Supplementary Material for EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection

Tengteng Huang[1⋆], Zhe Liu[1⋆], Xiwu Chen[1], and Xiang Bai[1⋆⋆]

Huazhong University of Science and Technology
{huangtengtng, zheliu1994, xiwuchen, xbai}@hust.edu.cn

1 More Quantitative Results on KITTI

Ablation analysis of the recall. In the main manuscript, we have provided the quantitative analysis of the impact of our LI-Fusion module and CF loss in terms of mAP. We further present the results in terms of recall to better understand how these components influence the 3D detection performance.

In Table 1, we give the results under two different settings with the IoU thresholds of 0.5 and 0.7, respectively. Under the IoU threshold of 0.7, our LI-Fusion module leads to an improvement of 3.10%. It indicates that fusing the semantic image features, which contains sufficient information (*e.g.*, color, shape, *etc.*), is important for detecting hard objects that LiDAR-based methods fail to detect. Besides, the model trained with the CF loss yields a recall of 70.69%, outperforming the result of the baseline model by 8.77%. The potential reason is that the CF loss effectively guarantees the consistency between the localization and classification confidence, and thus the boxes with large overlaps are well reserved. Our EPNet, which integrates both the LI-Fusion module and the CF loss, shows a promising recall of 71.92%, surpassing the baseline model by 10.00%. We observe consistent results under the IoU threshold of 0.5. Based on these quantitative results, we can conclude that 3D detection can benefit a lot from the proposed LI-Fusion module and CF loss.

Table 1. Quantitative analysis of the recall for our proposed method on the KITTI validation set (Cars). CF denotes the CF loss.

LI-Fusion	CF	Recall(IoU=0.5)	Recall(IoU=0.7)
×	×	92.23	61.92
✓	×	93.23	65.02
×	✓	93.20	70.69
✓	✓	**93.61**	**71.92**

⋆ equal contribution
⋆⋆ corresponding author

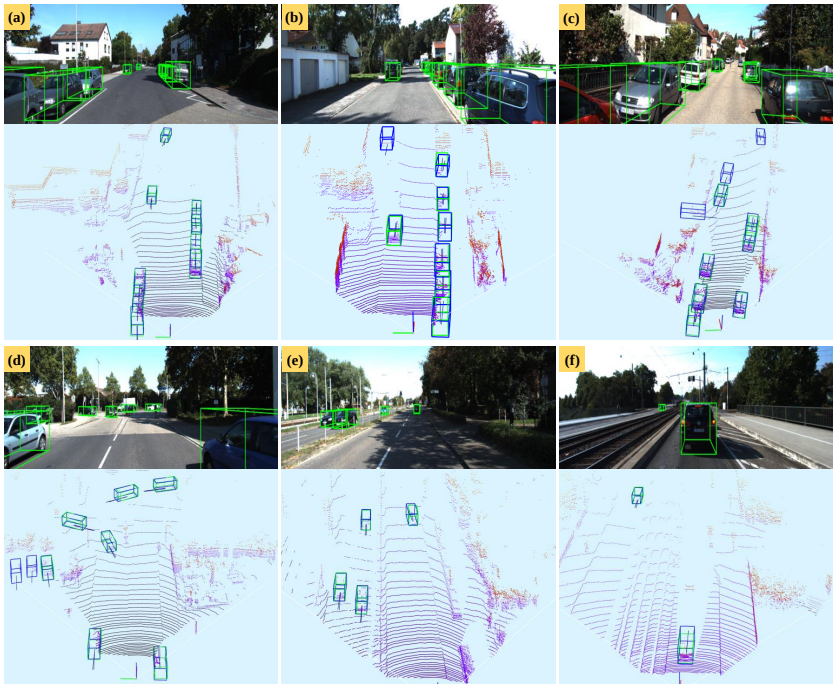

Fig. 1. Qualitative results of our approach on the KITTI validation set. For each pair, the first and the second row show the camera image and the representative view of LiDAR point cloud. The ground truth and detected boxes are highlighted with green and blue boxes, respectively.

2 More Qualitative Results

In this section, we first present more qualitative results on the KITTI and the SUN-RGBD dataset. Then we provide several qualitative analyses of the effect of the LI-Fusion module in the 3D object detection task.

2.1 KITTI Dataset

Fig. 1 illustrates the qualitative results on the KITTI validation set. Our method can detect the objects in the 3D scene accurately. On the one hand, our method produces precise boxes even under the extremely challenging cases where multiple cars are crowded together, as is shown in Fig. 1 (a), (b), and (c). On the other hand, our approach captures the cars far away well (*e.g.* Fig. 1 (d), (e), and (f)), although these objects are usually difficult to be recognized in the camera image and suffer from the sparsity of the point cloud. All these challenging cases persuasively demonstrate the effectiveness of our method.

