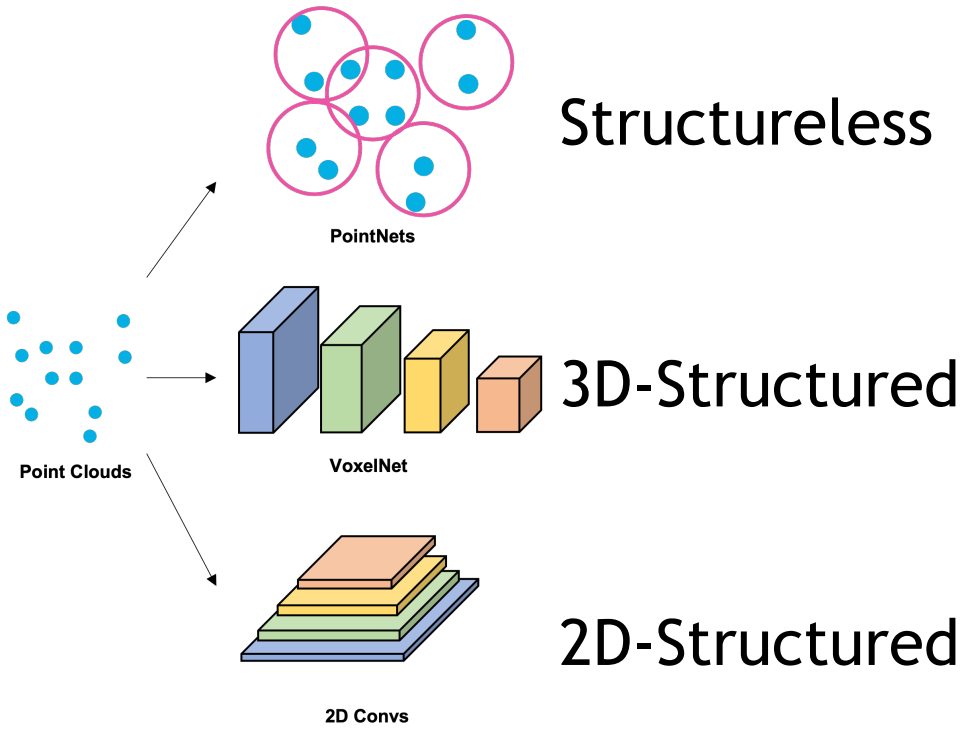
M3DETR: Multi-representation, Multi-scale, Mutual-relation 3D Object Detection with Transformers

Tianrui Guan*, Jun Wang*, Shiyi Lan, Rohan Chandra, Zuxuan Wu,
Larry S. Davis, Dinesh Manocha

