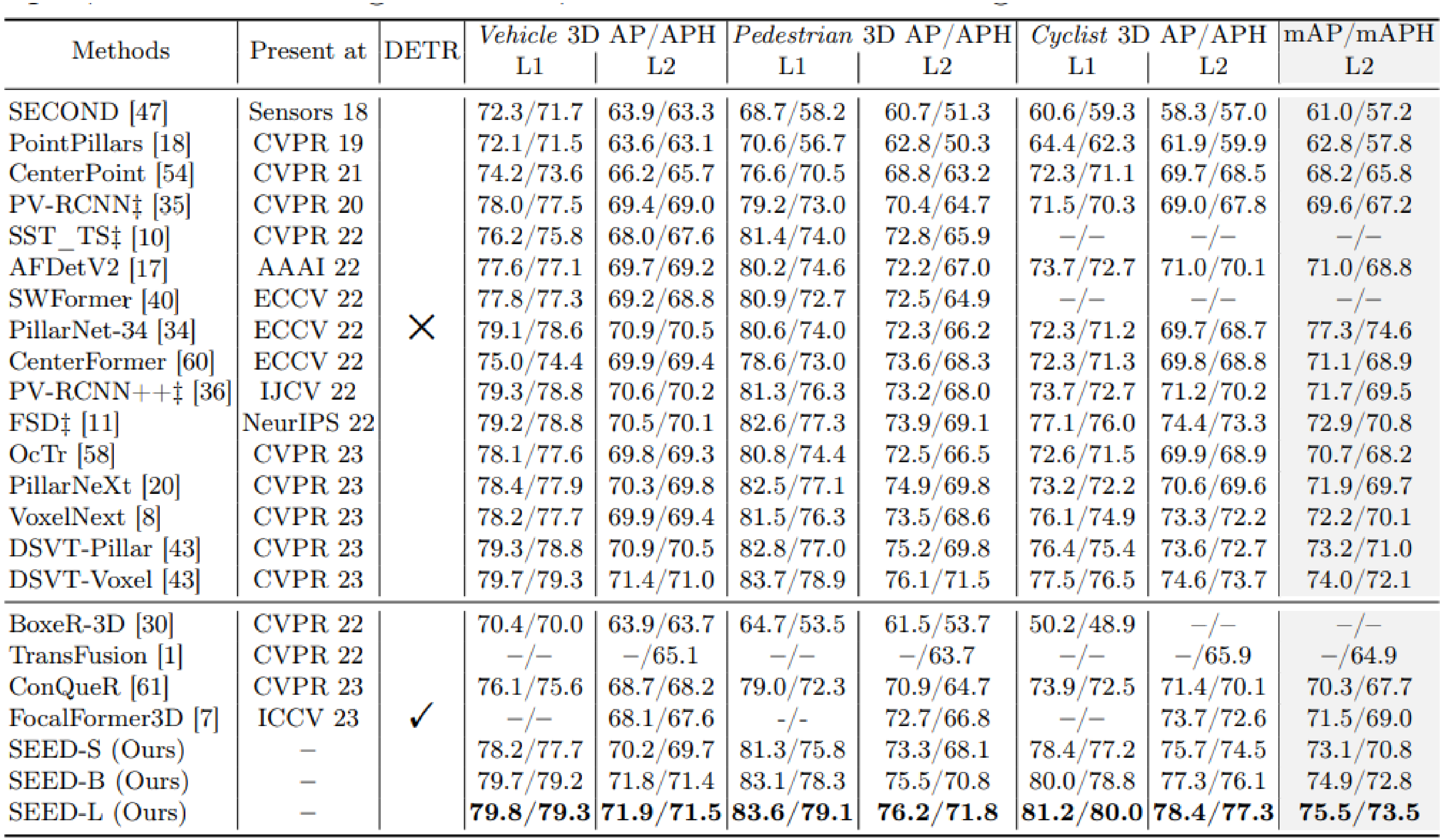
$$\boldsymbol{\mathcal{C}}_{neg} = -\alpha \cdot (1-\boldsymbol{S}_f)^{\gamma} \cdot \log \boldsymbol{S}_f,$$

