QCRAFT

PillarNeXt: Rethinking Network Designs for 3D Object Detection in LiDAR Point Clouds

Jinyu Li Chenxu Luo Xiaodong Yang

Introduction

Mainstream: Develop operators for point clouds

- Encoders
  - Point-based
    - PointNet
  - Grid-based
    - Voxel
    - Pillar
    - RangeView

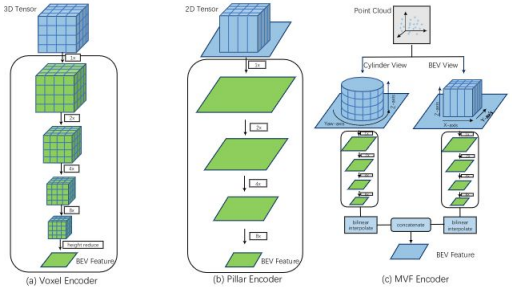
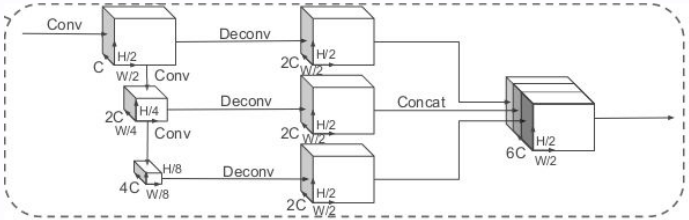


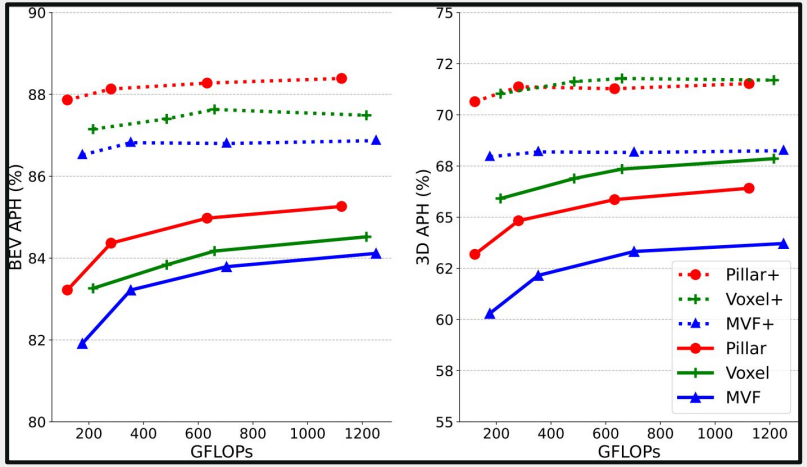
Figure 2. Comparison of the voxel, pillar and MVF encoders with a unified detection network. Each encoder is followed by a sparse convolution based ResNet-18. For the pillar and MVF encoders, we use 2D convolutions, and for the voxel encoder, we use 3D convolutions.

Unexplored: Network designs

- Networks
  - PointPillars
  - VoxelNet
  - SECOND



Local Point Aggregators Matter?

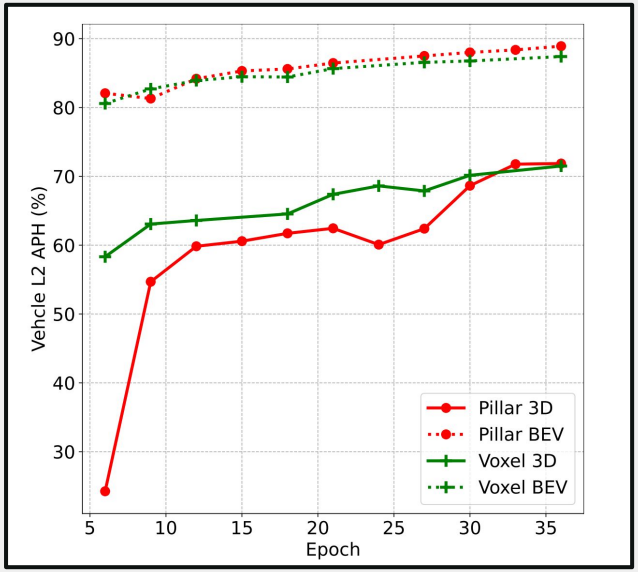


The simplest pillar based models outperform on BEV and are on-par with the fine-grained voxel and MVF based networks.