University of Maryland Fudan University

Motivation

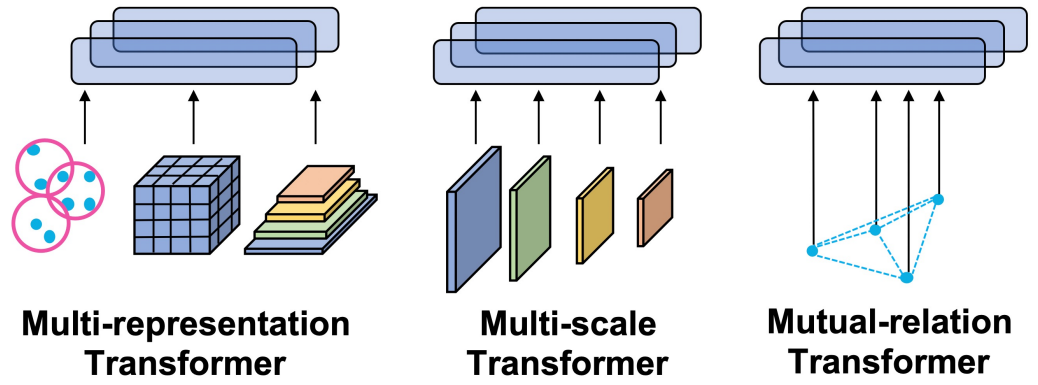


Structureless

3D-Structured

2D-Structured

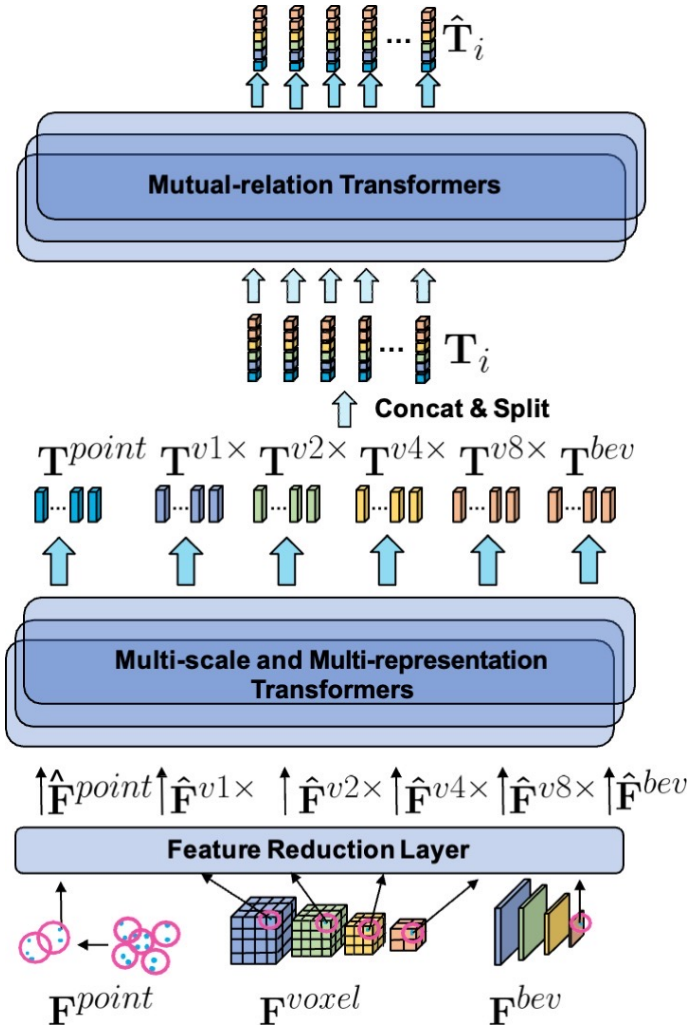
Representation Gap



M3 Transformer

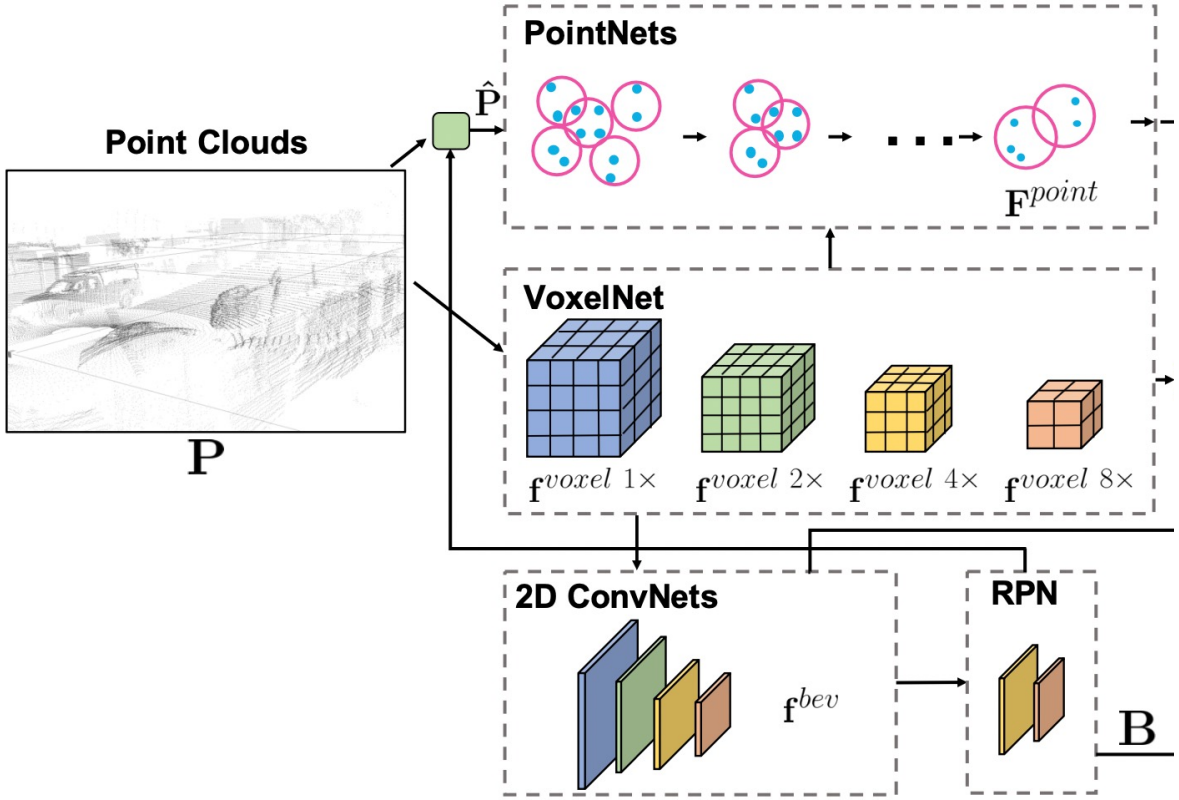
Main Contribution

- First unified architecture for 3D object detection with transformers that accounts for multi-representation, multi-scale, mutual-relation models of point clouds in an end-to-end manner.
- M3DETR is robust and insensitive to the hyper-parameters of transformer architectures.
- Our unified architecture achieves SOTA on KITTI and Waymo Open Dataset.

