BEV-SAN: Accurate BEV 3D Object Detection via Slice Attention Networks

Xiaowei Chi[1,2*] Jiaming Liu[1*] Ming Lu[1*] Rongyu Zhang[1]
Zhaoqing Wang[3] Yandong Guo[4] Shanghang Zhang[1†]
[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]The Chinese University of Hong Kong [3]The University of Sydney [4]AI2Robotics

In the supplementary material, we first present additional related work of transformer network in Sec .1 since we utilize dual-branch transformer module to fuse the global and local slices. In Sec .2, we then provide additional and detailed cross domain training strategy. In Sec .3, we explore the generalization ability of our proposed BEV-SAN by evaluating the performance on unseen and challenging data distribution. In Sec .4, we demonstrate the robustness of our method by comparing with baseline methods when encountering cameras malfunctioning.

1. Additional related works

Vision transformer. Transformer network was first introduced for neural machine translation tasks [15], and the encoder and decoder of transformer leverage self-attention mechanism to extract better feature representation and reserve contextual information [7,12,15]. Vision Transformer (ViT) [2,14] first brings a transferring in backbone architectures for computer vision, which is transferred from CNNs to Transformers. This seminal work has led to subsequent research that aims to improve its utility [10]. Meanwhile, Swin Transformer [9] is a practical backbone for various image recognition tasks, which adopts the inductive biases of locality, hierarchy and translation invariance. DeiT [14] focuses on improving the efficiency and practicality of transformer network, it proposes several training strategies that allows ViT to be effective when training on smaller image datasets. In this paper, we introduce a dual branches transformer block to fuse global an local-level BEV slices and generate the fused BEV feature map for task heads.

2. Additional implementation details

Our training process can be regarded as an end-to-end training. Firstly, in order to fully leverage the feature extraction ability of the model [5], we load the backbone of ImageNet pretrained parameters. Then we train the model with slice-attention module for 28 epochs with **CBGS** [16] and 40 epochs without. It should be noted that we freeze the backbone starting from epoch 23 and fine-tune the slice-attention module and detection head in the rest of the epochs. We adopt 256×704 as image input size and the same data augmentation methods as [5]. We apply AdamW [11] optimizer with 2e-4 learning rate. We decay the learning rate on epochs 19, 23, and 33 with ratio $\alpha = 1e-7$. As for further detailed image augmentation process, we follow BEVDepth and adopt random cropping, random scaling, random flipping, and random rotation. The BEV feature generated by the model is also augmented by random scaling, random flipping, and random rotation. All experiments are conducted on NVIDIA Tesla V100 GPUs.

3. Additional generalization exploration

Slice-attention module leverages the attention mechanism of Transformer to fuse the features from different global information to construct a more comprehensive BEV feature. Therefore, BEV-SAN is of better generalization ability in more display scenarios after integrating multiple levels of information. We conduct further experiments on some particular scenarios like rainy and night in NuSences dataset to demonstrate the superiority generalization ability of BEV-SAN.

As shown in Tab. 1, the baseline can only achieve 0.170 and 0.124 in NDS and mAP, respectively on the night validation set. Due to the faint light condition at night, the camera based method will encounter great challenges. However, we observe that BEV-SAN shows satisfying performance under such severe condition with 0.210 NDS and 0.129 mAP, respectively. As for rainy validation set, we notice that BEV-SAN also outperforms the baseline with significant margin by over 3% in NDS. These results verify the generalization ability of BEV-SAN.

4. Additional robustness exploration

Though there are lots of recent works on autonomous driving systems, only a few of them [6, 13] explore the robustness of the proposed methods. LSS [13] presents the

*Equal contribution: liujiaming.pku@gmail.com
†Corresponding author: shzhang.pku@gmail.com

Table 1. Comparisons of Generalization ability with different methods on the validation set of unseen environment [1]. The unseen environment includes night-time and rainy data. All methods utilize ResNet 50 [3] as backbone.

Test on	Method	Backbone	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
Night	BEVDepth [5]	R50	0.170	0.124	0.847	0.463	0.906	1.855	0.696
	BEV-SAN	R50	**0.210**	**0.129**	**0.827**	**0.466**	**0.670**	**1.655**	**0.584**
Rainy	BEVDepth [5]	R50	0.363	0.305	0.722	0.298	0.662	0.915	0.289
	BEV-SAN	R50	**0.396**	**0.314**	**0.711**	**0.296**	**0.629**	**0.664**	**0.242**

Table 2. Comparisons of the Robustness ability with different methods on the validation set [1]. We design a special experiment setting in which one camera breaks down or is occluded. And we occlude the front-view images in inference time.