$$\boldsymbol{\mathcal{C}}_{cls} = \boldsymbol{\mathcal{C}}_{pos} - \boldsymbol{\mathcal{C}}_{neg},$$

$$\boldsymbol{\mathcal{C}}_{\text{match}} = \lambda_{cls} \cdot \boldsymbol{\mathcal{C}}_{cls} + \lambda_{reg} \cdot \boldsymbol{\mathcal{C}}_{reg} + \lambda_{giou} \cdot \boldsymbol{\mathcal{C}}_{giou},$$

Results

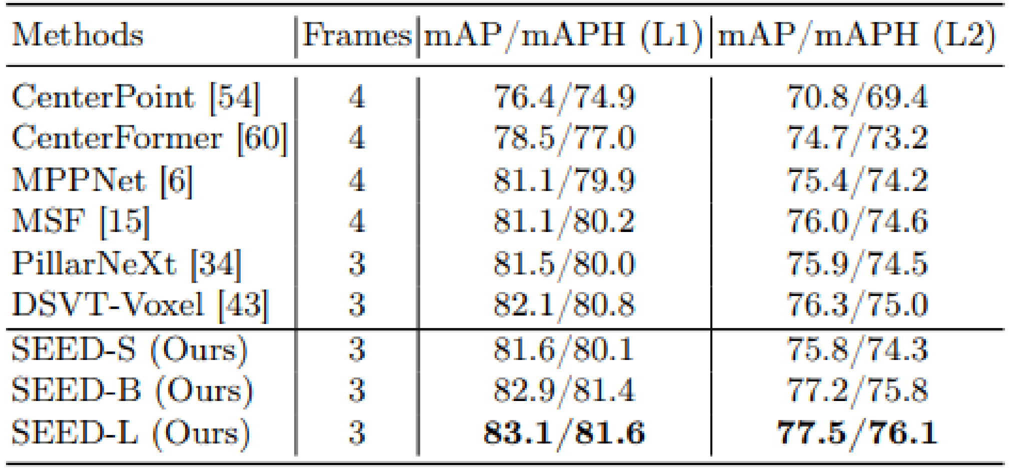
Comparison of other methods on the Waymo validation set. SEED achieves SOTA performance with 73.5 mAPH L2.