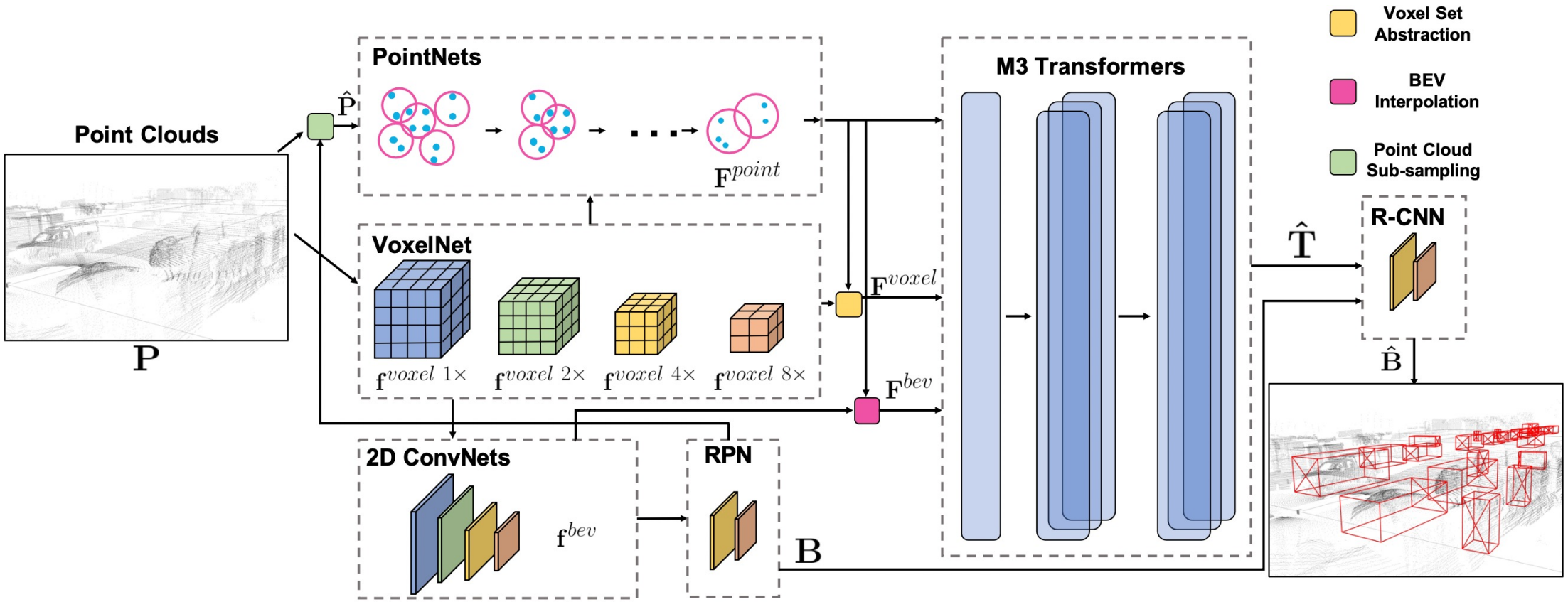


Detailed M3 Transformer

Proposed Architecture



Proposed Architecture



Results on Waymo Test Set

Vehicle, pedestrian, and cyclist classes:

Method	Vehicle				Pedestrian				Cyclist				All			
	L1 mAP	L1 mAPH	L2 mAP	L2 mAPH	L1 mAP	L1 mAPH	L2 mAP	L2 mAPH	L1 mAP	L1 mAPH	L2 mAP	L2 mAPH	L1 mAP	L1 mAPH	L2 mAP	L2 mAPH
PV-RCNN* [39]	76.89	76.30	67.99	67.46	65.43	55.85	59.56	50.74	66.38	64.68	63.42	61.81	69.57	65.61	63.65	60.00
M3DETR	77.75	77.17	70.63	70.06	68.10	58.87	60.57	52.37	67.28	65.69	65.31	63.75	71.05	67.09	65.50	61.92
Improvement	**+0.86**	**+0.87**	**+2.64**	**+2.6**	**+2.67**	**+3.02**	**+1.01**	**+1.63**	**+0.9**	**+1.01**	**+1.89**	**+1.94**	**+1.48**	**+1.48**	**+1.85**	**+1.92**

Vehicle class with different difficulty:

Method	3D mAP LEVEL_1				3D mAPH LEVEL_1				3D mAP LEVEL_2				3D mAPH LEVEL_2			
	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf
RCD [1]	69.59	87.2	67.8	46.1	-	-	-	-	-	-	-	-	-	-	-	-
StarNet [30]	61.50	82.20	56.60	32.20	61.00	81.70	56.00	31.80	54.9	81.3	49.5	23.0	54.50	80.80	49.00	22.70
PointPillars [19]	68.62	87.20	65.50	40.92	68.08	86.71	64.87	40.19	65.21	87.93	63.80	38.20	64.29	87.30	62.36	36.87
Det3D [63]	73.29	90.31	70.54	49.10	72.27	89.65	68.96	47.45	65.21	87.93	63.80	38.20	64.29	87.30	62.36	36.87
RangeDet [10]	75.83	88.41	73.83	55.31	75.38	87.95	73.38	54.84	67.12	87.53	67.99	44.40	66.73	87.08	67.58	44.01
PV-RCNN* [39]	76.89	92.27	75.51	55.35	76.30	91.82	74.79	54.27	67.99	89.18	69.39	42.80	67.46	88.75	68.70	41.95
M3DETR	**77.75**	**92.54**	**76.35**	**57.52**	**77.17**	**92.09**	**75.71**	**56.35**	**70.63**	**89.43**	**70.26**	**45.88**	**70.06**	**89.01**	**69.65**	**44.88**
Improvement	**+0.86**	**+0.27**	**+0.84**	**+2.17**	**+0.87**	**+0.27**	**+0.92**	**+1.51**	**+2.64**	**+0.25**	**+0.87**	**+1.48**	**+2.6**	**+0.26**	**+0.95**	**+0.87**

Results on Waymo Validation Set

Vehicle class with different difficulty:

