SA-BEV: Generating Semantic-Aware Bird's-Eye-View Feature for Multi-view 3D Object Detection
Supplementary Material

This supplementary material provides more implementation details on SA-BEV in Sec. A, more experiments results in Sec. B and additional visualizations in Sec. C.

A. More Implementation Details

A.1. Data Augmentation

We augment both images and BEV features following the operation applied in [1]. For images, they are first down-sampled to the desired resolution. Then they are processed by random scaling with a range of $[0.94, 1.11]$, random rotating with a range of $[-5.4°, 5.4°]$ and random flipping with a probability of 0.5. After that, the images are padded and cropped to a uniform shape. For BEV features, augmentation is applied on the virtual points whose features are cumulated to form BEV features. The coordinates of virtual points are processed by random scaling with a range of $[0.95, 1.05]$, random flipping of the X and Y axes with a probability of 0.5 and random rotating with a range of $[-22.5°, 22.5°]$. Augmenting virtual points rather than BEV features themselves can generate more accurate augmented BEV features because the bilinear sampling is not required by the former. The additional BEV data augmentation (BDA) used by BEV-Paste also follows the above settings.

A.2. Detection Configuration

We use the detection head of CenterPoint [10] to detect 3D objects from semantic-aware BEV features and follow the settings used in BEVDepth [3]. The LiDAR coordinate system of nuScenes is used to represent the coordinate of points in the BEV space. The X and Y coordinates are in the range of $[-51.2, 51.2]$, and the Z coordinate is in the range of $[-5, 3]$. The BEV space is divided into pillars for cumulating virtual point features. When the resolution of input images is 256×704, the pillars are in the size of $[0.8, 0.8, 8]$ and the BEV features are in the shape of 128×128. For larger input images, the pillars are in the size of $[0.4, 0.4, 8]$ and the BEV features are in the shape of 256×256.

Table A: Comparison with previous state-of-the-art multi-view 3D detectors on the nuScenes *val* set.

Method	Backbone	Resolution	mAP↑	NDS↑
FCOS3D [8]	ResNet-101	900×1600	0.343	0.415
DETR3D [9]	ResNet-101	900×1600	0.303	0.374
PGD [7]	ResNet-101	900×1600	0.369	0.428
PETR [5]	ResNet-101	512×1408	0.357	0.421
BEVFormer [4]	ResNet-101	900×1600	0.416	0.517
PETRv2 [6]	ResNet-101	900×1600	0.421	0.524
PolarFormer [2]	ResNet-101	900×1600	0.432	0.528
BEVDepth [3]	ResNet-101	512×1408	0.412	0.535
SA-BEV	ResNet-101	512×1408	**0.441**	**0.549**

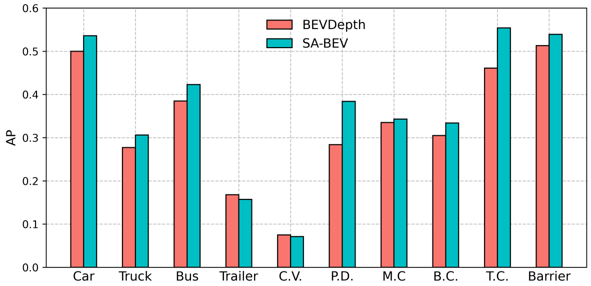

Figure A: Comparison of BEVDepth and SA-BEV on AP for each category. C.V., P.D, M.C., B.C. and T.C. are the abbreviations of construction vehicle, pedestrian, motorcycle, bicycle and traffic cone respectively.

B. More Experiment Results

We change the image backbone of SA-BEV to ResNet-101 when processing 512×1408 resolution images and compare it with other methods that also utilize ResNet-101 as their backbone. The results are shown in Table A. SA-BEV achieves the best mAP and NDS, 2.9% and 1.4% higher than its baseline (i.e. BEVDepth [3]). It also exceeds other start-of-the-art methods that take 900×1600 resolution images as input. This comparison further proves the effectiveness of SA-BEV.