Methods	Present at	DETR	*Vehicle* 3D AP/APH		*Pedestrian* 3D AP/APH		*Cyclist* 3D AP/APH		mAP/mAPH
			L1	L2	L1	L2	L1	L2	L2
SECOND [47]	Sensors 18	×	72.3/71.7	63.9/63.3	68.7/58.2	60.7/51.3	60.6/59.3	58.3/57.0	61.0/57.2
PointPillars [18]	CVPR 19		72.1/71.5	63.6/63.1	70.6/56.7	62.8/50.3	64.4/62.3	61.9/59.9	62.8/57.8
CenterPoint [54]	CVPR 21		74.2/73.6	66.2/65.7	76.6/70.5	68.8/63.2	72.3/71.1	69.7/68.5	68.2/65.8
PV-RCNN‡ [35]	CVPR 20		78.0/77.5	69.4/69.0	79.2/73.0	70.4/64.7	71.5/70.3	69.0/67.8	69.6/67.2
SST_TS‡ [10]	CVPR 22		76.2/75.8	68.0/67.6	81.4/74.0	72.8/65.9	−/−	−/−	−/−
AFDetV2 [17]	AAAI 22		77.6/77.1	69.7/69.2	80.2/74.6	72.2/67.0	73.7/72.7	71.0/70.1	71.0/68.8
SWFormer [40]	ECCV 22		77.8/77.3	69.2/68.8	80.9/72.7	72.5/64.9	−/−	−/−	−/−
PillarNet-34 [34]	ECCV 22		79.1/78.6	70.9/70.5	80.6/74.0	72.3/66.2	72.3/71.2	69.7/68.7	77.3/74.6
CenterFormer [60]	ECCV 22		75.0/74.4	69.9/69.4	78.6/73.0	73.6/68.3	72.3/71.3	69.8/68.8	71.1/68.9
PV-RCNN++‡ [36]	IJCV 22		79.3/78.8	70.6/70.2	81.3/76.3	73.2/68.0	73.7/72.7	71.2/70.2	71.7/69.5
FSD‡ [11]	NeurIPS 22		79.2/78.8	70.5/70.1	82.6/77.3	73.9/69.1	77.1/76.0	74.4/73.3	72.9/70.8
OcTr [58]	CVPR 23		78.1/77.6	69.8/69.3	80.8/74.4	72.5/66.5	72.6/71.5	69.9/68.9	70.7/68.2
PillarNeXt [20]	CVPR 23		78.4/77.9	70.3/69.8	82.5/77.1	74.9/69.8	73.2/72.2	70.6/69.6	71.9/69.7
VoxelNext [8]	CVPR 23		78.2/77.7	69.9/69.4	81.5/76.3	73.5/68.6	76.1/74.9	73.3/72.2	72.2/70.1
DSVT-Pillar [43]	CVPR 23		79.3/78.8	70.9/70.5	82.8/77.0	75.2/69.8	76.4/75.4	73.6/72.7	73.2/71.0
DSVT-Voxel [43]	CVPR 23		79.7/79.3	71.4/71.0	83.7/78.9	76.1/71.5	77.5/76.5	74.6/73.7	74.0/72.1
BoxeR-3D [30]	CVPR 22	✓	70.4/70.0	63.9/63.7	64.7/53.5	61.5/53.7	50.2/48.9	−/−	−/−
TransFusion [1]	CVPR 22		−/−	−/65.1	−/−	−/63.7	−/−	−/65.9	−/64.9
ConQueR [61]	CVPR 23		76.1/75.6	68.7/68.2	79.0/72.3	70.9/64.7	73.9/72.5	71.4/70.1	70.3/67.7
FocalFormer3D [7]	ICCV 23		−/−	68.1/67.6	-/-	72.7/66.8	−/−	73.7/72.6	71.5/69.0
SEED-S (Ours)	−		78.2/77.7	70.2/69.7	81.3/75.8	73.3/68.1	78.4/77.2	75.7/74.5	73.1/70.8
SEED-B (Ours)	−		79.7/79.2	71.8/71.4	83.1/78.3	75.5/70.8	80.0/78.8	77.3/76.1	74.9/72.8
SEED-L (Ours)	−		**79.8/79.3**	**71.9/71.5**	**83.6/79.1**	**76.2/71.8**	**81.2/80.0**	**78.4/77.3**	**75.5/73.5**

