DistillBEV: Boosting Multi-Camera 3D Object Detection with Cross-Modal Knowledge Distillation

Supplementary Material

Zeyu Wang[1,2*] Dingwen Li[1*] Chenxu Luo[1] Cihang Xie[2] Xiaodong Yang[1†]
[1]QCraft [2]UC Santa Cruz

Section A describes the architectures of teacher and student models based on CNNs and Transformers. Section B reports the training process and the extra training time induced by the proposed distillation. Section C analyzes the spatial attention maps generated by both teacher and student detectors. Section D provides more implementation details.

A. Architectures

As discussed in the main paper, the LiDAR based teacher model and the multi-camera BEV based student model are separately developed in their specific domains, resulting in different architectures.

We first illustrate the CNNs based architectures as well as the corresponding multi-scale distillation in Figure 4, where H indicates the pre-head layer and B2-B0 denote its three preceding intermediate layers. This figure shows the differences between teacher and student, such as feature map sizes, structures, connections, etc. We introduce the lightweight adaptation module to map the student features before aligning with the associated teacher features. It is also observed that distilling at B0 is detrimental, presumably because the representation gap between the two modalities remains large at the early stage.

To facilitate the cross-modal distillation for BEVFormer, we develop the Transformers based teacher model built on top of CenterPoint or MVP. As illustrated in Figure 5, the distillation is performed at the intermediate features between the encoder layers and the decoder layers.

B. Training Process

Next we take a closer look into the learning process of the student model before and after applying DistillBEV. As shown in Figure 6, the proposed cross-modal distillation approach brings consistent improvements over the baseline (exemplified with BEVDet4D). Comparing the two different teacher models, we observe that MVP (camera-LiDAR

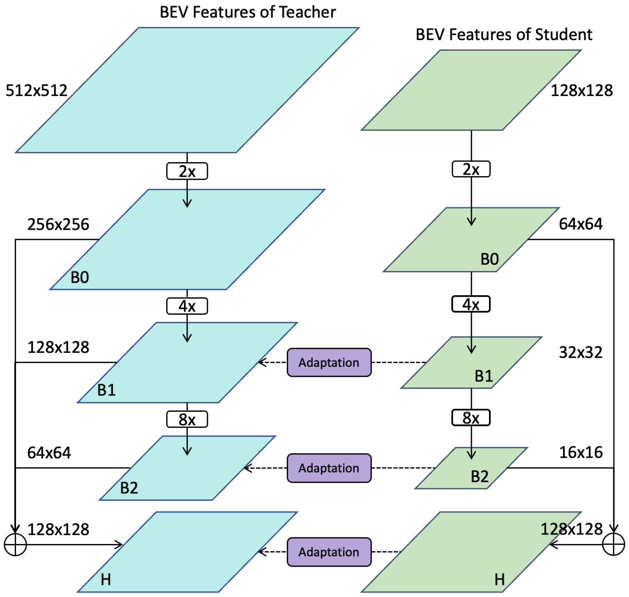

Figure 4. Illustration of the architecture details (BEV feature encoding parts) in teacher and student networks based on CNNs, as well as the multi-scale distillation performed at different levels.

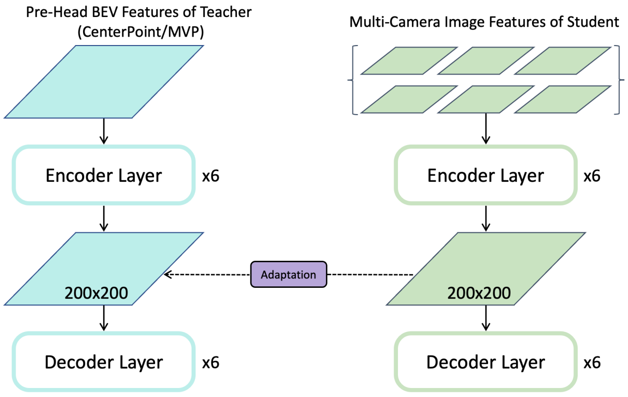
Pre-Head BEV Features of Teacher (CenterPoint/MVP)
Multi-Camera Image Features of Student
Encoder Layer
x6
Encoder Layer
x6
200x200
Adaptation
200x200
Decoder Layer
x6
Decoder Layer
x6

Figure 5. Illustration of the architecture details (encoder and decoder parts) in teacher and student networks based on Transformers, as well as the distillation performed at the corresponding level.