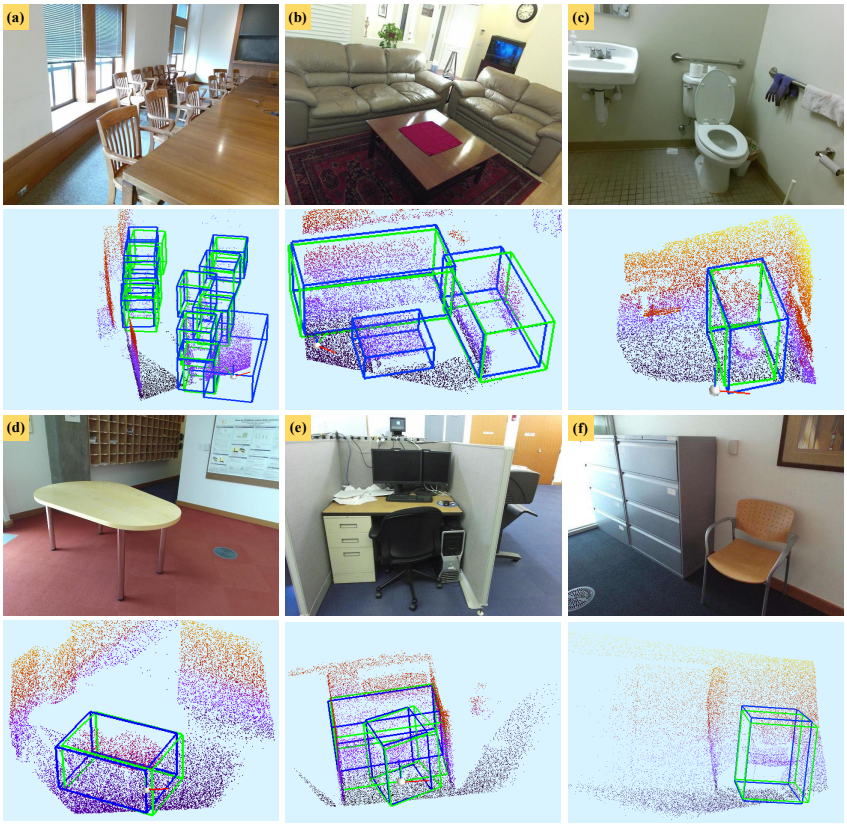

Fig. 2. Qualitative results of our approach on the SUN-RGBD test set. For each pair, the first and the second row show the camera image and the representative view of LiDAR point cloud. The ground truth and detected boxes are highlighted with green and blue boxes, respectively.

2.2 SUN-RGBD Dataset

We present the qualitative results on the SUN-RGBD test set in Fig. 2. Different from the KITTI dataset, SUN-RGBD is an indoor dataset which contains objects of many categories and various scales. As is shown in Fig. 2, our method and can accurately detect multiple kinds of objects with significant scale variations, including large objects (*e.g.*, bed, sofa) and small objects (*e.g.*, dresser, chair). Predicting the bounding boxes of objects in a crowded area is especially challenging. For example, Fig. 2 (a) are crowded with lots of chairs and increase the difficulty for detection significantly. Even under this challenging case, our method still outputs precise bounding boxes, demonstrating the robustness of our method for crowded objects.

2.3 Analysis on the LI-Fusion Module

As mentioned in the main manuscript, the LI-Fusion module can combine the abundant semantic information (*e.g.*, color) in the camera image and the geometric information encoded in the LiDAR point cloud. In this section, we provide a qualitative analysis on the effect of the LI-Fusion module.

We remove the LI-Fusion module from our EPNet and compare its results with those of EPNet. As shown in Fig. 3, the LI-Fusion module leads to more precise bounding boxes. The rationale behind is that the edge information and the color information embedded in the camera image help differentiate an object from its neighboring environment, for example, the desks in Fig. 3(a), as well as the bed and the dresser in Fig. 3(b). These results consistently verify the effectiveness of our LI-Fusion module in exploiting the semantic image information and LiDAR point cloud information for improving the 3D detection task.

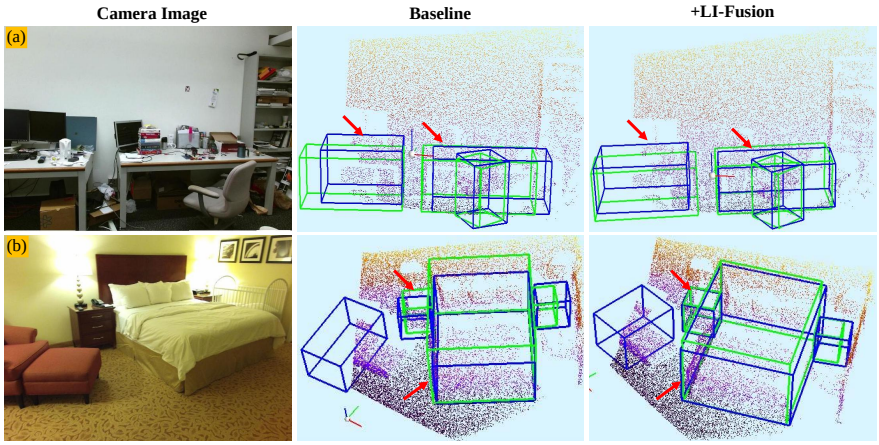

Fig. 3. Qualitative analysis of the effect of our LI-Fusion module on the SUN-RGBD test set. The LI-Fusion module effectively exploits the semantic image information, which is important for generating more bounding boxes .

2.4 Visualization for Images with Varying Illumination

In the main manuscript, we simulate the real environment by varying the illumination condition through a transformation function and verify the effectiveness of our LI-Fusion module. In Fig. 4, we provide some examples generated by the transformation function. It can be seen that the darkened images and the lightened images can well simulate the underexposure and the overexposure cases in the real scenes. Even under such severe illumination conditions, where the camera images bring much interfering information, our LI-Fusion can still effectively

Fig. 4. Visualization of the darkened images and lightened images generated by the illumination transformation to simulate the underexposure and overexposure cases in the real scenes.

enhance the point features and lead to improved detection performance as shown in the main manuscript. It demonstrates the superiority of the LI-Fusion module in adaptively selecting the beneficial features and suppressing the harmful features.