Model	Channels	#Params (M)	FLOPs (G)	Latency (ms)	Vehicle		Pedestrian	
					3D	BEV	3D	BEV
Pillar-T	[32, 64, 128, 128]	1.65	70	52	**62.03**	**82.26**	**67.63**	**75.76**
MVF-T	[32, 64, 128, 128]	3.44	78	137	59.16	81.33	64.10	73.42
Pillar-S	[42, 84, 168, 168]	2.83	121	79	63.18	**83.22**	68.12	**76.37**
Voxel-S	[12, 24, 48, 96]	1.53	121	169	**64.67**	82.45	**69.10**	76.29
MVF-S	[44, 88, 176, 176]	6.38	148	186	61.06	82.51	65.15	74.24
Pillar-B	[64, 128, 256, 256]	6.53	281	103	64.83	**84.37**	69.04	**76.96**
Voxel-B	[18, 36, 72, 144]	3.42	272	226	**66.00**	83.32	**69.45**	76.38
MVF-B	[68, 136, 272, 272]	15.02	353	291	62.15	83.22	66.12	75.00
Pillar-L	[96, 192, 384, 384]	14.63	632	194	65.86	**84.98**	68.42	76.67
Voxel-L	[28, 56, 112, 224]	8.27	660	299	**67.35**	84.17	**70.47**	**77.44**
MVF-L	[96, 192, 384, 384]	29.67	704	390	63.32	83.79	66.87	75.34
Pillar-S+	[42, 84, 168, 168]	2.83	121	79	**70.64**	**87.86**	73.48	79.95
Voxel-S+	[12, 24, 48, 96]	1.53	121	169	70.61	86.83	**74.26**	**80.37**
MVF-S+	[44, 88, 176, 176]	6.38	148	186	67.46	86.52	70.87	78.18
Pillar-B+	[64, 128, 256, 256]	6.53	281	103	**71.37**	**88.13**	73.93	80.28
Voxel-B+	[18, 36, 72, 144]	3.42	272	226	71.36	87.33	**74.76**	**80.74**
MVF-B+	[68, 136, 272, 272]	15.02	353	291	68.19	86.82	71.42	78.51

Even with similar FLOPs, the pillar based models run much faster than the voxel and MVF based networks.

Training Matters



Pillar based models need more epochs to converge on the height axis.

Receptive Fields Matter

Method	Vehicle L1		Vehicle L2		Pedestrian L1		Pedestrian L2	
	AP	APH	AP	APH	AP	APH	AP	APH
Neck of PillarNet [31]	91.39	90.58	84.54	83.72	**87.90**	**83.02**	81.93	77.20
FPN [17]	92.17	91.35	85.96	85.13	87.88	82.91	82.05	77.23
BiFPN [39]	92.71	91.90	86.92	86.09	87.86	82.88	82.05	77.23
Plain	91.01	90.19	83.86	83.04	87.59	82.61	81.52	76.71
Dilated Block [7]	92.70	91.90	86.61	85.79	87.84	82.91	**82.09**	**77.29**
ASPP [5]	**92.77**	**91.94**	**86.99**	**86.14**	87.74	82.85	82.00	77.26

Multi-scale feature fusion may not be essential on BEV space. We enlarge receptive fields and the performance can be boosted.

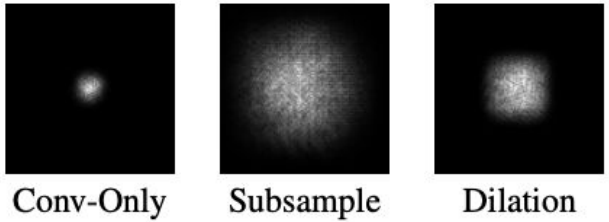
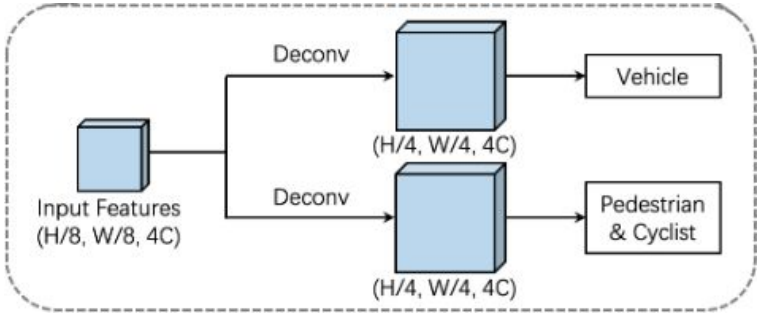


