QD-BEV : Quantization-aware View-guided Distillation for Multi-view 3D Object Detection (Supplementary Materials)

Yifan Zhang[1*], Zhen Dong[2*], Huanrui Yang[2], Ming Lu[3], Cheng-Ching Tseng[3], Yuan Du[1], Kurt Keutzer[2], Li Du[1†], Shanghang Zhang[3†]

[1]Nanjing University, [2]University of California, Berkeley,

[3]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University,

The supplementary materials contain additional implementation details, extra experimental results, ablation study, and visualization results.

A. Additional Implementation Details

A.1. Supplementary description on datasets

As we mentioned in the main body of the paper, the nuScenes [1] dataset has 750 scenarios as the training set, 100 scenarios as the validation set, and 150 scenarios as the test set. All our experiments are conducted on the nuScenes train set and tested on the nuScenes val set. Better results can be obtained using data enhancement and additional training, and some previous works [5, 2, 3] use additional training data in order to get better results on the test set. However, for the sake of fairness, we only train on the original training set and do not use techniques such as additional data and data enhancement.

A.2. Extra details on training strategy

Our experiments mainly use Tesla V100 32G GPU and Tesla A40 48G GPU to meet our video memory and computing power requirements. For the experiments of QD-BEV-Tiny, we use 8 pieces of Tesla A40 48G GPU with parallel computing, where the batch size is 6. For QD-BEV-Small and QD-BEV-Base experiments, we use 8 Tesla V100 GPU and 8 Tesla A40 GPU with parallel computing, where the batch size is 1. For the tiny, small, and base models, when the batch size is 1, the required single-card memory is 7G, 30G, and 47G, respectively.

For the training parameters, we generally follow the training configuration of the previous work [2, 5]. In progressive quantization-aware training, we use the initial learning rate of 2e-4, learning rate multiplier of the backbone is 0.1 in each stage. In view-guided distillation, we use an initial learning rate of 1e-5, and the learning rate multiplier of the backbone is 0.5.

For the temperature parameter τ of the view-guided distillation, our default configuration is $\tau = 1$. To evaluate the sensitivity of the final performance to the hyperparameter τ, we have conducted ablation study on our QD-BEV models with different τ in Section B.

Table 1: Ablation study on the temperature parameter τ in VGD.

Model	τ	NDS	mAP
QD-BEV-Tiny	1	0.372	0.255
	2	0.371	0.258
	4	0.374	0.258
QD-BEV-Small	1	0.479	0.374
	4	0.481	0.371
QD-BEV-Base	1	0.506	0.403
	4	0.509	0.406

B. Additional Ablation Study

In previous work [4], it is found that the temperature parameter has an obvious effect on the results of distillation. Therefore, we carried out a control experiment with different hyperparameter τ. We change the probability distribution of Softmax on the image feature and the BEV feature by selecting different τ.

* Equal contribution :
zhang_yifan@smail.nju.edu.cn, zhendong@berkeley.edu
† Corresponding authors :
ldu@nju.edu.cn, shanghang@pku.edu.cn

In Table 1, we can see different hyperparameters do not have a decisive impact on the results. There is indeed some improvement in the performance of the three models when $\tau = 4$, but the gap between a good result and a bad result is within 0.003 NDS, and the accuracies of all experiments are significantly and consistently higher than that of QAT.

C. Additional Visualization

C.1. Feature map visualization

We visualize the feature map of the QD-BEV-Base model with view-guided distillation in Figure 1. Here the teacher is BEVFormer-Base, and the student is our QD-BEV-Base model with 4-bit weights and 6-bit activations. Our view-guided distillation method considers the Image feature and the BEV feature at the same time to obtain a better convergence direction of the model.

C.2. Visualization on the oscillation during QAT

As discussed in the main contexts, we found that the quantization-aware training (QAT) has severe stability issues, with accuracy curves of both mAP and NDS oscillating up and down throughout the process. We visualize this phenomenon in Figure 2. As shown in Figure 2c, when conducting QAT for W4A6, the standard QAT sometimes suffers from gradient explosion, causing the training to collapse after a few epochs. In contrast, the progressive QAT has a better curve, but the stability issue still exists, and it has the drawback that it is difficult to achieve higher accuracy, as shown in Figure 2b. On the other hand, the VGD curve maintains stability while continuously improving accuracy, eventually achieving excellent results, as shown in Fig. 2a.

To explain this phenomenon, we visualize the changes in weight distribution during training. Figure 3 shows the weight distribution of the image neck during training. As can be seen, for standard QAT in Figure 3a, it collapses since the weights converge to zero during the training. Compared with progressive QAT in Figure 3b, VGD in Figure 3c gradually learns weight distributions that can extract more features, while the weight distribution of progressive QAT does not change significantly.

