Pillar-based Object Detection for Autonomous Driving

Anonymous ECCV submission

Paper ID 3892

Supplementary Material

Longer training

One recent work [1] shows object detection models exhibit better performance if it is trained longer. We find the methods we compared to use different training scheduling: StarNet [3] is trained for 75 epochs; MVF [5] and PointPillars [2] re-implemented by [5] are trained for 100 epochs, while ours in the main text is only trained for 30 epochs. To have a fair comparison to these methods and investigate the effect of training schedule, we report the results of the model trained for 75 epochs. Table 1 (for vehicles) and Table 2 (for pedestrians) show that when being trained longer, the proposed method has additional performance boosting. Also, compared to MVF [5] which is trained for 100 epochs, our method improves by 6.19 3D mAP for vehicle and 5.32 3D mAP for pedestrian.

Parameter Specification

In this section, we provide details on the parameters of the model. The model consists of three parts: a multi-view feature learning network; a birds-eye view pillar backbone network; and a detection head. We show the pipeline in Figure 1 and the additional parameter specification in Table 3.

Method	BEV mAP (IoU=0.7)				3D mAP (IoU=0.7)			
	Overall	0 - 30m	30 - 50m	50m - Inf	Overall	0 - 30m	30 - 50m	50m - Inf
StarNet [3] (75 epochs)	-	-	-	-	53.7	-	-	-
PointPillars¶ [2]	80.4	92.0	77.6	62.7	62.2	81.8	55.7	31.2
PointPillars† [2] (100 epochs)	75.57	92.1	74.06	55.47	56.62	81.01	51.75	27.94
MVF [5] (100 epochs)	80.4	93.59	79.21	63.09	62.93	86.3	60.2	36.02
Ours (30 epochs)	**86.14**	**95.59**	**83.62**	**70.85**	**67.71**	**87.46**	**64.38**	**39.88**
Ours (75 epochs)	**86.65**	**95.41**	**84.51**	**71.19**	**69.12**	**88.26**	**65.67**	**41.44**
Improvements (over MVF)	**+6.25**	**+1.82**	**+5.30**	**+8.1**	**+6.19**	**+1.96**	**+5.47**	**+5.42**

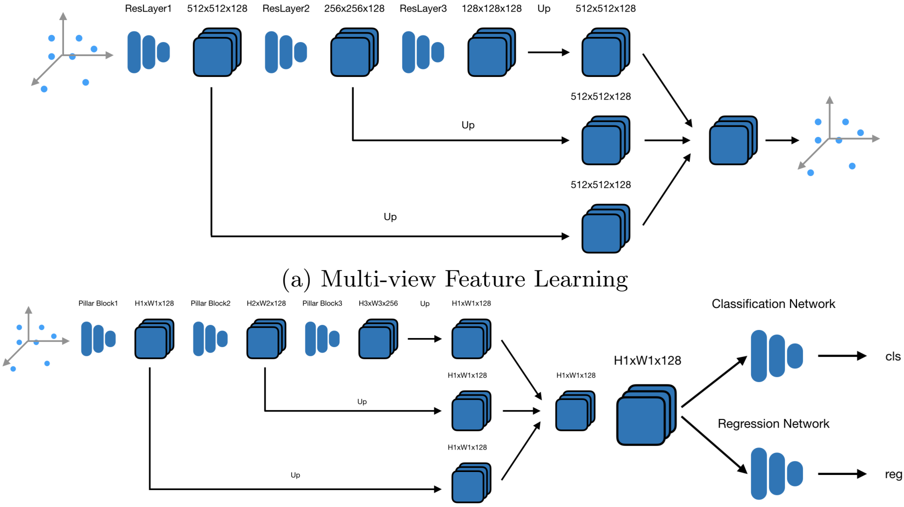
Table 1. Results on vehicle. ¶: re-implemented by [4], the feature map in the first PointPillars block is two times as big as in others; †: re-implemented by [5].

Method	BEV mAP (IoU=0.7)				3D mAP (IoU=0.7)			
	Overall	0 - 30m	30 - 50m	50m - Inf	Overall	0 - 30m	30 - 50m	50m - Inf
StarNet [3] (75 epochs)	-	-	-	-	66.8	-	-	-
PointPillars¶ [2]	68.7	75.0	66.6	58.7	60.0	68.9	57.6	46.0
PointPillars† [2] (100 epochs)	68.57	75.02	67.11	53.86	59.25	67.99	57.01	41.29
MVF [5] (100 epochs)	74.38	80.01	72.98	62.51	65.33	72.51	63.35	50.62
Ours (30 epochs)	**76.45**	**82.42**	**74.38**	**64.91**	**67.69**	**75.42**	**64.88**	**51.48**
Ours (75 epochs)	**76.85**	**82.9**	**77.15**	**63.82**	**70.65**	**77.87**	**70.37**	**54.48**
Improvements (over MVF)	**+2.47**	**+2.89**	**+4.17**	**+1.31**	**+5.32**	**+5.36**	**+7.02**	**+3.86**

Table 2. Results on pedestrian. ¶: re-implemented by [4]. †: re-implemented by [5].

Stage	Vehicle Model		Pedestrian Model	
	Kernel	Output Size	Kernel	Output Size
Multi-view Feature Learning	3x3, 128, stride 1	512x512x128	3x3, 128, stride 1	512x512x128
	3x3, 128, stride 2	256x256x128	3x3, 128, stride 2	256x256x128
	3x3, 128, stride 2	128x128x128	3x3, 128, stride 2	128x128x128
Pillar Backbone Block1	3x3, 128, stride 2	256x256x128	3x3, 128, stride 1	512x512x128
	{3x3, 128, stride 1}x3	256x256x128	{3x3, 128, stride 1}x3	512x512x128
Pillar Backbone Block2	3x3, 128, stride 1	256x256x128	3x3, 128, stride 2	256x256x128
	{3x3, 128, stride 1}x5	256x256x128	{3x3, 128, stride 1}x5	256x256x128
Pillar Backbone Block3	3x3, 256, stride 2	128x128x256	3x3, 256, stride 2	128x128x256
	{3x3, 256, stride 1}x5	128x128x256	{3x3, 256, stride 1}x5	128x128x256
Detection Head	{3x3, 256, stride 1}x4	256x256x256	{3x3, 256, stride 1}x4	512x512x256

Table 3. Parameters of convolutional kernels and feature map sizes.



(a) Multi-view Feature Learning

(b) Pillar Backbone

(c) Detection Head

Fig. 1. Details of the proposed model: (a) the multi-view feature learning module, we show the network for one view; (b) Pillar backbone network; (c) the detection head, we show both the classification network and the regression network. For details on the parameters and the feature map sizes, refer to Table 3.

References

1. He, K., Girshick, R.B., Dollár, P.: Rethinking imagenet pre-training (2019)
2. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
3. Ngiam, J., Caine, B., Han, W., Yang, B., Chai, Y., Sun, P., Zhou, Y., Yi, X., Alsharif, O., Nguyen, P., Chen, Z., Shlens, J., Vasudevan, V.: Starnet: Targeted computation for object detection in point clouds. ArXiv **abs/1908.11069** (2019)
4. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tak-wing Tsui, P., Guo, J.C.Y., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z.F., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. ArXiv **abs/1912.04838** (2019)
5. Zhou, Y., Sun, P., Zhang, Y., Anguelov, D., Gao, J., Ouyang, T., Guo, J., Ngiam, J., Vasudevan, V.: End-to-end multi-view fusion for 3d object detection in lidar point clouds. In: The Conference on Robot Learning (CoRL) (2019)