Supplementary Material for “VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection”

Yin Zhou
Apple Inc
yzhou3@apple.com

Oncel Tuzel
Apple Inc
otuzel@apple.com

1. Random sampling threshold

The main purpose of subsampling points, for the voxels that have more than a predefined maximum number of points, is to form an efficient tensor representation for computation on a GPU. In our experiments, the maximum number of points is set to a large enough value, $T = 35$ for car detection and 45 for pedestrian/cyclist detection, such that the aggregated statistics within the voxels, e.g. mean or other pooling operations, are close to the true statistics. On the datasets presented, only 0.17% of non-empty voxels contain more points than our threshold. For the pedestrian/cyclist detection, increasing the maximum number of points from $T = 35$ to $T = 45$ resulted in $\sim 1\%$ AP improvement.

2. Ablation Study

Comparing VFE vs. hand-crafted features

To fully demonstrate the effectiveness of our feature learning network, we present experimental results obtained by connecting VFE with the 2D convolution architecture of our hand-crafted baseline, *i.e.,* VFE + 2D Conv. As shown in Table 1, VFE + 2D Conv significantly improves over the hand-crafted features + 2D Conv baseline. Our full model (VFE + 3D Conv) has $\sim 1\%$ higher AP than VFE + 2D Conv, demonstrating that the proposed 3D feature aggregation is also helpful.

Contribution of data augmentation

For completeness, we study the individual contribution from each of the proposed data augmentation strategies. As listed in Table 2, for the 3D detection task, scaling and rotation transformations improve by $\sim 1\%$ each, and the box perturbation improves by 1 - 4% over no-augmentation. The improvements are not additive and the combined effect of all augmentations is approximately 1 - 5%. For our experiments, the same augmentations are applied to the hand-crafted feature baseline.

Method	Birds Eye View			3D Detection		
	Easy	Moderate	Hard	Easy	Moderate	Hard
Hand-crafted + 2D Conv	88.26	78.42	77.66	71.73	59.75	55.69
VFE + 2D Conv	89.11	83.67	78.23	80.70	64.77	62.39

Table 1. Comparison of features for car detection on KITTI val set.

Method	Birds Eye View			3D Detection		
	Easy	Moderate	Hard	Easy	Moderate	Hard
No augmentation	87.89	79.43	76.80	76.31	62.98	55.88
Scaling	88.85	80.31	77.28	77.12	63.14	56.81
Rotation	89.55	82.58	78.39	78.28	63.71	56.86
Box perturbation	89.05	82.43	78.09	81.44	64.89	62.15
All	**89.60**	**84.81**	**78.57**	**81.97**	**65.46**	**62.85**

Table 2. Ablation study on data augmentation for car detection on KITTI val set.