Figure 4b shows the weight distribution of the classification branch during training. We can see that VGD in Figure 4c is more stable and does not experience irregular oscillations in weight distribution like progressive QAT in Figure 4b.

C.3. Additional 3D object detection visualization

Here in Figure 5, we visualize our model on more samples. In Figure 5a, the upper part is the viewing angle of the three cameras in front of the car, and the lower half is the viewing angle of the three cameras behind the car. From top to bottom are the Ground Truth, BEVFormer-Tiny results, and our QD-BEV-Base results, respectively. We can find that QD-BEV-Base predicts the results more accurately than BEVFormer-T, and the error rate is also significantly lower.

In Figure 5b, the BEV visualization of the QD-BEV-Base is on the left and the BEVFormer-Tiny result is shown on the right as a comparison. In these two pictures, the color blue represents the predicted results and the color green represents the Ground Truth. As can be seen in the figures, QD-BEV-Base (32.9MB) made a much clearer and more accurate prediction than BEVFormer-Tiny (126.8MB) with only 1/4 of the model size.

D. Video Demo

To showcase the efficacy of QD-BEV models, we made a 1080P HD video demo of QD-BEV-Base (32.9 MB) for about 20 seconds, including its comparison with BEVFormer-T (126.8 MB) and Ground Truth. The video is attached to the zip file.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[2] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022.

[3] 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.

[4] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021.

[5] 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.

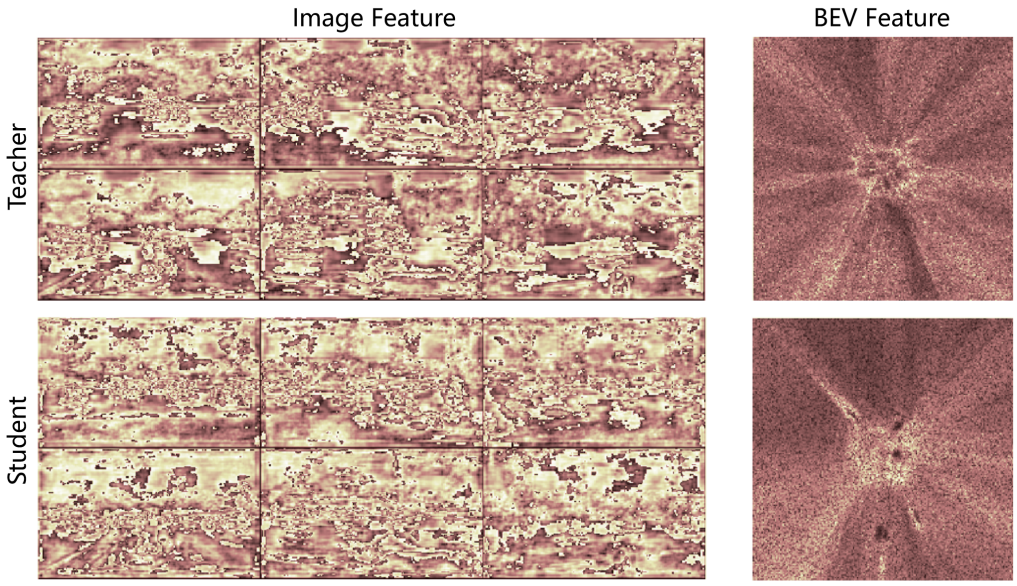
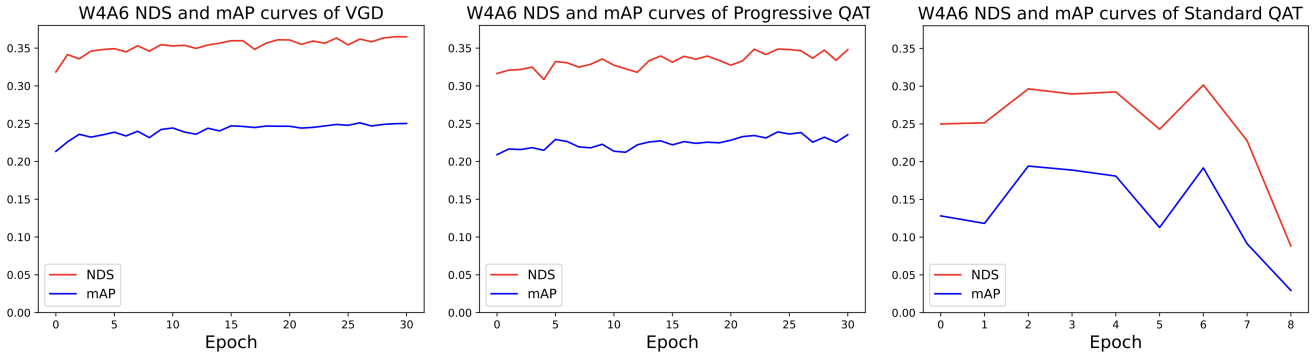
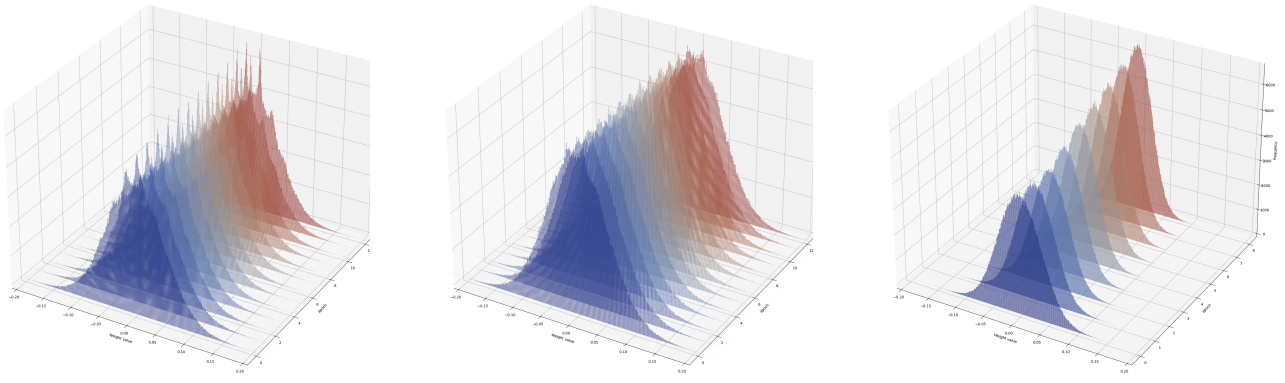


