FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer – Supplementary Material –

A.1. Full Results of Single-Stage 3D Detectors on Waymo Open Dataset

	#Frames	Mean L1 (mAP/APH)	Mean L2 (mAP/APH)	Vehicle L1 (mAP/APH)	Vehicle L2 (mAP/APH)	Pedestrian L1 (mAP/APH)	Pedestrian L2 (mAP/APH)	Cyclist L1 (mAP/APH)	Cyclist L2 (mAP/APH)
SECOND [13][3]	1	67.2 / 63.1	61.0 / 57.2	72.3 / 71.7	63.9 / 63.3	68.7 / 58.2	60.7 / 51.3	60.6 / 59.3	58.3 / 57.0
PointPillars [5][3]	1	69.0 / 63.5	62.8 / 57.8	72.1 / 71.5	63.6 / 63.1	70.6 / 56.7	62.8 / 50.3	64.4 / 62.3	61.9 / 59.9
○ CenterPoint [14][1]	1	74.0 / 71.3	67.9 / 65.5	74.8 / 74.2	66.7 / 66.2	75.8 / 69.7	68.3 / 62.6	71.3 / 70.2	68.7 / 67.6
● VoTr-SSD [7]	1	–	–	69.0 / 68.4	60.2 / 59.7	–	–	–	–
● SST [2][2]	1	74.8 / 71.1	68.2 / 64.8	73.6 / 73.1	64.8 / 64.4	80.0 / 70.6	71.7 / 63.0	70.7 / 69.6	68.0 / 66.9
● SST-Center [2]	1	75.6 / 72.3	69.3 / 66.3	75.1 / 74.6	66.6 / 66.2	80.1 / 72.1	72.4 / 65.0	71.5 / 70.2	68.9 / 67.6
● VoxSet [4]	1	75.4 / 72.2	69.1 / 66.2	74.5 / 74.0	66.0 / 65.6	80.0 / 72.4	72.5 / 65.4	71.6 / 70.3	69.0 / 67.7
○ PillarNet [9]	1	76.1 / 72.9	69.9 / 67.2	78.2 / 77.7	70.4 / 69.9	79.8 / 72.6	71.6 / 64.9	70.4 / 69.3	67.8 / 66.7
● **FlatFormer** (Ours)	1	76.1 / 73.4	69.7 / 67.2	77.5 / 77.1	69.0 / 68.6	79.6 / 73.0	71.5 / 65.3	71.3 / 70.1	68.6 / 67.5
○ CenterPoint [14][1]	2	75.9 / 74.2	70.1 / 68.4	75.7 / 75.2	67.7 / 67.2	78.3 / 74.6	71.0 / 67.5	73.8 / 72.9	71.5 / 70.5
○ PillarNet [9]	2	77.3 / 75.8	71.5 / 70.0	79.6 / 79.1	71.6 / 71.1	82.1 / 78.8	74.5 / 71.4	70.4 / 69.6	68.3 / 67.5
● **FlatFormer** (Ours)	2	78.9 / 77.3	72.7 / 71.2	79.1 / 78.6	70.8 / 70.3	81.6 / 78.2	73.8 / 70.5	76.1 / 75.1	73.6 / 72.6
○ CenterPoint [14]	3	–	–	–	–	–	–	–	–
○ CenterPoint++ [15][1]	3	79.1 / 77.6	73.0 /71.6	79.7 / 79.2	71.8 / 71.4	81.5 / 78.6	73.5 / 70.8	76.0 / 75.1	73.7 / 72.8
● SST [2][2]	3	78.1 / 76.3	73.6 / 70.4	75.2 / 74.7	66.5 / 66.1	83.2 / 79.2	76.2 / 72.3	76.0 / 75.1	73.6 / 72.8
● SST-Center [2][†]	3	79.0 / 77.0	72.8 / 71.2	77.0 / 76.4	68.8 / 68.2	82.7 / 78.5	75.8 / 71.8	77.3 / 76.0	74.4 / 73.3
● **FlatFormer** (Ours)	3	79.6 / 78.0	73.5 / 72.0	79.7 / 79.2	71.4 / 71.0	82.0 / 76.1	74.5 / 71.3	77.2 / 76.1	74.7 / 73.7

