Method	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	GFLOPs↓	FPS↑
BEVDepth-R50H (E)	54.28	42.64	55.11	26.17	37.57	32.56	19.03	1288.70	2.9
BEVDepth-R50L (A)	42.66	30.79	73.13	28.91	61.74	42.49	21.05	322.17	9.3
+ FD3D(Ours)	45.93 (↑3.27)	32.95 (↑2.16)	71.87	27.28	48.12	38.45	19.73		
BEVDepth-R50L‡(A)	48.68	35.96	63.84	27.27	46.70	34.32	20.90		
+ FD3D(Ours)	**51.08 (↑2.40)**	**38.10 (↑2.14)**	**61.62**	**26.35**	**40.45**	**31.86**	**19.48**		

Table 5. 3D object detection enhancement for BEVDepth on nuScenes *val* set. The symbol "E" and "A" denote "expert" and "apprentice" respectively. The symbol "‡" represents adopting data augmentation and CBGS strategy. We present their efficiency metric with GFLOPS and FPS, which are measured on RTX 2080TI.

A. Experiments on BEVDepth

We also conduct experiments on BEVDepth. The main results and the corresponding settings are depicted in Tab. 5 and Tab. 6 respectively. The following are our experimental setting, implementation details and analysis of the results.

A.1. setting

The setting of the expert-apprentice pair is illustrated in Tab. 6, where both models utilize the identical ResNet-50 backbone that has been pre-trained on ImageNet. The expert network, referred to as BEVDepth-R50H, accepts high-resolution images and generates high-resolution Bird's Eye View (BEV) features, while the apprentice model processes low-resolution inputs. Specifically, the input image resolution and BEV grid size of the apprentice network are half that of the expert counterpart. The achieved gains in efficiency are highlighted by the FPS and GFLOPS values, as demonstrated in Tab. 5, which underscore the substantial improvements resulting from the compression of image and BEV grid resolutions.

A.2. Implementation details

Training of BEVDepth. The implementation is borrowed from BEVDet repository. Two versions of the apprentice network BEVDepth-R50L are developed, distinguished by the presence or absence of data augmentation and the Class Balanced Group Sampling (CBGS) strategy. The impact of knowledge distillation is evaluated separately for each version. The models are trained using the AdamW optimizer for a duration of 24 epochs, with an initial learning rate of 2e-4. The learning rate undergoes a 10-fold reduction following the 16th and 22nd epochs. The training protocol for BEVDepth remains consistent, irrespective of the presence of knowledge distillation. All models are trained using 8 NVIDIA A100 GPUs.

A.3. Results

As shown in Tab. 5, the proposed distillation method improves the performance of the apprentice model BEVDepth-R50L by 3.27 NDS. The observed improvements are mainly attributed to enhancements in mAP, orientation and velocity estimation accuracy. The proposed method yields gains of 2.16, 13.62, 4.04 points in terms of mAP, mAOE and mAVE, respectively. And even for BEVDepth-R50L(‡) which is well optimized via data augmentation and CBGS strategy, the proposed distillation method can still produced an impressive increase of 2.40 in NDS and 2.14 in mAP. The improvements indicate an overall enhancement of the distilled model in localization, classification, and attribute estimation. The attained notable improvement provides compelling evidence that the proposed distillation approach can be extended to BEVDepth and showcases its ability to generalize across diverse 3D object detectors.

Method	BEVDepth-R50H (E)	BEVDepth-R50L (A)
Backbone	ResNet-50	ResNet-50
Image Resolution	512×1408	256×704
BEV Resolution	256×256	128×128

Table 6. The setting of expert-apprentice pairs. The symbol "E" and "A" denote "expert" and "apprentice" respectively. The expert model and apprentice model adopt the identical backbone ResNet-50. "R50H" represents taking high-resolution image with the shape of 512×1408 as input, while "R50L" indicates taking low-resolution image with the shape of 256×704 as input.

B. Experiments on nuScenes *test* set

The effectiveness of the proposed distillation method, FD3D, is assessed on the nuScenes *test* dataset, as indicated in Tab. 7. The observed improvements on the *test* set align with those observed on the nuScenes *val* dataset. The consistency demonstrates the generalizability of the proposed distillation method at the dataset level.

Method	NDS	mAP
BEVFormer-Base* (E)	48.78	38.72
BEVFormer-Tiny (A)	40.54	29.09
+FD3D (Ours)	45.02 (↑4.48)	33.20 (↑4.11)
DETR3D-R101 (E)	41.76	35.78
DETR3D-R50 (A)	35.96	28.91
+FD3D (Ours)	38.56 (↑2.60)	32.10 (↑3.19)

Table 7. 3D object detection improvement on nuScenes *test* set. The results demonstrate that the proposed distillation method benefits BEVFormer and DETR3D on nuScenes *test* set.