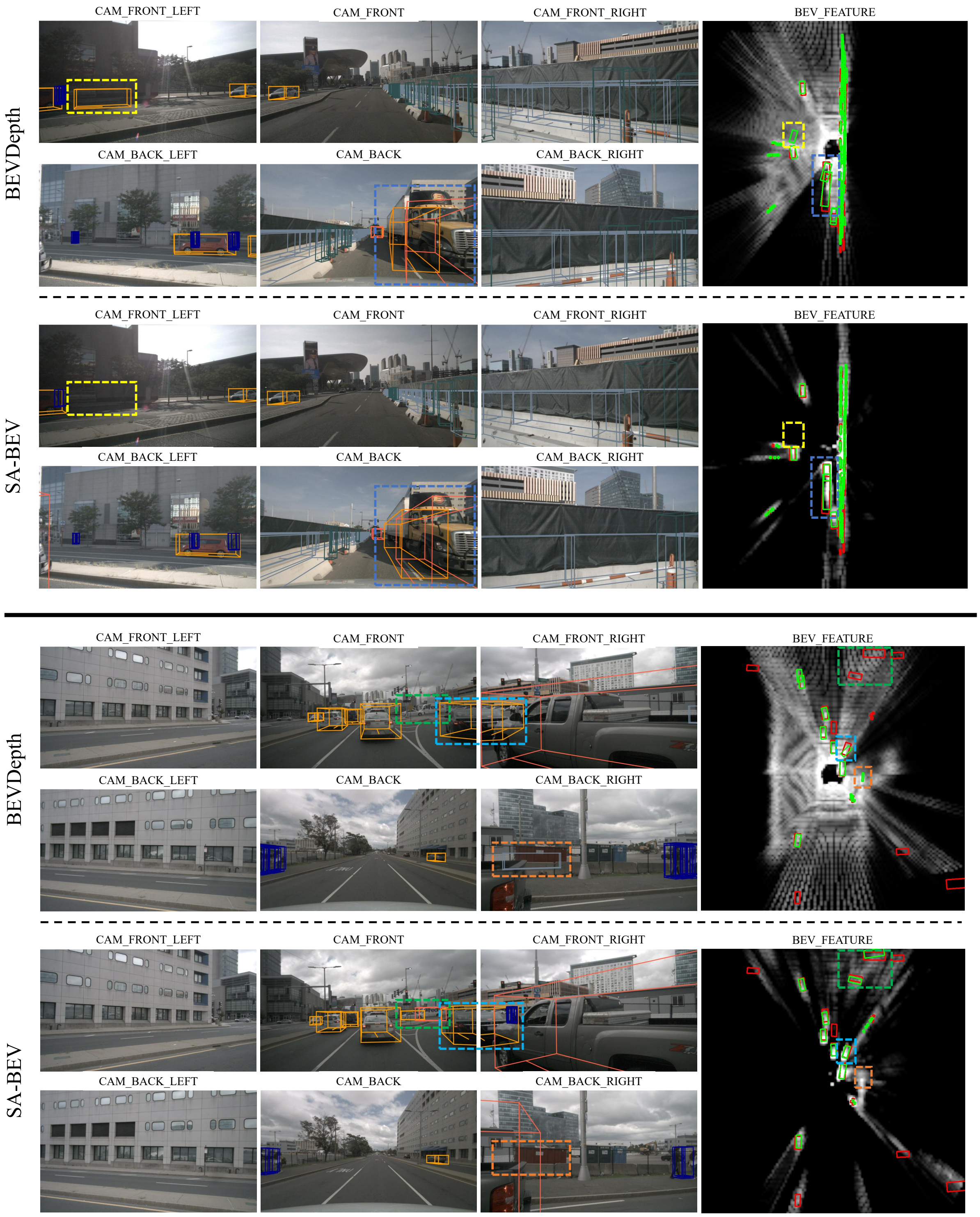

Figure B: Visualization results on images and BEV features. The red boxes and green boxes on BEV features represent the ground truth and the predicted boxes, respectively. The dashed rectangles illustrate that the prediction of SA-BEV is more precise than BEVDepth.

We also compare the detection precision of BEVDepth and SA-BEV in each category and show the results in Fig. A. SA-BEV achieves better precision than BEVDepth in most of the categories. For instance, the APs on pedestrian and traffic cone are increased by about 10%, and the APs on car, truck, bus and bicycle are increased by about 3%. The greater improvement in pedestrian and traffic cone categories indicates that the semantic-aware BEV features effectively preserve the information of small scale objects that is more likely to be submerged by the large proportion of background information.

C. More Visualization Results

We provide more visualization results of BEVDepth and SA-BEV in Fig. B. With the help of semantic-aware BEV features, SA-BEV can recall objects in the far distance and identify the false truth precisely. Besides, SA-BEV generally predicts more accurate locations and directions of the objects, which is also important in actual practice.

References

[1] Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1

[2] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*, 2022. 1

[3] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 1

[4] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022. 1

[5] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 531–548. Springer, 2022. 1

[6] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 1

[7] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 1

[8] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021. 1

[9] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 1

[10] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 1