Table A1. Full results of single-stage 3D detectors on Waymo Open Dataset (validation set). Markers ○ and ● refer to sparse convolutional models and point cloud transformers, respectively. Methods with <60 L2 mAPH are marked gray. ([†]: reproduced by us, [1]: from CenterPoint authors, [2]: from SST authors, [3]: from FSD paper)

A.2. Full Results of Two-Stage 3D Detectors on Waymo Open Dataset

	#Frames	Mean L1 (mAP/APH)	Mean L2 (mAP/APH)	Vehicle L1 (mAP/APH)	Vehicle L2 (mAP/APH)	Pedestrian L1 (mAP/APH)	Pedestrian L2 (mAP/APH)	Cyclist L1 (mAP/APH)	Cyclist L2 (mAP/APH)
○ LiDAR R-CNN [6][†]	1	71.9 / 67.0	65.8 / 61.3	76.0 / 75.5	68.3 / 67.9	71.2 / 58.7	63.1 / 51.7	68.6 / 66.9	66.1 / 64.4
○ PV-RCNN [10][†]	1	73.4 / 69.6	66.8 / 63.3	77.5 / 76.9	69.0 / 68.4	75.0 / 65.6	66.0 / 57.6	67.8 / 66.4	65.4 / 64.0
○ Part-A^2 [12][†]	1	73.6 / 70.3	66.9 / 63.8	77.1 / 76.5	68.5 / 68.0	75.2 / 66.9	66.2 / 58.6	68.6 / 67.4	66.1 / 64.9
○ PV-RCNN++ [11][†]	1	74.8 / 71.0	68.4 / 64.9	78.8 / 78.2	70.3 / 69.7	76.7 / 67.2	68.5 / 59.7	69.0 / 67.6	66.5 / 65.2
○ CenterFormer [16]	1	75.4 / 73.0	71.2 / 69.0	75.2 / 74.7	70.2 / 69.7	78.6 / 73.0	73.6 / 68.3	72.3 / 71.3	69.8 / 68.8
○ FSD-SpConv [3]	1	79.6 / 77.4	72.9 / 70.8	79.2 / 78.8	70.5 / 70.1	82.6 / 77.3	73.9 / 69.1	77.1 / 76.0	74.4 / 73.3
● **FlatFormer**+FSD (Ours)	1	79.4 / 77.1	72.7 / 70.5	78.6 / 78.1	69.8 / 69.4	82.9 / 77.5	74.3 / 69.3	76.6 / 75.6	73.9 / 72.8
○ CenterFormer [16]	2	78.3 / 76.7	74.3 / 72.8	77.0 / 76.5	72.1 / 71.6	81.4 / 78.0	76.7 / 73.4	76.6 / 75.7	74.2 / 73.3
● **FlatFormer**+FSD (Ours)	2	81.4 / 79.9	75.2 / 73.8	79.9 / 79.4	71.4 / 71.0	84.6 / 81.5	76.9 / 73.9	79.8 / 78.8	77.3 / 76.4
○ CenterFormer [16]	4	78.5 / 77.0	74.7 / 73.2	78.1 / 77.6	73.4 / 72.9	81.7 / 78.6	77.2 / 74.2	75.6 / 74.8	73.4 / 72.6
○ MPPNet [1]	4	81.1 / 79.8	75.4 / 74.2	81.5 / 81.1	74.1 / 73.6	84.6 / 81.9	77.2 / 74.7	77.2 / 76.5	75.0 / 74.4
● **FlatFormer**+FSD (Ours)	3	82.2 / 80.7	76.2 / 74.8	80.8 / 80.3	72.5 / 72.1	85.0 / 82.1	77.7 / 74.8	80.7 / 79.8	78.3 / 77.4