	3D mAP LEVEL_1				3D mAPH LEVEL_1				3D mAP LEVEL_2				3D mAPH LEVEL_2			
Method	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf	Overall	0-30m	30-50m	50m-Inf
LaserNet [28]	52.11	70.90	52.90	29.60	50.05	68.70	51.40	28.60	-	-	-	-	-	-	-	-
PointPillars [19]	56.62	81.00	51.80	27.90	-	-	-	-	-	-	-	-	-	-	-	-
RCD [1]	69.59	87.20	67.80	46.10	69.16	86.80	67.40	45.50	-	-	-	-	-	-	-	-
RangeDet [10]	72.85	87.96	69.03	48.88	-	-	-	-	-	-	-	-	-	-	-	-
PV-RCNN [39]	70.30	91.90	69.20	42.20	69.69	91.34	68.53	41.31	65.36	91.58	65.13	36.46	64.79	91.00	64.49	35.70
M3DETR	**75.71**	**92.69**	**73.65**	**52.96**	**75.08**	**92.22**	**72.94**	**51.80**	**66.58**	**91.92**	**65.73**	**40.44**	**66.02**	**91.45**	**65.10**	**39.52**
Improvement	**+2.86**	**+0.79**	**+4.45**	**+3.98**	**+5.39**	**+0.88**	**+4.41**	**+6.3**	**+1.22**	**+0.34**	**+0.6**	**+3.94**	**+1.23**	**+0.45**	**+0.61**	**+3.82**

[19] PointPillars: Fast Encoders for Object Detection from Point Clouds, CVPR 2019

[39] PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection, CVPR 2020

Results on KITTI Test Set

Vehicle and Cyclist classes with different difficulty:

Method	Car			Cyclist		
	Easy	Mod	Hard	Easy	Mod	Hard
F-PointNet [34]	72.27	56.12	49.01	82.19	69.79	60.59
VoxelNet [62]	77.47	65.11	57.73	61.22	48.36	44.37
SECOND [52]	83.34	72.55	65.82	75.83	60.82	53.67
PointPillars [19]	82.58	74.31	68.99	77.10	58.65	51.92
PointRCNN [40]	86.96	75.64	70.70	74.96	58.82	52.53
STD [56]	87.95	79.71	75.09	78.69	61.59	55.30
HotSpotNet [5]	87.60	78.31	73.34	82.59	65.95	59.00
PVRCNN [39]	90.25	81.43	76.82	78.60	63.71	57.65
M3DeTR	**90.28**	**81.73**	**76.96**	**83.83**	**66.74**	**59.03**
Improvement	**+0.03**	**+0.3**	**+0.14**	**+1.24**	**+0.79**	**+0.03**

[56] Std: Sparse-to-dense 3d object detector for point cloud, CVPR 2019

[5] Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots, ECCV 2020

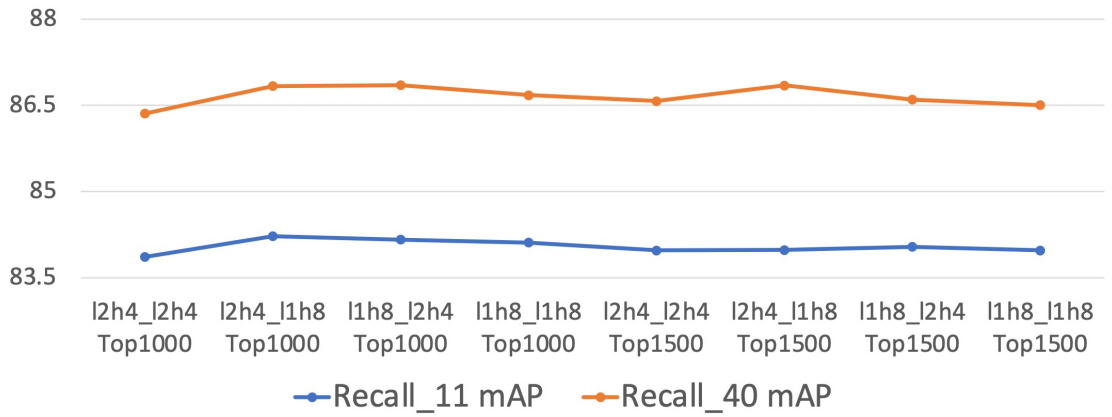
[39] PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection, CVPR 2020

Ablation Studies

Rel. Trans.	Rep. and Scal. Trans.	Recall_11 Easy	Recall_11 Mod	Recall_11 Hard	Recall_40 Easy	Recall_40 Mod	Recall_40 Hard
x	x	88.66	79.07	78.49	91.17	82.61	82.06
✓	x	88.82	83.23	78.64	91.37	84.40	82.34
x	✓	88.93	83.63	78.59	91.72	84.68	82.39
✓	✓	**89.28**	**84.16**	**79.05**	**92.29**	**85.41**	**82.85**

