PointAugmenting: Cross-Modal Augmentation for 3D Object Detection —Supplementary Material—

Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{weiwei0224, chaoma, droplet-to-ocean, xkyang}@sjtu.edu.cn

In this supplementary material, we provide more details and experimental results to complement the manuscript.

A. Data Augmentation for LiDAR Points

For LiDAR points $\mathcal{P}$ in the scene, we transform the point $p = (x, y, z)$ into the LiDAR spherical coordinate system as (r, θ, ϕ) by the following equations:

$$r = \sqrt{x^2 + y^2 + z^2},\ \theta = \arccos(\frac{z}{r}),\ \phi = \arctan(\frac{y}{x}) \tag{1}$$

Given an object and its ground-truth box with eight corners $\mathcal{C} = \{(r_k, \theta_k, \phi_k), k = 1, 2, ..., 8\}$, the perspective V of an object is represented by the range of θ and ϕ, where

$$\begin{aligned} V &= ([\theta_{min}, \theta_{max}], [\phi_{min}, \phi_{max}]) \\ \theta_{min}, \theta_{max} &= \min(\theta_k), \max(\theta_k), k = 1, 2, ..., 8 \\ \phi_{min}, \phi_{max} &= \min(\phi_k), \max(\phi_k), k = 1, 2, ..., 8 \end{aligned} \tag{2}$$

For a LiDAR point $p = (r_i, \theta_i, \phi_i) \in \mathcal{P}$, if $\theta_i \in [\theta_{min}, \theta_{max}]$ and $\phi_i \in [\phi_{min}, \phi_{max}]$, we denote the point p in the perspective of this object.

When pasting virtual objects into current training scene, we restrict the perspective overlap between objects. Specifically, we denote the perspective area of an object as $\Delta\theta \times \Delta\phi$, where $\Delta\theta = \theta_{max} - \theta_{min}$ and $\Delta\phi = \phi_{max} - \phi_{min}$. For a candidate virtual object M, it will be pasted if its perspective IoU with any object N in current scene is within a threshold T as illustrated in Equation 3. In experiments, we set T to be 0.7.

$$V_{IoU} = \frac{Area(M \cap N)}{Min(Area(M), Area(N))} \leq T \tag{3}$$

B. Discussion of 2D Network.

In our PointAugmenting, we employ CNN features from pretrained 2D detection networks as image representation to fuse with LiDAR points for 3D object detection. In this section, additional discussion about the choice and training scheme of 2D network are provided.

Methods	Pretrained 2D model	Joint training	mAP
CenterPoint	-	-	37.6
PointAugmenting	-	✓	38.3
PointAugmenting	✓	-	47.3
PointAugmenting	✓	✓	47.7

Table S1. Training schemes on the 2D network. Using pretrained 2D detection network and finetuning the 2D network (joint training) with the supervision of 3D labels achieve the best mAP.

Pretrained 2D network. Table S1 shows that the pretrained 2D model facilitates overall 3D detection accuracy. This is due to the straightforward semantics provided for point clouds. Without the pretrained model, image features are implicitly learned from 3D labels from scratch, leading to significant performance drop.

Joint training. Jointly training 2D and 3D networks together could achieve better results (+0.4% mAP) than freezing the pretrained model. With the use of the supervision information from 3D labels, the 2D network generates better image features, but this consumes larger memory and longer time during training.

Pretrained ImageNet. During our experiments, we find that a pretrained ImageNet classification network can also boost 3D detection by providing semantic information. Due to the domain gap between datasets and tasks, using the pretrained classification network yields a performance drop of -2.1% in comparison with the detection counterpart on the 1/8 nuScenes dataset.

C. Additional Experimental Results

We visualize the qualitative results of our PointAugmenting on the nuScenes and Waymo datasets. Figure S1 and Figure S2 show that our approach successfully detects objects in the challenging scenes, where red boxes denote ground truth and yellow boxes denote our predictions.