Figure from *Understanding the Effective Receptive Field in Deep Convolutional Neural Networks*

Resolutions Matter

In Size	Backbone ↓	Head ↑	Out Size	Veh	Ped	Latency
0.3	1	1	0.3	65.0	67.2	255
0.075	8	1	0.6	62.8	66.6	131
0.075	8	2	0.3	64.8	69.0	173

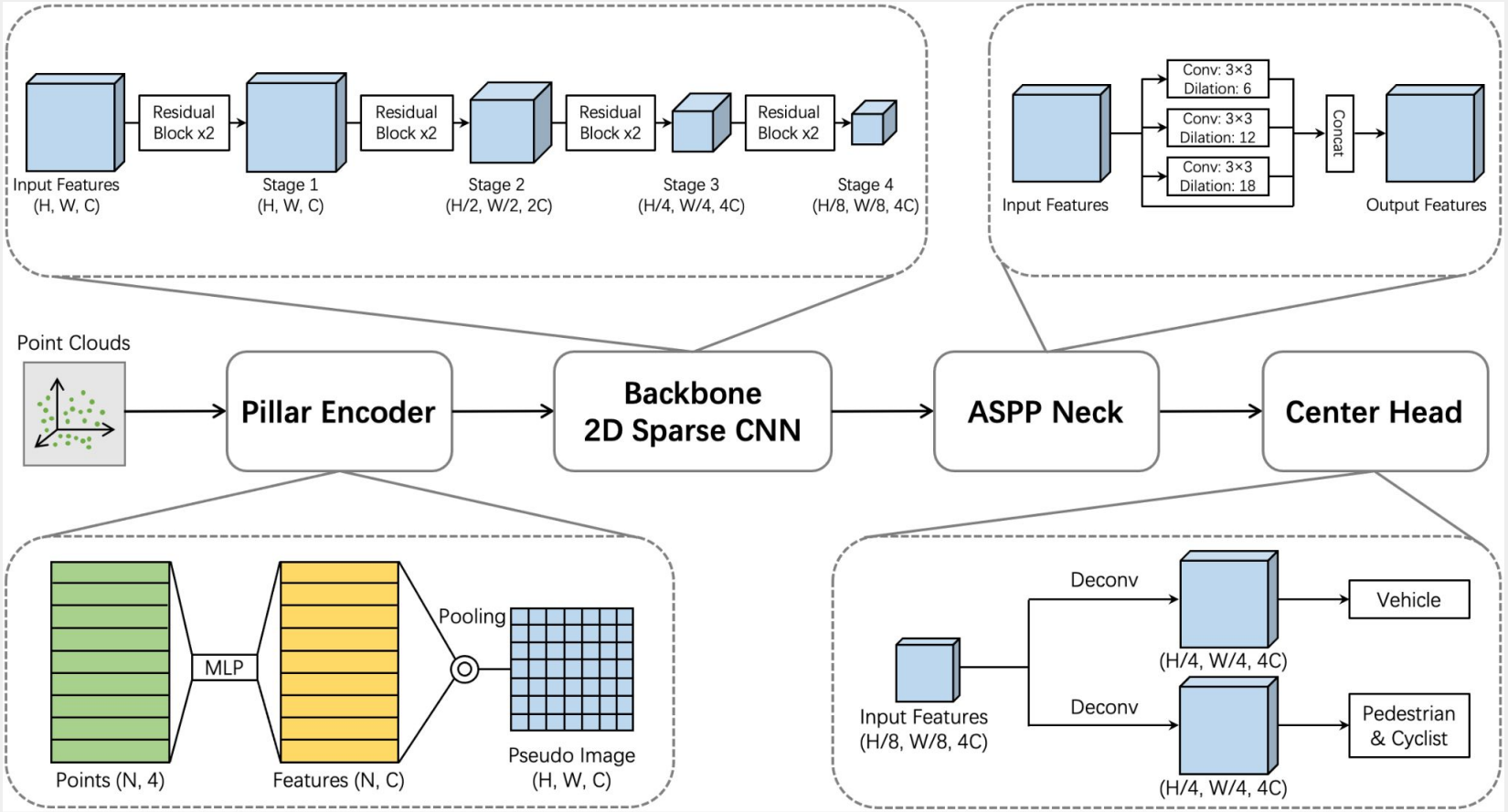
To attain a tradeoff between latency and accuracy, we put up with a upsampling layer to address information loss during downsampling.