Figure 1: Visualizaton of QD-BEV-Base feature map. On the left is the Image features of the teacher and student model, while on the right we show the BEV features.



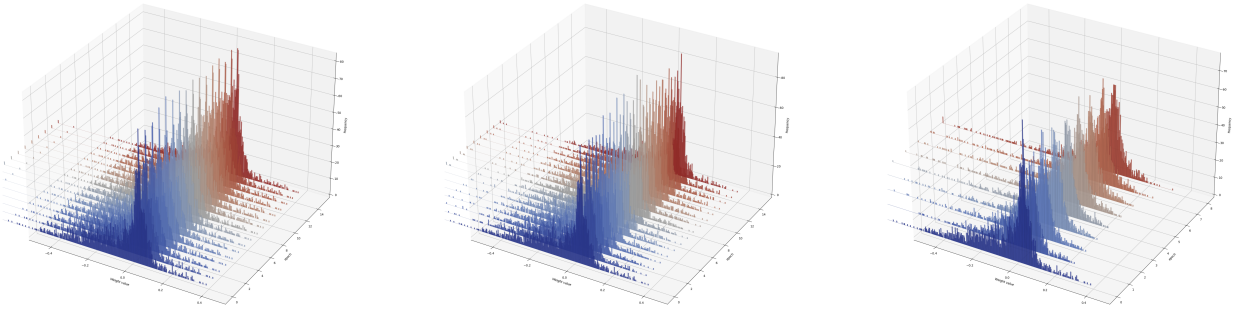
(a) Oscillation of VGD (b) Oscillation of Progressive QAT (c) Oscillation of Standard QAT

Figure 2: Visualization on the oscillation of accuracy during VGD, compared to Progressive QAT and Standard QAT.