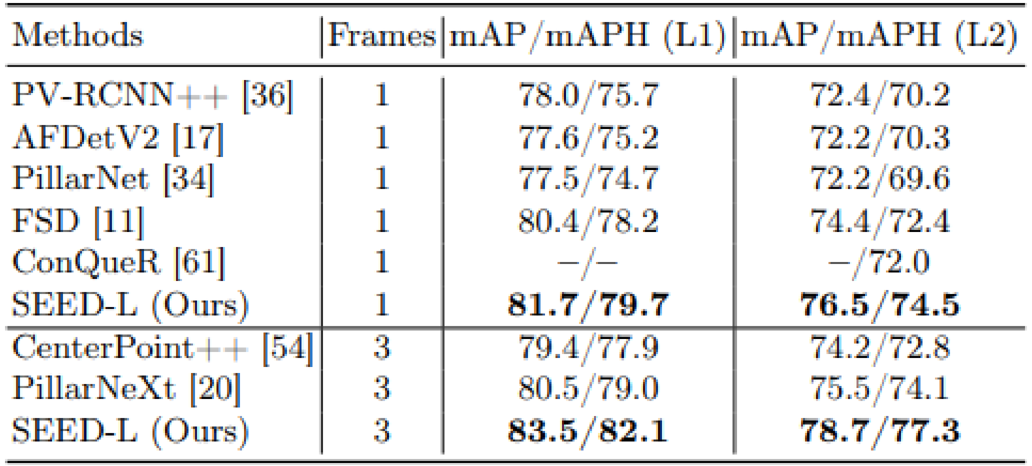
Results

Effectiveness of our SEED with multiple frames as inputs on the Waymo Open Dataset validation and test split. SEED achieves SOTA performance.



Methods	Frames	mAP/mAPH (L1)	mAP/mAPH (L2)
CenterPoint [54]	4	76.4/74.9	70.8/69.4
CenterFormer [60]	4	78.5/77.0	74.7/73.2
MPPNet [6]	4	81.1/79.9	75.4/74.2
MSF [15]	4	81.1/80.2	76.0/74.6
PillarNeXt [34]	3	81.5/80.0	75.9/74.5
DSVT-Voxel [43]	3	82.1/80.8	76.3/75.0
SEED-S (Ours)	3	81.6/80.1	75.8/74.3
SEED-B (Ours)	3	82.9/81.4	77.2/75.8
SEED-L (Ours)	3	**83.1/81.6**	**77.5/76.1**

(a) Effectiveness of SEED on *validation* split.

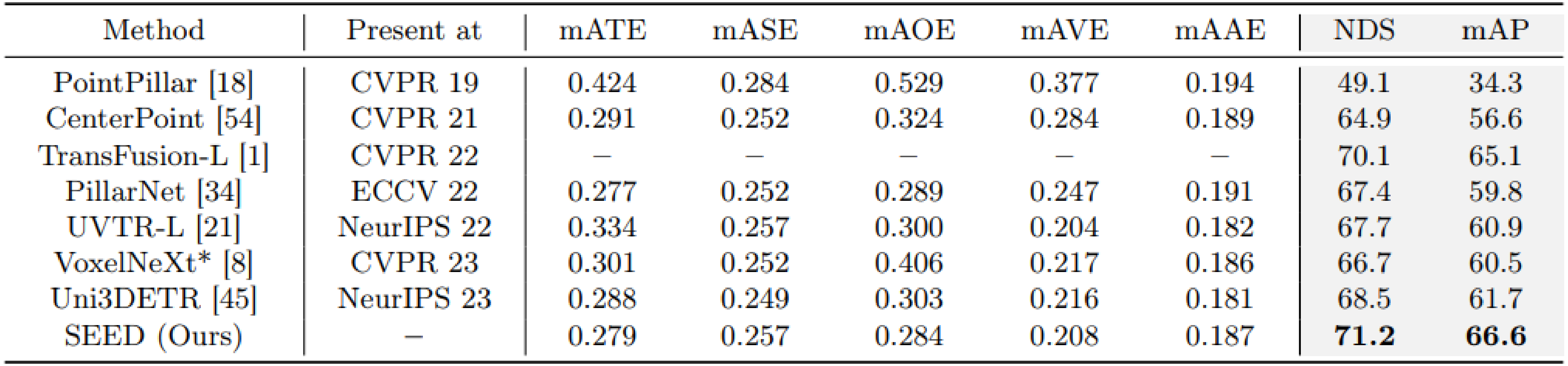


Methods	Frames	mAP/mAPH (L1)	mAP/mAPH (L2)
PV-RCNN++ [36]	1	78.0/75.7	72.4/70.2
AFDetV2 [17]	1	77.6/75.2	72.2/70.3
PillarNet [34]	1	77.5/74.7	72.2/69.6
FSD [11]	1	80.4/78.2	74.4/72.4
ConQueR [61]	1	–/–	–/72.0
SEED-L (Ours)	1	**81.7/79.7**	**76.5/74.5**
CenterPoint++ [54]	3	79.4/77.9	74.2/72.8
PillarNeXt [20]	3	80.5/79.0	75.5/74.1
SEED-L (Ours)	3	**83.5/82.1**	**78.7/77.3**

(b) Effectiveness of SEED on *test* benchmark.