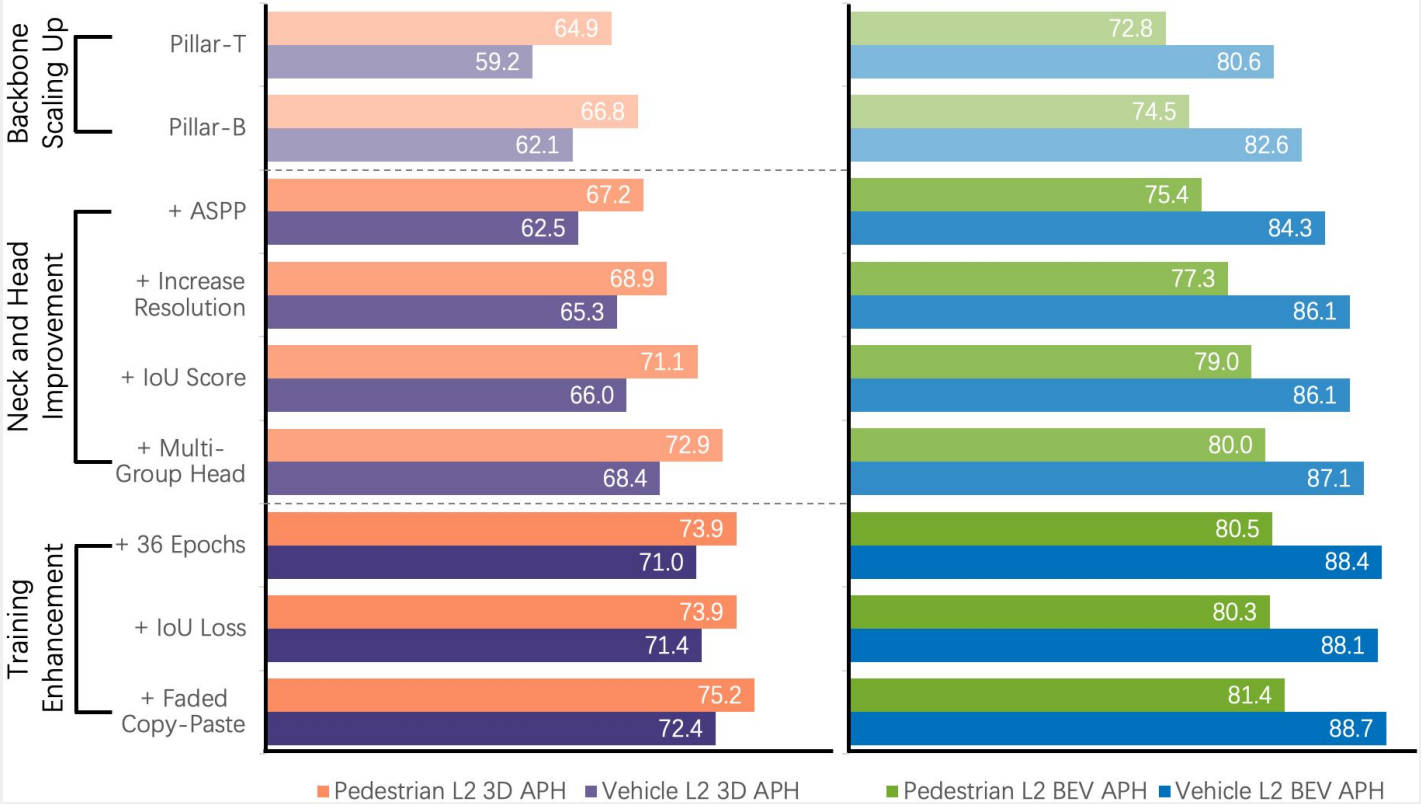
Takeaways

- Fine-grained local operators are not necessary. Pillar is more efficient and effective under similar cost.
- Enlarging receptive field is the key.
- A simple upsampling layer is sufficient for fine-grained modeling.
- Training matters.