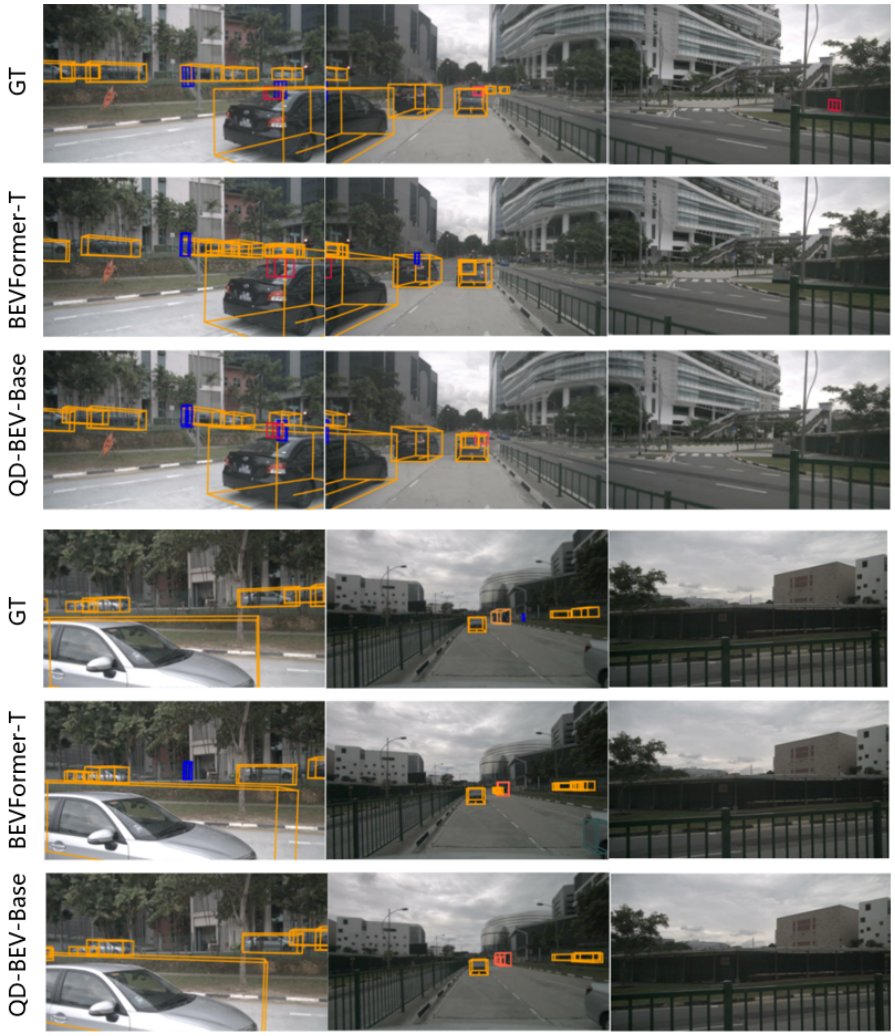
(a) weight distribution of VGD (b) weight distribution of Progressive QAT (c) weight distribution of Standard QAT

Figure 3: Visualization of the weight distribution shift during VGD, progressive QAT, and standard QAT, measured on the image neck. The x-axis stands for the weight value, the y-axis shows the index of the current epoch during training (either VGD, progressive QAT, or standard QAT), and the z-axis is the frequency of the corresponding weight value.

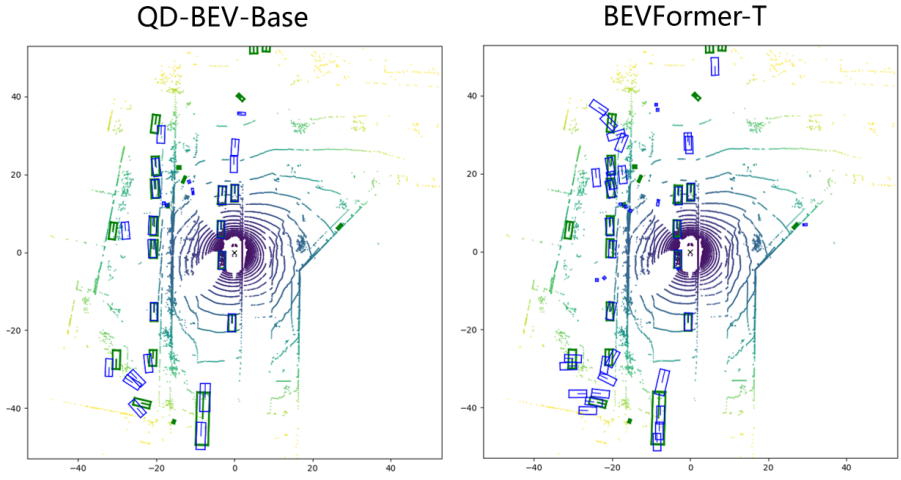


(a) weight distribution of VGD (b) weight distribution of Progressive QAT (c) weight distribution of Standard QAT

Figure 4: Visualization of the weight distribution shift during VGD, progressive QAT, and standard QAT, measured on the classification branch. The x-axis stands for the weight value, the y-axis shows the index of the current epoch during training (either VGD, progressive QAT, or standard QAT), and the z-axis is the frequency of the corresponding weight value.



(a) Visualization of 3D detection results of QD-BEV-Base, BEVFormer-T and Ground Truth



(b) BEV visualization of QD-BEV-Base and BEVFormer-T

Figure 5: Visualization of QD-BEV-Base results and the comparison with results obtained by BEVFormer-T and Ground Truth. The upper figures are from front cameras, and the lower figures are from back cameras.