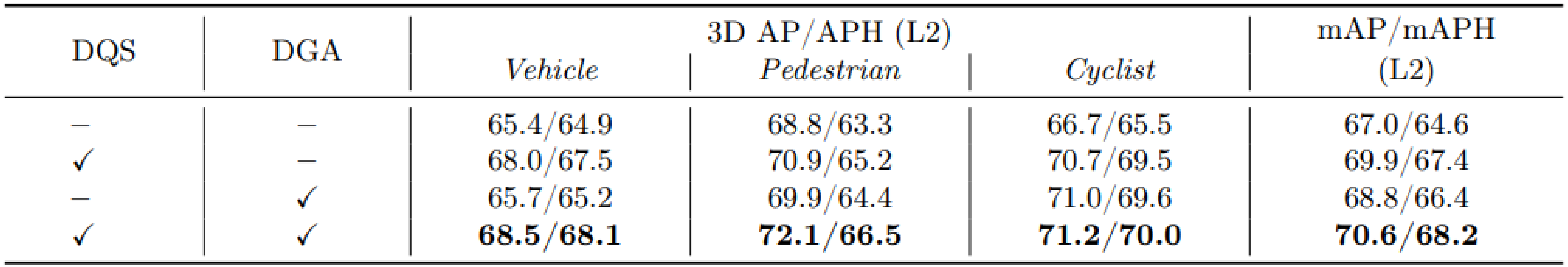
Results

Effectiveness of our SEED on the nuScenes validation set. SEED achieves SOTA performance.



Method	Present at	mATE	mASE	mAOE	mAVE	mAAE	NDS	mAP
PointPillar [18]	CVPR 19	0.424	0.284	0.529	0.377	0.194	49.1	34.3
CenterPoint [54]	CVPR 21	0.291	0.252	0.324	0.284	0.189	64.9	56.6
TransFusion-L [1]	CVPR 22	–	–	–	–	–	70.1	65.1
PillarNet [34]	ECCV 22	0.277	0.252	0.289	0.247	0.191	67.4	59.8
UVTR-L [21]	NeurIPS 22	0.334	0.257	0.300	0.204	0.182	67.7	60.9
VoxelNeXt* [8]	CVPR 23	0.301	0.252	0.406	0.217	0.186	66.7	60.5
Uni3DETR [45]	NeurIPS 23	0.288	0.249	0.303	0.216	0.181	68.5	61.7
SEED (Ours)	–	0.279	0.257	0.284	0.208	0.187	**71.2**	**66.6**

