







FSD-BEV: Foreground Self-Distillation for Multi-view 3D Object Detection

Supplementary Material

Zheng Jiang^{1,3,4*} , Jinqing Zhang^{2*} , Yanan Zhang² , Qingjie Liu^{1,2,5†} ,
Zhenghui Hu^{1,2†} , Baohui Wang³, and Yunhong Wang^{1,2} 

¹ Hangzhou Innovation Institute, Beihang University, Hangzhou, China

² State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, Beijing, China

³ School of Software, Beihang University, Beijing, China

⁴ Shanghai ZEEKR Blue New Energy Technology Co., Ltd.

⁵ Zhongguancun Laboratory, Beijing, China

{jzzzz, zhangjinqing, zhangyanan, qingjie.liu}@buaa.edu.cn,
zhenghuihu2013@163.com, {wangbh, yhwang}@buaa.edu.cn

This supplementary material provides more implementation details on FSD-BEV in Sec. A, more experiment results in Sec. B and visualization results in Sec. C.

A More Implementation Details

A.1 Data Augmentation

We first perform random scaling on the input images with a scaling factor in the range of $[0.5, 1.25]$. Then, we crop the images according to the input size, followed by flipping operations with a probability of 0.5. Finally, we rotate the images within the range of $[-5.4^\circ, 5.4^\circ]$ to obtain the augmented input images. Similar to BEVDepth [1], we also perform data augmentation on BEV features. The rotation range is $[-22.5^\circ, 22.5^\circ]$, the scaling factor ranges from $[0.95, 1.05]$, and flipping is applied independently along the X and Y axes with a probability of 0.5.

A.2 Details of Training

During training, we generate ground truth heatmaps by drawing elliptical Gaussian distributions on the original image size and then performing rigid body transformations similar to those applied to the image. We compute the depth loss using Cross Entropy Loss. The center head in Centerpoint [2] is employed as the detection head, using Gaussian Focal Loss to supervise the heatmap of BEV features and L_1 Loss as the regression loss.

We set the detection region along the X and Y axes to $[-51.2, 51.2]$ and along the Z axis to $[-5, 3]$. When our image input size is 256×704 , the BEV features are divided into sizes of 128×128 . However, when we use larger image input sizes, the BEV size is increased to 256×256 .

* Equal contribution. † Corresponding author.

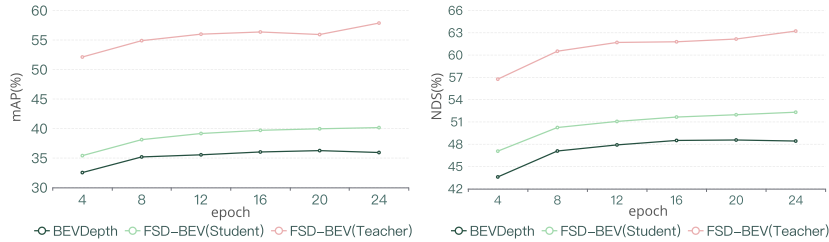


Fig. 1: Comparison of performance between baseline (BEVDepth) and FSD-BEV during training. FSD-BEV is divided into student and teacher branches, and we evaluate mAP and NDS on the nuScenes *val* set.

B More Experiment Results

B.1 Performance Analysis of the Training Process

We conduct a comparative analysis of the performance between the baseline (BEVDepth) and FSD-BEV over the entire training process. Both of them use ResNet50 as the backbone network and are trained for 24 epochs with the CBGS strategy, and their performance is depicted in Fig. 1. During the training process, it is observed that the precision of the FSD-BEV’s teacher branch increases and consistently provides high-quality guidance to the student branch, resulting in significant improvement compared to the baseline. After training for 20 epochs, there is a slight decrease in the precision of BEVDepth, indicating the occurrence of overfitting. On the contrary, FSD-BEV continues to demonstrate a growth trend, which is attributed to the accurate depth information provided by the teacher branch.

B.2 Statistics of Point Cloud Intensification

We quantify the efficacy of Point Cloud Intensification (PCI) by counting the number of intensified ground truth (GT) 3D boxes in the nuScenes *train* dataset. The proportions of the benefited GT boxes are illustrated in Fig. 2. It can be observed that 10.4% of the GT boxes do not carry LiDAR points, which is detrimental to generating high-quality teacher BEV. After applying PPA, 2.6% of the GT boxes are appropriate for supplementing pseudo points, mitigating the loss of objects information to a certain extent. If the Frame Combination (FC) is first applied, the proportion of GT boxes without LiDAR points drops to 6.9%, and the proportion of the GT boxes benefit from PPA drops to 1.5%, demonstrating that FC can perform some of the functions of PPA. However, there are still 5.4% of the GT boxes can not be supplemented by PCI. They could have bad visibility or be located too distant, and boldly intensifying them may introduce inaccurate information.