Occlude	Method	Backbone	NDS ↑	mAP ↑
	BEVDepth [5]	R50	0.336	0.296
Front	BEVDepth [5]	R50	0.318	0.228
	Ours(BEVDepth)	R50	**0.325**	**0.258**
Front-Left	BEVDepth [5]	R50	0.331	0.265
	Ours(BEVDepth)	R50	**0.332**	**0.279**
Front-Right	BEVDepth [5]	R50	0.326	0.242
	Ours(BEVDepth)	R50	**0.330**	**0.271**

performance under extrinsic noises and camera dropout at test time. Following previous work, we aim to give a qualitative analysis of our method under camera missing condition. Camera image missing occurs when one camera breaks down or is occluded. Multi-view images provide panoramic visual information, yet it can also face the condition when one of them is absent in the real-world. Therefore, it is necessary to evaluate the robustness of our method when encountering camera view missing.

As shown in Tab. 2, among six cameras of nuScenes dataset, front-view data are the most important, and their absence leads to a drop of 1.8% NDS and 6.8% mAP on BEVDepth [5]. In term of our proposed method, front-view camera missing only leads to a drop of 1.1% NDS and 3.8% mAP, which demonstrates that BEV-SAN has great potential on robustness. For other views missing, the results show a similar tendency.

5. additional comparative result

5.1. Ablation study of LiDAR-guided slicing

In order to construct the local slices, we need to slice the overall height range [-6,4] into several bins. Instead of uniform slicing, we propose to use the statistics of LiDAR points along the height dimension (Fig. 3) to guide the slicing. Specifically, we accumulate the histogram and choose the local slices from the accumulated distribution. We call this strategy as LiDAR-guided slicing (LiDAR-guided sampling in the draft). LiDAR-guided slicing can make the local slices focus on the informative foreground due to the localization advantage of LiDAR points. To better evaluate the effectiveness of LiDAR-guided slicing, we compare it against uniform slicing under different numbers of bins in Tab. 3. As can be seen, LiDAR-guided slicing consistently outperforms uniform slicing, demonstrating its effectiveness.

Table 3. Ablation study of LiDAR-guided slicing.

Statistics Local	NDS	mAP
3 Local Bins (Uniform)	0.352	0.298
4 Local Bins (Uniform)	0.349	0.299
5 Local Bins (Uniform)	0.352	0.307
6 Local Bins (Uniform)	0.359	0.310
7 Local Bins (Uniform)	0.358	0.300
8 Local Bins (Uniform)	0.359	0.305
6 Local Bins (LiDAR)	**0.366**	**0.310**

5.2. Detection results of each object

Tab.4 shows the results of local slices and global slices on each object category. As can be seen, local slices can benefit all the categories (except Bus). This is because Bus is taller than other categories.

5.3. Comparison with more established methods

Tab.5 shows the comparison with more established methods [8] [5] [6] under the same image backbone for a more comprehensive evaluation. As can be seen, our method is still competitive with well-established methods.

Table 4. 3D object detection results (mAP) of each object category on nuScenes val set.

Method	Truck	trailer	Car	Bus	Pedestrian	Motorcycle	Bicycle	Barrier	Traffic cone
BEVDepth [5]	0.237	0.153	0.466	0.332	0.247	0.289	0.267	0.417	0.465
SANet(Local)	0.240	0.176	0.476	0.345	0.257	0.296	0.283	0.498	0.432
SANet(Global)	0.250	0.156	0.471	0.333	0.248	0.300	0.274	0.479	0.409
SANet	0.244	0.165	0.491	0.350	0.265	0.302	0.272	0.432	0.503

Table 5. Comparison with more established methods under the same image backbone without CBGS.

Method	Backbone	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVFormer [6]	R50	0.359	-	-	-	-	-	-
PETRv2 [8]	R50	0.350	-	-	-	-	-	-
BEVDepth [5]	R50	0.336	0.296	0.732	0.283	0.713	1.218	0.396
BEV-SAN	R50	**0.366**	**0.310**	**0.705**	**0.278**	**0.608**	**1.070**	**0.300**

5.4. Computation Cost

Recently, BEVPoolv2 [4] presents an engineering optimization for LSS operation. We re-implement our method based on BEVPoolv2 and evaluate the computational cost in Tab. 6. As can be seen, the efficiency can be significantly improved with engineering optimization. Our method can further improve the efficiency since our current implementation still repeats the naive LSS operation.

Table 6. We re-implement our method and evaluate the computational cost. -O denotes the engineering optimization based on BEVPoolv2 [4].

Method	FPS	Backbone	Pooling	Fusion
BEVDepth (R50) [5]	22.6	41.84	2.35	0
SANet (R50)	15.4	42.00	20.02	2.99
SANet-O (R50)	19.1	41.79	7.45	3.00

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

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. 3

[5] 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, 2, 3

[6] 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. 1, 2, 3

[7] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017. 1

[8] 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. 2, 3

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[10] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 1

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[12] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016. 1

[13] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 1

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[16] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 1