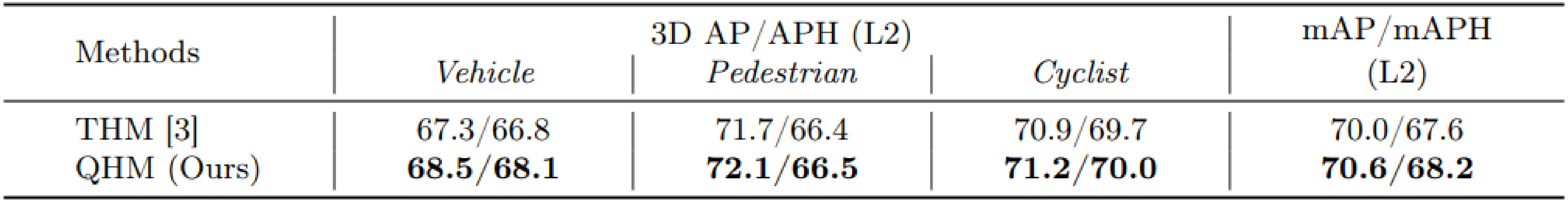
Ablation Studies



DQS	DGA	3D AP/APH (L2)			mAP/mAPH
		Vehicle	*Pedestrian*	*Cyclist*	(L2)
–	–	65.4/64.9	68.8/63.3	66.7/65.5	67.0/64.6
✓	–	68.0/67.5	70.9/65.2	70.7/69.5	69.9/67.4
–	✓	65.7/65.2	69.9/64.4	71.0/69.6	68.8/66.4
✓	✓	**68.5/68.1**	**72.1/66.5**	**71.2/70.0**	**70.6/68.2**

- DQS is effective for selecting out high-quality queries
- DGA is effective for achieving better feature interaction

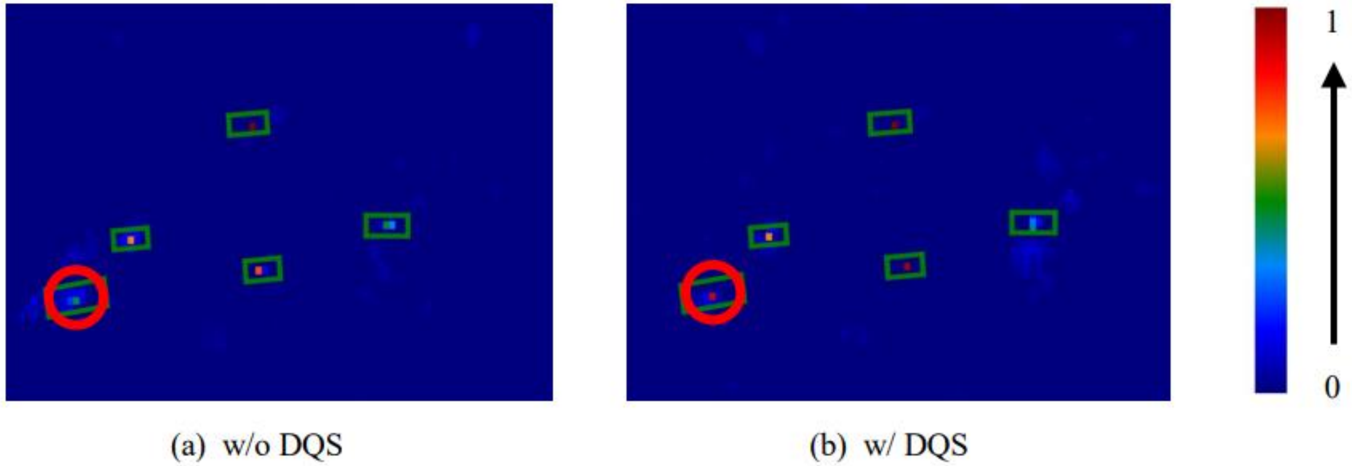
Ablation Studies



Methods	3D AP/APH (L2) *Vehicle*	 *Pedestrian*	 *Cyclist*	mAP/mAPH (L2)
THM [3]	67.3/66.8	71.7/66.4	70.9/69.7	70.0/67.6
QHM (Ours)	**68.5/68.1**	**72.1/66.5**	**71.2/70.0**	**70.6/68.2**

Taking the quality scores of 3D objects into account when computing classification cost in Hungarian Matching is effective.