Ablation studies of transformers in Car class on KITTI:

- Rel. Trans: mutual-relation transformer
- Rep. and Scal. Trans: multi-representation and multiscale transformer



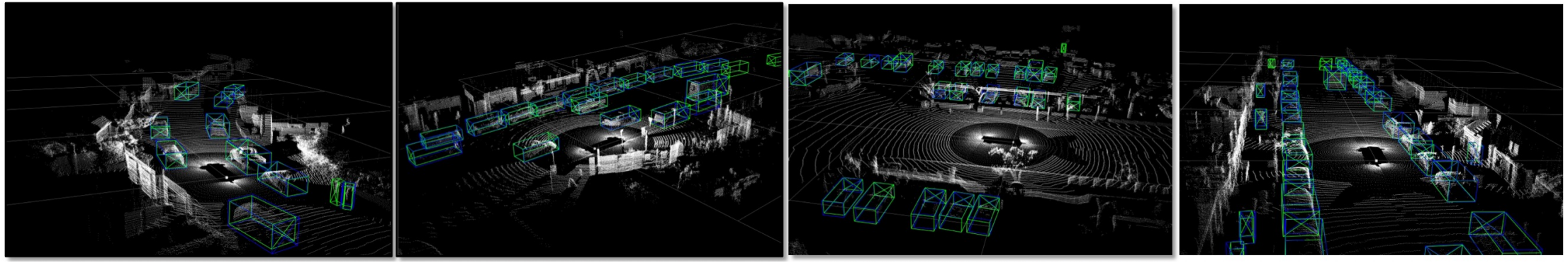
Robustness of M3DETR:

Note that ”l” and ”h” represent layer number and head dimension of M3 Transformers, respectively.

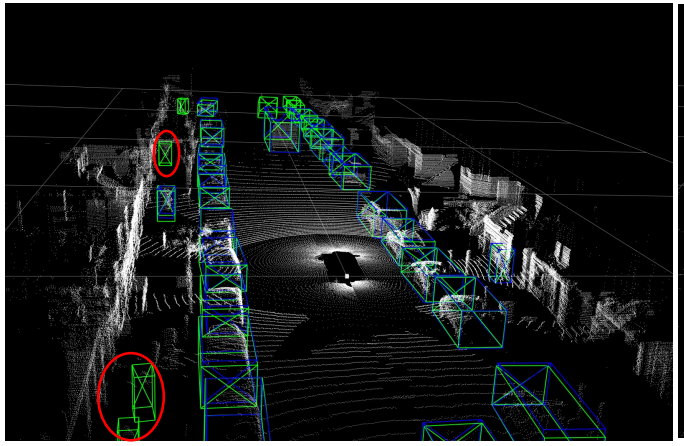
”Top” denotes the number of proposals used for keypoint sampling from RPN stage.

Qualitatively Results

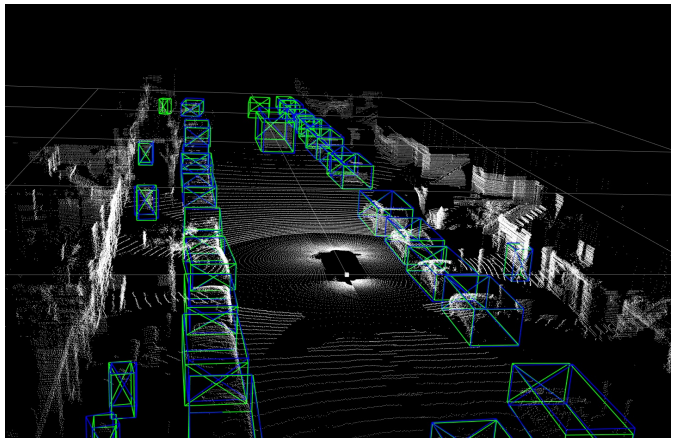
3D detection results of M3DeTR on the Waymo Dataset. The 3D ground truth bounding boxes are in green, while the detection bounding box are in blue.



PV-RCNN



M3DeTR



Thank You!