Learning High-resolution Vector Representation from Multi-Camera Images for 3D Object Detection

Supplementary Material

Zhili Chen[1†], Shuangjie Xu[1], Maosheng Ye[1], Zian Qian[1], Xiaoyi Zou[2], Dit-Yan Yeung[1], and Qifeng Chen[1✉]

[1]HKUST [2]DeepRoute.AI

{zchenei, shuangjie.xu, myeag, zqianaa}@connect.ust.hk,
xiaoyizou@deeproute.ai, {dyyeung, cqf}@cse.ust.hk

1 Visualization of Foreground Proposals and Deformable Offsetting

We demonstrate the chosen crucial positions in which we construct the sparse HR BEV queries in Fig. 1. As shown in the middle of Fig. 1, the red spots indicate the positions that are likely to contain objects. The initial proposals are close to each other. After applying the deformable offsetting, the proposals dispersed to the regions of interest and covered broader regions, as demonstrated in the right of Fig. 1.

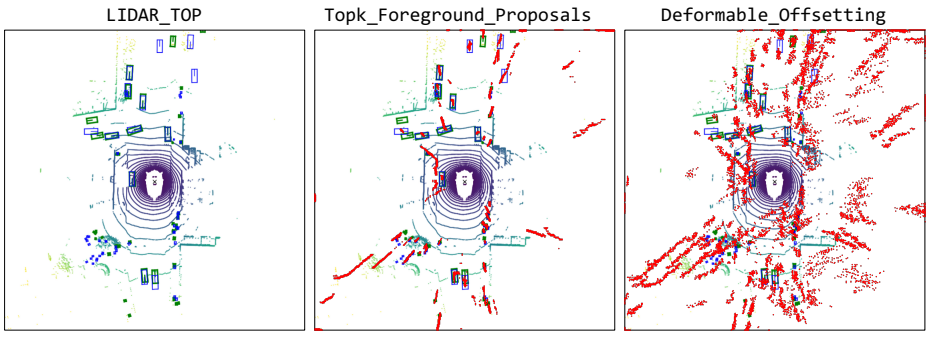

Fig. 1: Visualization of recognizing foreground regions by taking the directional top-k on the predicted heatmap and then applying deformable offsetting based on these proposals. The bird's-eye-view of LiDAR with detection predictions and the corresponding ground truths is shown on the left. As presented in the middle, we obtain the foreground proposals (red spots) by taking top-k along the $x-$axis and $y-$axis. We construct the sparse HR BEV query for the proposal positions (red spots) after applying deformable offsetting, as demonstrated on the right.

2 Loss Details

We followed similar loss designs of BEVFormer [3] for bounding boxes' categories classification and attributes regression. The same losses will be computed for the decoded Vector queries from the intermediate encoding layers. We denote the overall classification loss as $\mathcal{L}_{cls}$ and the regression loss as $\mathcal{L}_{reg}$, respectively.

As discussed in Sec. 3.2, we apply a Gaussian focal loss to supervise the heatmap predictions, which are denoted as $\mathcal{L}_{hm} = \mathcal{L}_{focal}(\mathcal{H}) + \mathcal{L}_{focal}(\mathbf{h}')$.

The overall training loss is:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{cls} + \lambda_2 * \mathcal{L}_{reg} + \lambda_3 * \mathcal{L}_{hm}, \tag{1}$$

where $\lambda_1 = 2.0$, $\lambda_2 = 0.25$, and $\lambda_3 = 0.5$.

3 Results Details

We present the 3D detection results of each object category in Tab. 1. Compared to BEVFormer-S [3] under the small setting, our VectorFormer-S shows significant improvements on almost all of the object classes. Regarding the base setting with a larger model, we demonstrate a leading performance in general.

Method	Car	Truck	Bus	Trailer	C. V.	Ped.	Motor.	Bicycle	T. C.	Barrier
BEVFormer-S [3]	56.0	28.3	43.1	**16.5**	10.5	44.0	36.2	33.8	54.1	47.8
VectorFormer-S	**59.8**	**33.4**	**45.6**	16.1	**13.5**	**47.0**	**41.5**	**36.6**	**58.0**	**54.0**
BEVFormer-B [3]	**61.8**	**37.0**	**44.5**	17.1	**12.9**	49.4	43.1	39.8	58.4	52.5
VectorFormer-B	**61.8**	36.2	43.3	**19.2**	10.2	**50.6**	**44.4**	**43.2**	**59.8**	**56.4**

Table 1: 3D detection results of each object category on nuScenes [1] validation set. C.V., Ped., Motor., and T. C. represent the construction vehicle, pedestrian, motorcycle, and traffic cone, respectively. The best results compared to the baselines are in bold.

4 Additional Ablation Studies

The settings of the following experiments are consistent with Sec. 4.3, which used VectorFormer-S by default and training with 12 epochs.

Effect of Heatmap Loss Weights We introduced the heatmap loss to the overall training loss and conducted an ablation study on the weights of the heatmap loss in Tab. 2. Among different settings, the performance reach the best when we set $\lambda_3 = 0.5$ in Eqn. 1.

Heatmap Loss Weight	NDS	mAP	mATE	mASE	mAOE
$\lambda_3 = 0.25$	49.54	39.17	70.86	27.58	39.68
$\lambda_3 = 0.5$	**49.73**	39.76	70.92	**27.37**	**38.77**
$\lambda_3 = 1.0$	49.71	**40.19**	**68.87**	27.58	43.34

Table 2: The ablation study on the weights of heatmap loss in Eqn. 1.