[*] Equal contribution

[†] Correspondence to xiaodong@qcraft.ai

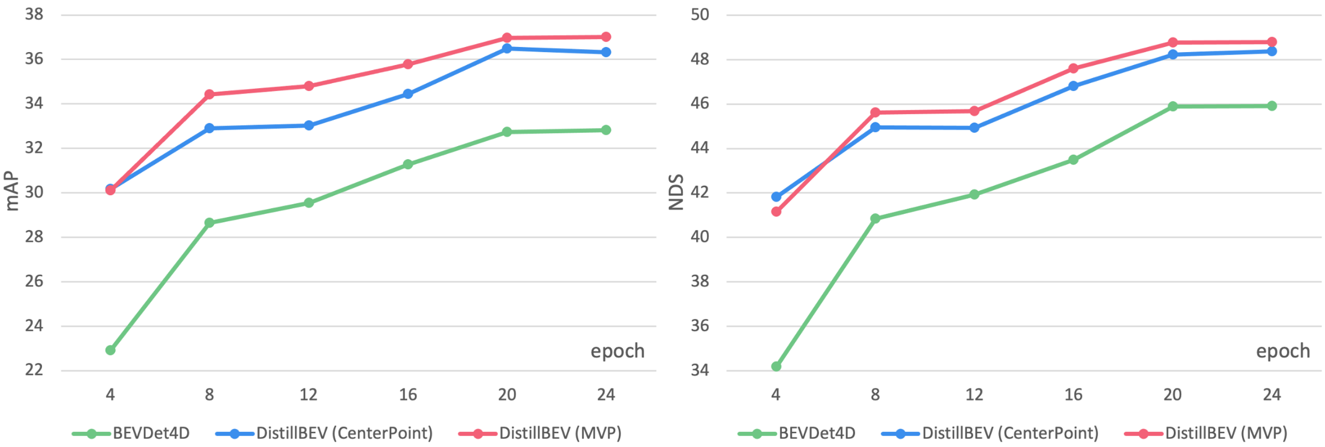

Figure 6. Comparison of the training process between the student model (BEVDet4D) and the distilled versions using CenterPoint and MVP as the teacher models. We report the results of mAP and NDS on the validation set of nuScenes.

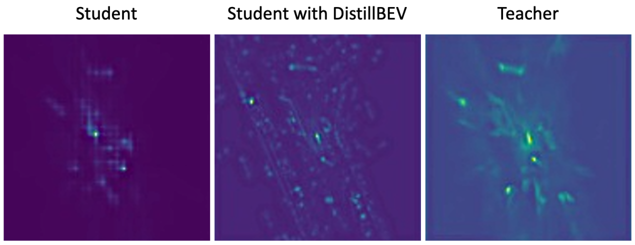

Figure 7. Visualization of the spatial attention maps generated by the teacher detector (MVP) and student detector (BEVDepth) before and after using DistillBEV.

fusion) is more effective than CenterPoint (LiDAR only) to perform distillation in general, and the performance gains are not diminishing along with the training.

As for the extra training time induced by our approach, training with 8 V100 GPUs, the student model takes 42.3 hours, and DistillBEV conducted on this model uses 49.0 hours (+15.8%), which is a relatively low extra training cost compared to the large performance gain.

C. Attention Visualization

To investigate the cross-modal distillation effect to the change of student features, we visualize the spatial attention maps generated by the teacher and student models (with and without DistillBEV) following Equations (4-5) in the main paper. As shown in Figure 7, the spatial attention map generated by the student exhibits a drastically different pattern compared to the one by the teacher model. The former concentrates on the central area (i.e., close to the ego-vehicle), and rarely activates in some distant yet important areas. After training with DistillBEV, the attention map produced by student becomes sharper and is more similar to the one of teacher in both nearby and faraway regions.

D. More Implementation Details

In all experiments, we set the region decomposition and spatial attention related hyper-parameters as $\eta = 20$, $\tau = 0.5$, $\gamma = 0.1$. As for the loss related hyper-parameters, α = 6e-3, β = 4e-2, λ = 2.5e-3 for the networks based on CNNs, α = 3e-3, β = 4e-2, λ = 1e-3 for the networks based on Transformers (CenterPoint used as the teacher), α = 5e-3, β = 4e-3, λ = 5e-4 for the networks based on Transformers (MVP used as the teacher).