Table A2. Full results of two-stage 3D detectors on Waymo Open Dataset (validation set). Markers ○ and ● refer to sparse convolutional models and point cloud transformers, respectively. ([†]: from FSD paper)

A.3. Visualization

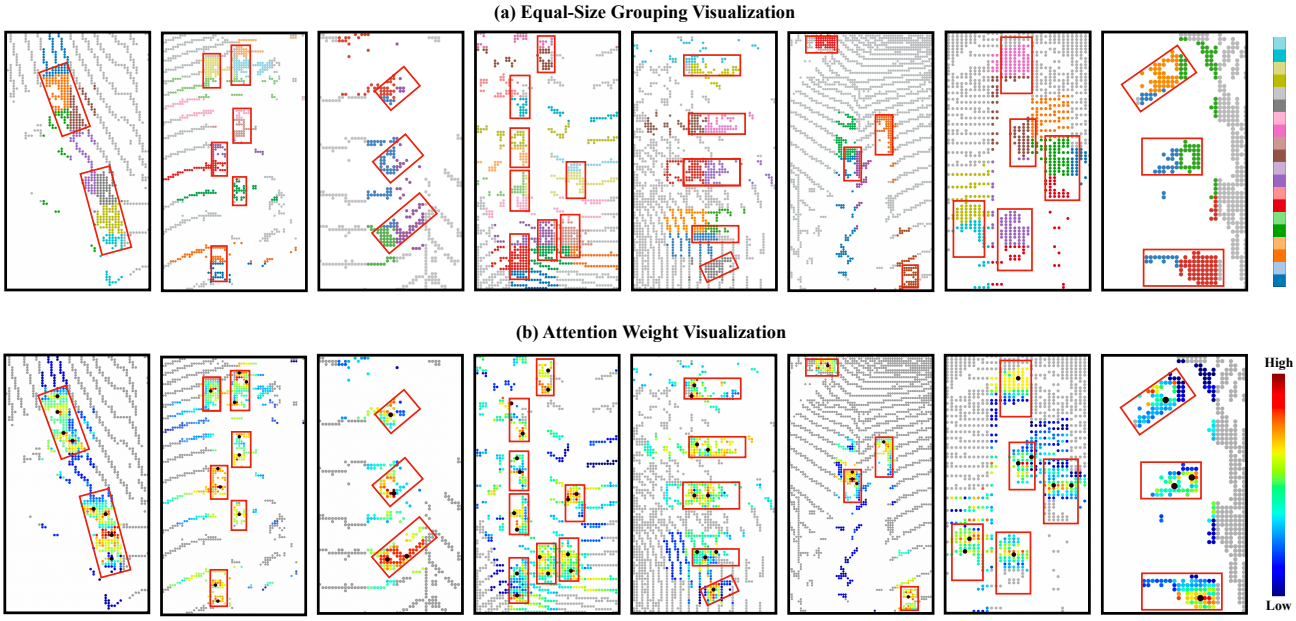

Figure A1. (a): Visualization of equal-size grouping in FlatFormer for vehicles. Different colors represent distinct groups. (b): Visualization of corresponding attention weights in FlatFormer for vehicles. Points on foreground objects have higher attention weights.

To further understand why FlatFormer works, we randomly choose eight local regions around vehicles and visualize our equal-size grouping (Figure A1a) and corresponding attention weights (Figure A1b) of each region in each column. In Figure A1a, we assign different colors to distinct groups. Groups with ≤ 15 points belonging to vehicles are colored gray. For each colored group, we pick one query point on the vehicle, color it black and show the attention weight distribution between the query point and all other points in the same group in Figure A1b. The colors represent the scale of attention weights, where warmer colors indicate larger attention weights. We also visualize the ground truth bounding boxes in red.