Visualization

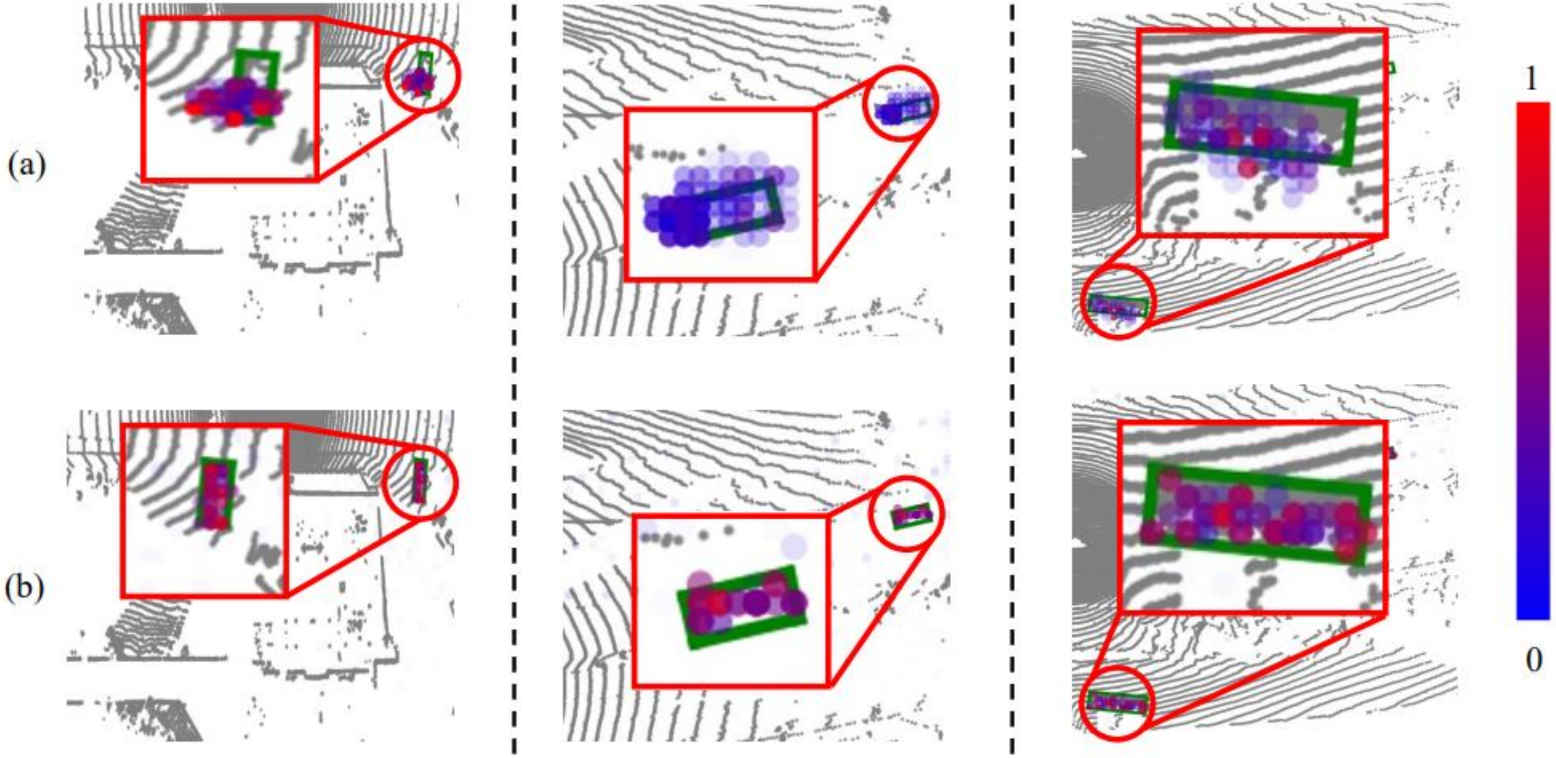


(a) w/o DQS

(b) w/ DQS

Visualization of SEED without DQS (the first row) and with DQS (the second row). It can be observed that after utilizing DQS, some hard queries are successfully captured, which indicates that DQS can enhance the confidence score of some potential hard objects.

Visualization



Comparison of attention map without DGA (a) and with DGA (b) on the Waymo validation set. After utilizing DGA, SEED can capture the geometric information of 3D objects in a flexible receptive field and achieve better query interaction.

Thanks!