PointPainting: Sequential Fusion for 3D Object Detection

Supplementary Material

A. PointPainting: 3D results

In this section, we present 3D results of PointPainting on the KITTI validation and test sets.

Validation Set Similar to bird's-eye view results (Table 1), we see that PointPainting substantially improves 3D detection performance on the validation set. As seen in Table 7, 23 out of 27 comparisons (3 experiments x 3 classes x 3 strata) were improved by PointPainting.

Test Set In the test set (Table 8), we observe that PointPainting consistently improves 3D detection results of PointRCNN on the pedestrians and cyclists classes across all difficulty strata (easy, medium and hard). However, we see that the 3D results on the car class drops substantially. We think this could be because of overfitting on our minival set which was very small (see Section 3.1).

B. PointPainting Latency

In Section 5.2, we concluded that Consecutive matching (see Figure 6) can minimize the latency introduced by PointPainting without any drop in detection performance. Here we provide a detailed breakdown of the latency introduced by PointPainting in the case of Consecutive matching.

Projection This step involves transforming the pointcloud to the coordinate system of the ego-vehicle in the previous frame followed by a projection into the camera images to get the segmentation scores. This operation only adds a latency of 0.15 ms.

Encoding The Painted PointPillars encoder operates on an 18 dimensional decorated pointcloud as opposed to the 7 dimensional pointcloud in the original PointPillars architecture. We measure the runtimes for both the encoders in TensorRT and find that PointPainting adds an additional latency of 0.6 ms in the encoding stage.

Thus, Painted PointPillars only introduces an additional latency of 0.75 ms over PointPillars when Consecutive matching is used. This makes Painted PointPillars a strong candidate for realtime camera-lidar fusion.

Class	Recall (%)	Precision (%)
Car	94	89
Bus	71	92
Construction Vehicle	40	58
Trailer	39	79
Truck	69	76
Motorcycle	89	87
Bicycle	58	84
Pedestrian	80	86
Barrier	81	80
Traffic Cone	78	84

Table 6. Class-wise Precision and Recall of the semantic segmentation network trained on the nuImages dataset.

C. nuImages Semantic Segmentation

Here we present some stats on the semantic segmentation network that we trained on the nuImages dataset. The mean intersection over union (mIoU) on the validation set was 0.65. The class-wise precision and recall on the validation set is shown in Table 6. Our model performs the best on the car class and worst on the construction vehicle and trailer classes.

Method	mAP	Car			Pedestrian			Cyclist		
	Mod.	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PointPillars [11]	66.96	87.22	76.95	73.52	65.37	60.66	56.51	82.29	63.26	59.82
Painted PointPillars	69.03	86.26	76.77	70.25	71.50	66.15	61.03	79.12	64.18	60.79
Delta	+2.07	-0.96	-0.18	-3.27	**+6.13**	**+5.49**	**+4.52**	-3.17	+0.92	+0.97
VoxelNet [34, 29]	67.12	86.85	76.64	74.41	67.79	59.84	52.38	84.92	64.89	58.59
Painted VoxelNet	68.01	87.15	76.66	74.75	68.57	60.93	54.01	85.61	66.44	64.15
Delta	+0.89	+0.3	+0.02	+0.34	+0.78	+1.09	+1.63	+0.69	+1.55	**+5.56**
PointRCNN [21]	67.01	86.75	76.05	74.30	63.29	58.32	51.59	83.68	66.67	61.92
Painted PointRCNN	70.34	88.38	77.74	76.76	69.38	61.67	54.58	85.21	71.62	66.98
Delta	**+3.33**	**+1.63**	**+1.69**	**+2.46**	+6.09	+3.35	+2.29	**+1.53**	**+4.95**	+5.06

Table 7. PointPainting applied to state of the art lidar based object detectors. All lidar methods show an improvement in 3D mean average precision (mAP) of car, pedestrian, and cyclist on KITTI *validation* set, moderate split.

Method	Modality	mAP	Car			Pedestrian			Cyclist		
		Mod.	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MV3D[3]	L & I	N/A	74.97	63.63	54.00	N/A	N/A	N/A	N/A	N/A	N/A
AVOD-FPN[10]	L & I	54.86	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
F-PointNet[18]	L & I	56.02	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
F-ConvNet[26]	L & I	**61.61**	87.36	76.39	66.69	52.16	43.38	38.80	**81.98**	**65.07**	**56.54**
MMF[13]	L, I & M	N/A	**88.40**	77.43	70.22	N/A	N/A	N/A	N/A	N/A	N/A
PointPillars[11]	L	58.29	82.58	74.31	68.99	51.45	41.92	**38.89**	77.10	58.65	51.92
STD[32]	L	61.25	87.95	**79.71**	75.09	**53.29**	**42.47**	38.35	78.69	61.59	55.30
PointRCNN[21]	L	57.94	86.96	75.64	**70.70**	47.98	39.37	36.01	74.96	58.82	52.53
Painted PointRCNN	L & I	58.82	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89
Delta	Δ I	**+0.88**	-4.85	-3.94	-3.62	**+2.34**	**+1.6**	**+1.86**	**+2.67**	**+4.96**	**+3.36**

Table 8. Results on the KITTI test 3D detection benchmark. The modalities are lidar (L), images (I), and maps (M). The delta is the difference due to Painting, ie Painted PointRCNN minus PointRCNN. We don't include a few entries from Table 2 because LaserNet[17] did not publish 3D results and SECOND[29], IPOD[31] no longer have their entries on the public leaderboard since KITTI changed to a 40 point interpolated AP metric instead of 11.