PillarNeXt Architecture



Summary


Backbone Scaling Up
Pillar-T
64.9
59.2
72.8
80.6
Pillar-B
66.8
62.1
74.5
82.6
Neck and Head Improvement
+ ASPP
67.2
62.5
75.4
84.3
+ Increase Resolution
68.9
65.3
77.3
86.1
+ IoU Score
71.1
66.0
79.0
86.1
+ Multi-Group Head
72.9
68.4
80.0
87.1
Training Enhancement
+ 36 Epochs
73.9
71.0
80.5
88.4
+ IoU Loss
73.9
71.4
80.3
88.1
+ Faded Copy-Paste
75.2
72.4
81.4
88.7
Pedestrian L2 3D APH
Vehicle L2 3D APH
Pedestrian L2 BEV APH
Vehicle L2 BEV APH

Results on Waymo Validation Set (3D)

QCRAFT

Method	Frames	Vehicle L1		Vehicle L2		Pedestrian L1		Pedestrian L2		Cyclist L1		Cyclist L2	
		AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
SST-TS* [11]	1	76.22	75.79	68.04	67.64	81.39	74.05	72.82	65.93	-	-	-	-
SWFormer [36]	1	77.8	77.3	69.2	68.8	80.9	72.7	72.5	64.9	-	-	-	-
PillarNet-18 [31]	1	78.24	77.73	70.40	69.92	79.80	72.59	71.57	64.90	70.40	69.29	67.75	66.68
AFDetV2 [13]	1	77.64	77.14	69.68	69.22	80.19	74.62	72.16	66.95	73.72	72.74	71.06	70.12
PV-RCNN++* [32]	1	79.25	78.78	70.61	70.18	81.83	76.28	73.17	68.00	73.72	72.66	71.21	70.19
PillarNeXt-B	1	78.40	77.90	70.27	69.81	82.53	77.14	74.90	69.80	73.21	72.20	70.58	69.62
PillarNet-18 [31]	2	79.59	79.06	71.56	71.08	82.11	78.82	74.49	71.35	70.41	69.57	68.27	67.46
PillarNet-34 [31]	2	79.98	79.47	72.00	71.53	82.52	79.33	75.00	71.95	70.51	69.69	68.38	67.58
PV-RCNN++* [32]	2	80.17	79.70	72.14	71.70	83.48	80.42	75.54	72.61	74.63	73.75	72.35	71.50
RSN* [37]	3	78.4	78.1	69.5	69.1	79.4	76.2	69.9	67.0	-	-	-	-
SST-TS* [11]	3	78.66	78.21	69.98	69.57	83.81	80.14	75.94	72.37				
SWFormer [36]	3	79.4	78.9	71.1	70.6	82.9	79.0	74.8	71.1	-	-	-	-
PillarNeXt-B	3	**80.58**	**80.08**	**72.89**	**72.42**	**85.04**	**82.11**	**78.04**	**75.19**	**78.92**	**77.94**	**76.71**	**75.74**
CenterFormer [49]	8	78.8	78.3	74.3	73.8	82.1	79.3	77.8	75.0	75.2	74.4	73.2	72.3
MPPNet [8]	16	82.74	82.28	75.41	74.96	84.69	82.25	77.43	75.06	77.28	76.66	75.13	74.52
3DAL† [29]	ALL	84.50	-	-	-	82.88	-	-	-	-	-	-	-

Table 4. Comparison of PillarNeXt-B and the state-of-the-art methods under the 3D metrics on the validation set of WOD. * denotes the two-stage models and † indicates the off-board method.