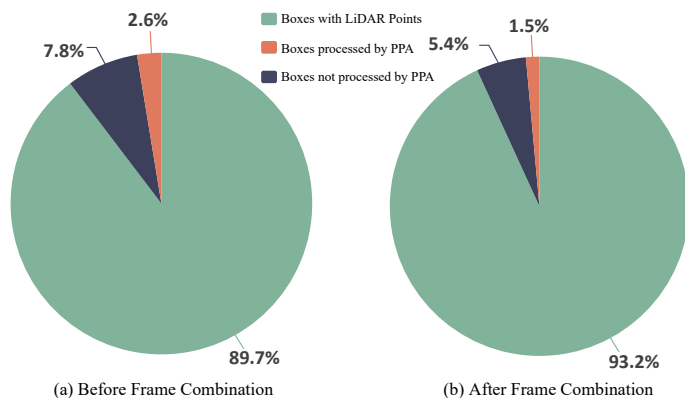


Fig. 2: Statistics of the GT boxes modified by Point Cloud Intensification. (a) illustrates the proportion of benefited GT boxes solely after applying PPA, while (b) shows the proportion of benefited GT boxes under the combined action of FC and PPA.

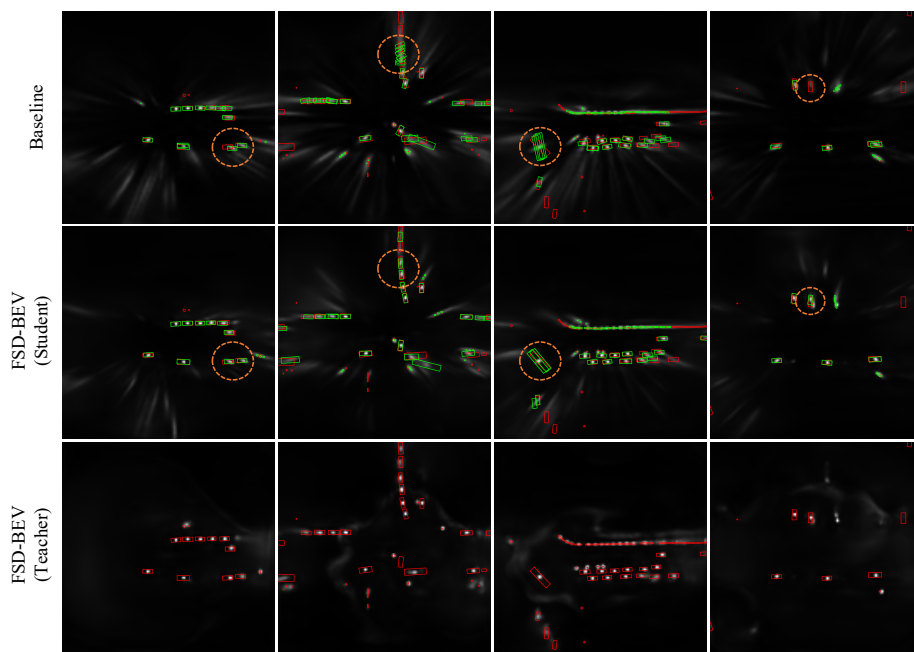


Fig. 3: Visualization results of FSD-BEV and baseline (BEVDepth) on BEV heatmaps. The red and green boxes represent the ground truth and predicted results, while orange circles denote improvement examples of FSD-BEV compared to the baseline.

C Visualization

As shown in Fig. 3, we visualize the predicted BEV heatmaps and bounding boxes of different models. The BEV heatmaps predicted by the teacher branch of FSD-BEV match the GT boxes well, reflecting the effectiveness of hard labels. Compared with the baseline, the heatmaps predicted by the student branch of FSD-BEV are closer to the teacher’s high-quality BEV heatmaps, which leads to more precise predicted boxes. Since the heatmaps are obtained from encoded $\hat{\mathbf{B}}_s$ and $\hat{\mathbf{B}}_t$ mentioned in the main text, it indicates that our distillation scheme works well on forcing $\hat{\mathbf{B}}_s$ to imitate $\hat{\mathbf{B}}_t$.

References

1. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevddepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1477–1485 (2023) [1](#)
2. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021) [1](#)