As in Figure A1a, the groups are limited to a small spatial region within or around a vehicle. As such, our equal-size grouping avoids interactions between each point and very faraway points. Moreover, Figure A1b shows that query points on the vehicle are usually highly attended to nearby points on the same car, while faraway points have very small learned attention weights. Such an observation can partially explain the effectiveness of FWA: even if equal-size grouping introduces some outlier points in the background to each group (*e.g.*, the second column), FlatFormer can still learn to reduce the importance of these points and lay emphasis on important foreground points within each group.

A.4. Distance-Conditioned Vehicle Detection Performance

	#Frames	Vehicle L1 mAPH				Vehicle L2 mAPH			
		0-30m	30-50m	50m-inf	Overall	0-30m	30-50m	50m-inf	Overall
CenterPoint [14][1]	1	91.0	72.5	50.2	74.2	90.1	66.6	39.1	66.2
SST-Center [2]	1	91.6	73.0	50.0	74.6	90.3	66.4	38.5	66.2
FlatFormer (Ours)	1	92.5 (+0.9)	75.3 (+2.3)	54.1 (+3.9)	77.1 (+2.5)	91.2 (+1.1)	68.8 (+2.2)	42.0 (+2.9)	68.6 (+2.4)

Table A3. Distance-conditioned vehicle detection performance among CenterPoint, SST-Center and FlatFormer on Waymo Open Dataset (validation set). FlatFormer significantly boosts the performance of CenterPoint and SST-Center in long-range metrics. ([1]: from CenterPoint authors)

Table A3 shows that FlatFormer significantly outperforms SST and CenterPoint in long-range perception. Compared with SST-Center, our equal-size grouping ensure that each point is attended to a fixed number of points. Equal-window grouping in SST, in contrast, creates groups with very few points in faraway regions, where the point cloud is much sparser. This difference can partially explain our advantages on long-range metrics: SST almost degenerates to MLP/PointNet [8] in faraway regions, providing limited capability in modeling distant objects.

References

[1] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li. MPPNet: Multi-Frame Feature Intertwining with Proxy Points for 3D Temporal Object Detection. In *ECCV*, 2022. 1

[2] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing Single Stride 3D Object Detector with Sparse Transformer. In *CVPR*, 2022. 1, 2

[3] Lue Fan, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Fully Sparse 3D Object Detection. In *NeurIPS*, 2022. 1

[4] Chenhang He, Ruihuang Li, Shuai Li, and Lei Zhang. Voxel Set Transformer: A Set-to-Set Approach to 3D Object Detection from Point Clouds. In *CVPR*, 2022. 1

[5] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, and Jiong Yang. PointPillars: Fast Encoders for Object Detection from Point Clouds. In *CVPR*, 2019. 1

[6] Zhichao Li, Feng Wang, and Naiyan Wang. LiDAR R-CNN: An Efficient and Universal 3D Object Detector. In *CVPR*, 2021. 1

[7] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel Transformer for 3D Object Detection. In *ICCV*, 2021. 1

[8] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *CVPR*, 2017. 3

[9] Guangsheng Shi, Ruifeng Li, and Chao Ma. PillarNet: Real-Time and High-Performance Pillar-based 3D Object Detection. In *ECCV*, 2022. 1

[10] 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 *CVPR*, 2020. 1

[11] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN++: Point-Voxel Feature Set Abstraction With Local Vector Representation for 3D Object Detection. *arXiv*, 2021. 1

[12] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From Points to Parts: 3D Object Detection from Point Cloud with Part-aware and Part-aggregation Network. *TPAMI*, 2020. 1

[13] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 2018. 1

[14] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D Object Detection and Tracking. In *CVPR*, 2021. 1, 2

[15] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. CenterPoint++ Submission to the Waymo Real-time 3D Detection Challenge. Technical report, 2022. 1

[16] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. CenterFormer: Center-based Transformer for 3D Object Detection. In *ECCV*, 2022. 1