QCRAFT

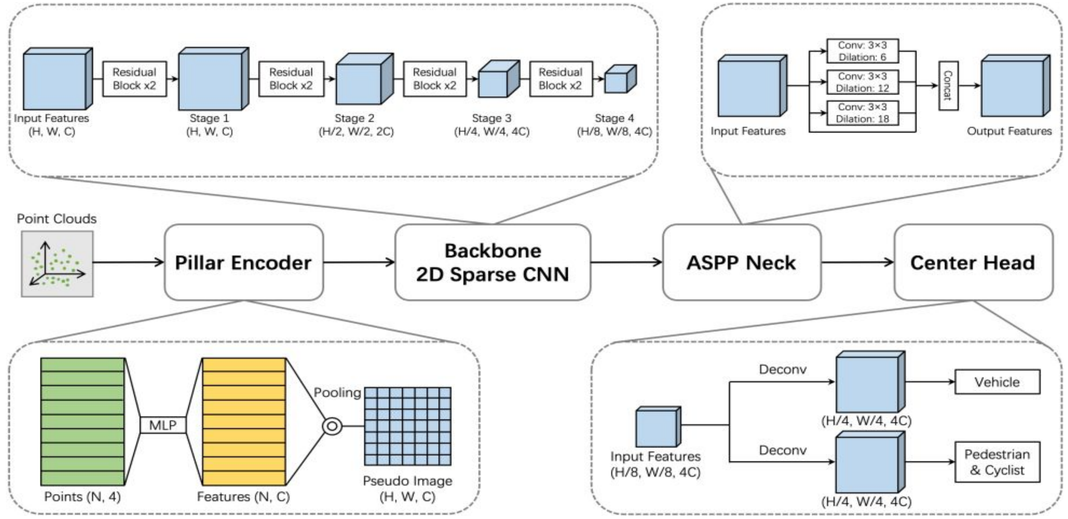
PillarNeXt: Rethinking Network Designs for 3D Object Detection in LiDAR Point Clouds

Jinyu Li Chenxu Luo Xiaodong Yang

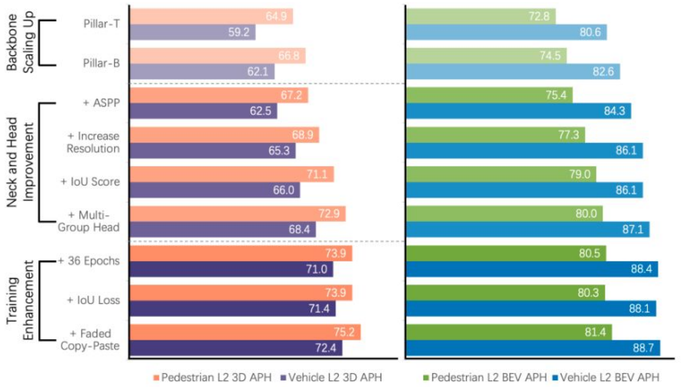
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- Existing works on 3D object detection from point cloud mainly focus on designing sophisticated local aggregators
- Network architecture for 3D object detection has been less studied.
- Enlarging receptive field accounts more than modeling detailed structures.
- compare models under similar #params or FLOPs is a common practice in 2D domain, yet has not been widely adopted in 3D.
- Training matters

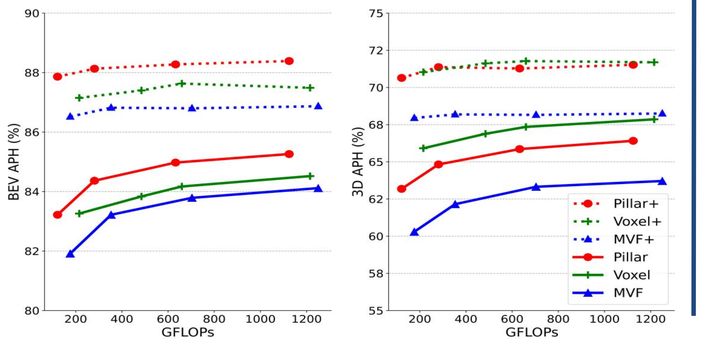


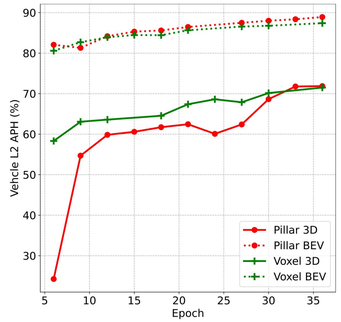
Summary



Do We Need Sophisticated Aggregators?

- simplest pillar-based model performs surprisingly well considering both accuracy and latency.
- Pillar needs longer time to converge on height dimension





Upsample to Address Information Loss

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

Receptive Fields Matter

- Multi-scale feature is not necessary
- Enlarging receptive field is crucial

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	81.51	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

Results

Waymo (test)

Method	Frames	All L2 mAP	All L2 mAPH	Vehicle L1 AP	Vehicle L1 APH	Vehicle L2 AP	Vehicle L2 APH	Pedestrian L1 AP	Pedestrian L1 APH	Pedestrian L2 AP	Pedestrian L2 APH	Cyclist L1 AP	Cyclist L1 APH	Cyclist L2 AP	Cyclist L2 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.

nuScenes

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).