Results on Waymo Validation Set (BEV)

QCRAFT

Method	Frames	Vehicle L1		Vehicle L2		Pedestrian L1		Pedestrian L2		Cyclist L1		Cyclist L2	
		AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
PV-RCNN++* [32]	1	91.57	-	-	-	85.43	-	-	-	75.94	-	-	-
PillarNeXt-B	1	93.30	92.60	87.26	86.53	88.19	82.13	81.77	75.82	75.67	74.61	72.97	71.95
SWFormer [36]	3	92.60	-	-	-	87.50	-	-	-	-	-	-	-
PillarNeXt-B	3	**94.41**	93.73	89.36	88.66	**90.20**	86.94	84.66	81.36	81.35	80.32	79.23	78.22
3DAL† [29]	ALL	93.30	-	-	-	86.32	-	-	-	-	-	-	-

Table 5. Comparison of PillarNeXt-B and the state-of-the-art methods under the BEV metrics on the validation set of WOD. * denotes the two-stage models and † indicates the off-board method.

Results on Waymo Test Set (3D)

QCRAFT

Method	Frames	All L2		Vehicle L1		Vehicle L2		Pedestrian L1		Pedestrian L2		Cyclist L1		Cyclist L2	
		mAP	mAPH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
SWFormer [36]	3	-	-	82.89	82.49	75.02	74.65	82.13	78.13	75.87	72.07	-	-	-	-
PillarNet-34† [31]	3	73.98	72.48	83.23	82.80	76.09	75.69	82.38	79.02	76.66	73.46	71.44	70.51	69.20	68.29
CenterPoint++ [45]	3	74.20	72.80	82.80	82.30	75.50	75.10	81.00	78.20	75.10	72.40	74.40	73.30	72.00	71.00
AFDetV2 [13]	2	74.60	73.12	81.65	81.22	74.30	73.89	81.26	78.05	75.47	72.41	76.41	75.37	74.05	73.04
PV-RCNN++* [32]	2	75.00	73.52	83.74	83.32	76.31	75.92	82.60	79.38	76.63	73.55	74.44	73.43	72.06	71.09
PillarNeXt-B	3	**75.53**	**74.10**	83.28	82.83	76.18	75.76	84.40	81.44	78.84	75.98	73.77	72.73	71.56	70.55

Table 6. Comparison of PillarNeXt-B and the state-of-the-art methods under the 3D metrics on the test set of WOD. * denotes the two-stage model and † indicates using test-time augmentations.

Results on nuScenes

QCRAFT

Method	Encoder	Grid Size	NDS	mAP	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint [45]	V	0.075	66.8	59.6	0.292	0.255	0.302	0.259	0.193
OHS [6]	V	0.1	66.0	59.5	-	-	-	-	-
PillarNet-18 [31]	P	0.075	67.4	59.9	-	-	-	-	-
Transfusion-L [1]	V	0.075	66.8	60.0	-	-	-	-	-
UVTR-L [15]	V	0.075	67.7	60.9	0.334	0.257	0.300	0.204	0.182
VISTA [9]	V+R	0.1	68.1	60.8	-	-	-	-	-
PillarNeXt-B	P	0.075	**68.8**	**62.5**	0.278	0.251	0.269	0.248	0.201
Our Voxel-B	V	0.075	68.2	62.4	0.278	0.250	0.308	0.263	0.198

Table 7. Comparison of PillarNeXt-B and the state-of-the-art methods on the validation set of nuScenes. P/V/R denotes the pillar, voxel and range view based grid encoder, respectively. Most leading methods adopt the voxel based representations.

Method	Car	Truck	Bus	Trailer	CV	Ped	Motor	Bicycle	TC	Barrier	mAP
PillarNeXt-B	84.8	58.6	66.5	35.3	21.4	87.2	68.0	56.4	77.0	69.8	62.5
Our Voxel-B	84.3	58.3	69.3	37.1	21.4	87.4	67.6	54.7	75.0	69.2	62.4

Table 8. Comparison of our proposed pillar and voxel based models under per-class AP and mAP on the validation set of nuScenes. Abbreviations are construction vehicle (CV), pedestrian (Ped), motorcycle (Motor), and traffic cone (TC).

QCRAFT

Thanks!