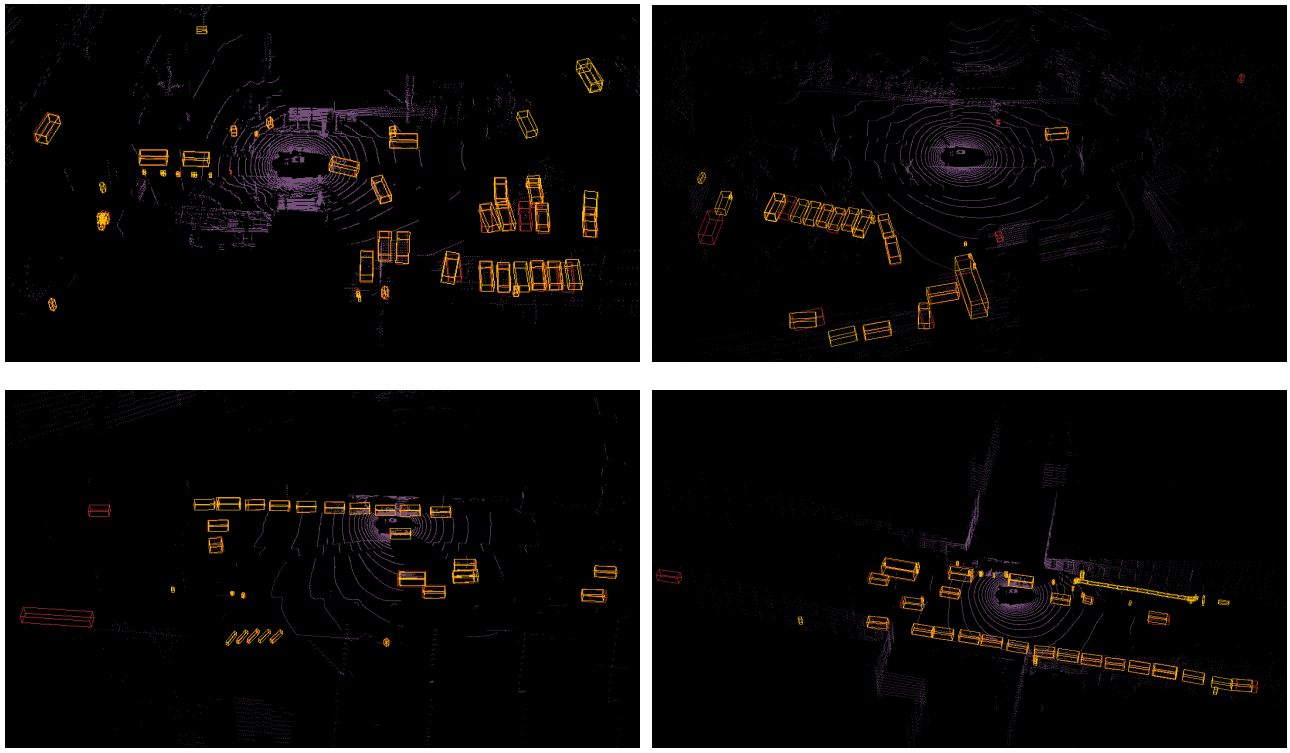

Figure S1. Qualitative results on the nuScenes dataset. Red: Ground Truth. Yellow: Predictions by our PointAugmenting.

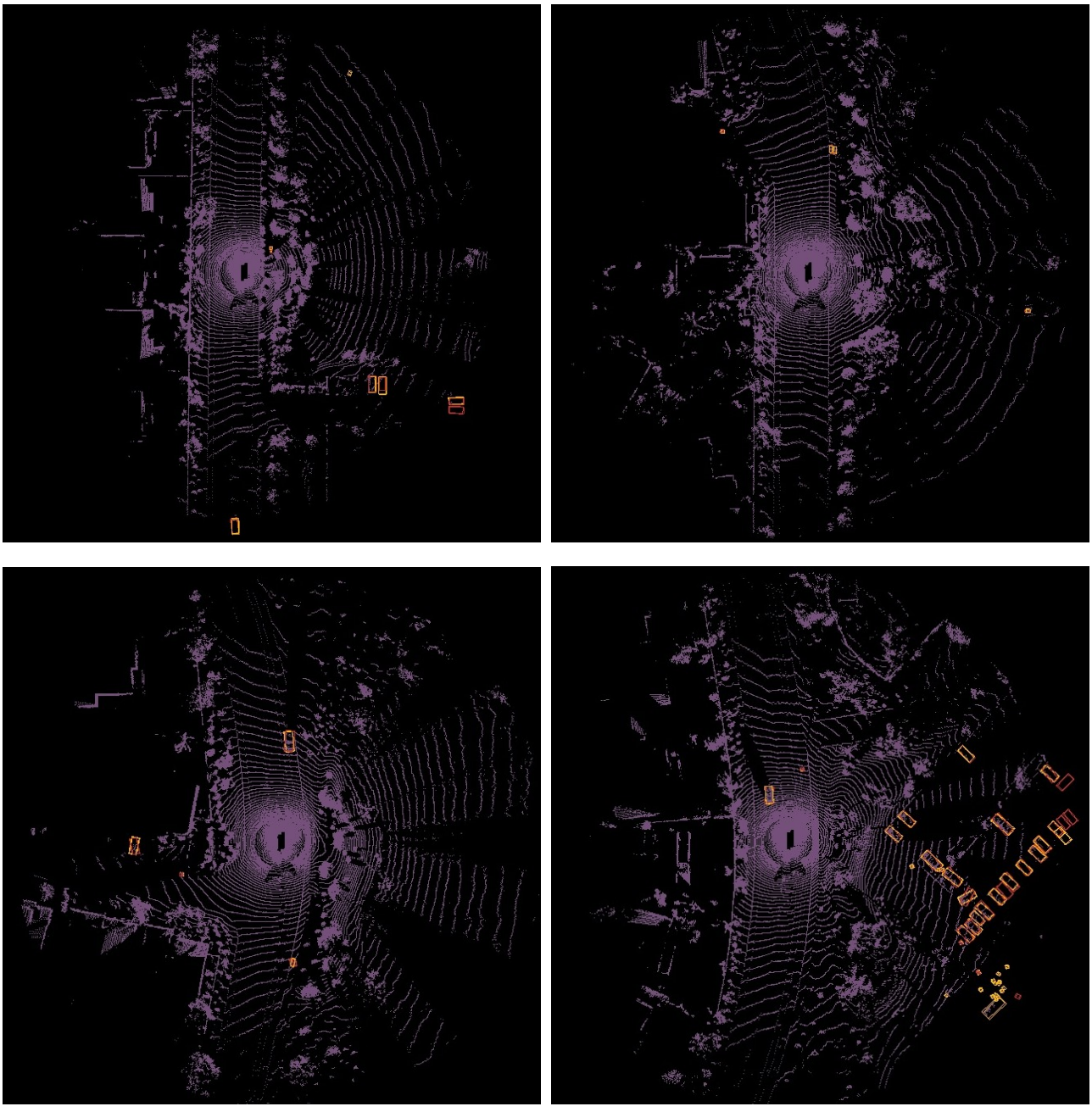

Figure S2. Qualitative results on the Waymo dataset. Red: Ground Truth. Yellow: Predictions by our PointAugmenting.