Effect of Selecting Decoding Queries from BEV Guided by the Heatmap This ablation study is based on the settings of adding the heatmap supervision to the baseline (the second row of Tab. 6 in the main paper and Tab. 3 in supplementary). In contrast to this setting that used the initialized embeddings as decoding queries, we further experimented with using the Top-k BEV queries as the decoding queries, which were obtained according to the Top-k objectiveness score indicated by the heatmap prediction. It is observed that the experiment of decoding from the BEV queries achieves the worst results at the third row of Tab. 3. We conjecture that selecting object queries by the heatmap might not be stable and cause the decoding queries to barely receive supervision signals while training, which leads to a worse result.

	NDS	mAP	mATE	mASE	mAOE
Baseline	48.4	38.1	72.8	27.8	39.9
+ Heatmap	48.9	38.8	72.1	27.8	41.3
+ Decoding from the BEV	47.2	37.0	72.3	27.7	41.7
Ours	**49.7**	**39.8**	**70.9**	**27.4**	**38.8**

Table 3: This is the ablation study using the Top-k BEV queries as the decoding queries. The second row adds heatmap supervision to the baseline while using the randomly initialized embeddings as the decoding queries. The third row uses the Top-k BEV queries selected by the heatmap guidance.

Effect of Intermediate Supervision As discussed in Sec. 3.5, the Vector queries with the representative scene context priors will be sent to the decoder as the decoding queries. Inspired by the DETR training strategy proposed in [2], we also decode the Vector queries from the intermediate encoding layers for additional supervision. As shown in Tab. 4, this training strategy can boost training convergence, resulting in better performance.

Intermed. Sup.	NDS	mAP	mATE	mASE	mAOE
-	49.0	39.4	**70.1**	27.8	42.1
✓	**49.7**	**39.8**	70.9	**27.4**	**38.8**

Table 4: The ablation study on decoding the intermediate Vector query while training.

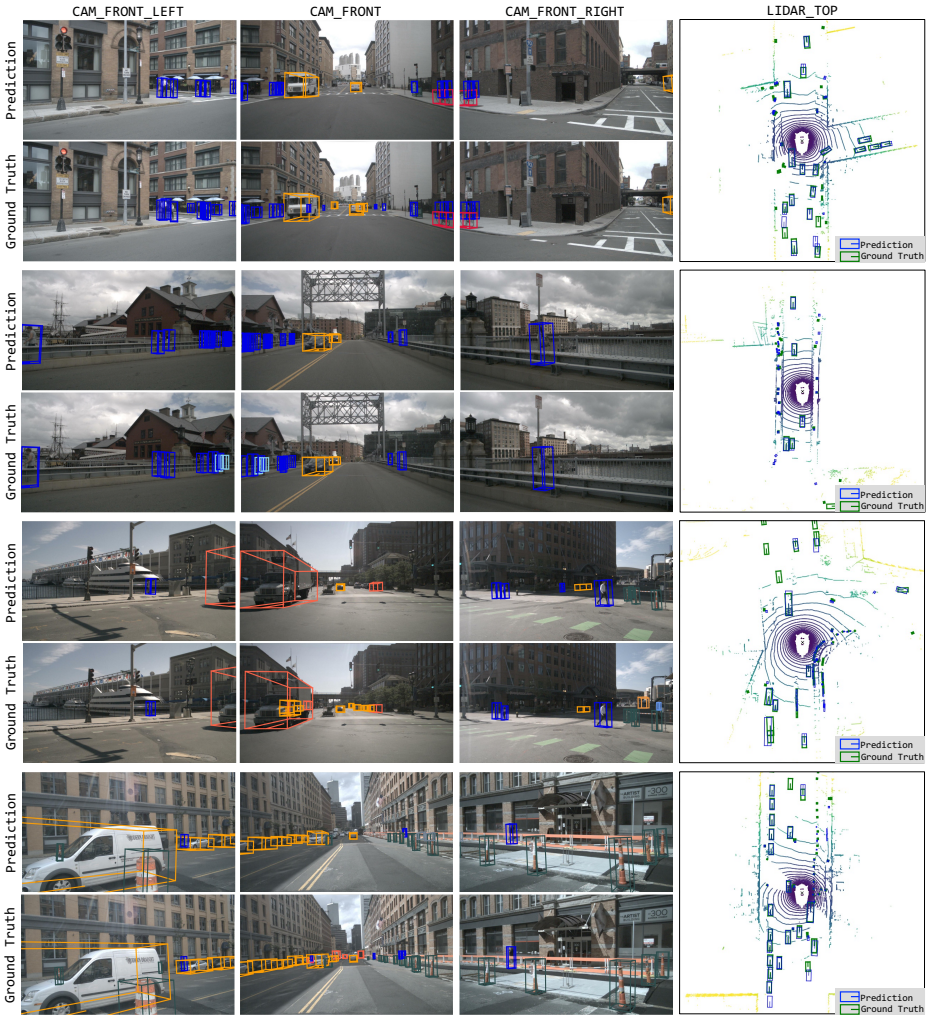

Fig. 2: Additional visualization results of VectorFormer on nuScenes [1] validation set. Detection predictions with ground truth in multi-view camera images are shown on the left and in bird's-eye-view is shown on the right.

5 Additional Qualitative Results

We present additional visualization results in Fig. 2. The VectorFormer achieves overall outstanding detection performance and accurately recognizes objects of small-scale remarkably. We observed some cases in which the VectorFormer failed to detect distant objects or the objects being severely occluded.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H.: Detrs with hybrid matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19702–19712 (2023)
3. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022)