BePo: Efficient Dual Representation for 3D Scene Understanding

Yunxiao Shi[1] Hong Cai[1] Jisoo Jeong[1] Yinhao Zhu[1] Steve Han[1] Amin Ansari[2] Fatih Porikli[1]

[1]Qualcomm AI Research* [2]Qualcomm Technologies, Inc.

{yunxshi, hongcai, jisojeon, yinhaoz, shizhan, amina, fporikli}@qti.qualcomm.com

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc

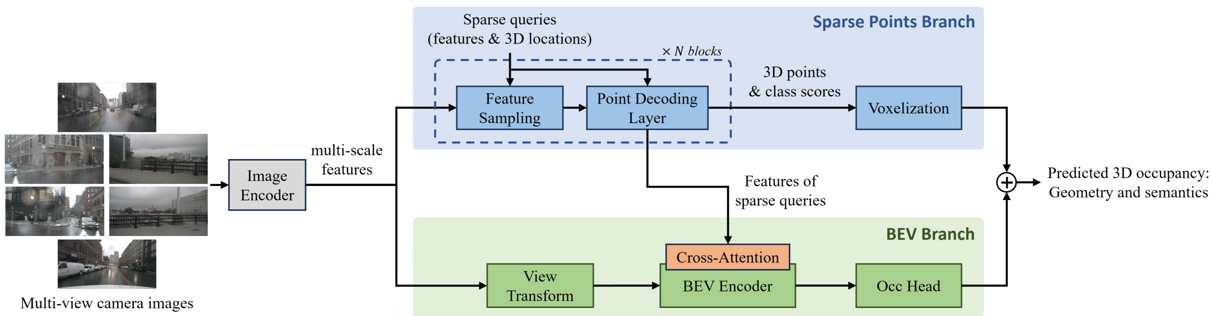

Figure 1. Proposed dual representation using Birds-Eye View (BEV) and 3D sparse points, providing efficient, complementary modeling capabilities to capture both objects in 3D and flat surfaces, which is critical for autonomous perception.

1. Introduction

3D scene understanding [5, 9, 13, 18, 24] forms the foundation of autonomous systems, such as self-driving vehicles and navigation robots. Recently, 3D occupancy prediction has emerged as a new paradigm for scene understanding, which aims to infer fine-grained 3D geometry and semantics from camera images [2, 6, 7, 11, 14, 15, 17, 19–23]. It provides critical scene information with a level of granularity beyond depth estimation and 3D object detection, which is crucial for downstream tasks such as motion planning.

Many existing solutions adopt dense voxel grid [11, 19, 23] as scene representation, followed by cross-attention to aggregate image features, which are then mapped to 3D occupancy. Such design entails significant memory footprint and computational cost, making it difficult to deploy on resource-constrained platforms. To avoid the high computational costs, recent works have adopted the Birds-Eye-View (BEV) representation [4, 22] and demonstrated much improved inference runtime. However, small objects are poorly captured by BEV, as their feature representation after being projected onto the BEV plane is very limited. To mitigate this, another line of research proposed learning the 3D scene as a set of sparse points with learnable queries [10, 17], which demonstrated competitive accuracy and latency. Yet, it is still not sensible to use sparse representation to capture flat surfaces such as the road, which would require a large number of points.

In this work, we propose a new approach, named **BePo**, which combines the advantages of BEV and sparse representations. As shown in Fig. 1, we advocate a dual-branch design, where one branch first adopts effic ient view transform to BEV followed by fast operations such as 2D convolutions for processing, and the other leverages sparse 3D points and a coarse-to-fine decoding scheme. To enable information flow between the two branches, we utilize cross-attention to transfer knowledge from features learned in the points branch to enrich the BEV features. Such learned 3D information from the sparse points can effectively inject more learning signals especially of small objects that have limited feature representation on BEV.

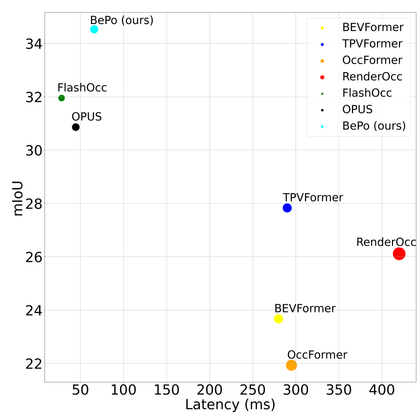

Figure 2. Accuracy (mIoU on Occ3D-nuScenes [1, 15]) vs. inference latency (ms) measured on a single NVIDIA A100 GPU.

By leveraging the dual representation of BEV and sparse points, BePo maintains high efficiency; meanwhile, its stronger 3D modeling power leads to better 3D occupancy prediction performance, as summarized in Figure 2.

Table 1. 3D occupancy prediction results on Occ-ScanNet validation set [21]. **Bold**/Underline: Best/second best results.

Method	IoU	mIoU	Ceiling	Floor	Wall	Window	Chair	Bed	Sofa	Table	TVs	Furniture	Objects
MonoScene [2]	41.60	24.62	15.17	44.71	22.41	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84
ISO [21]	42.16	28.71	19.88	41.88	22.37	16.98	29.09	42.43	42.00	29.60	10.62	36.36	24.61
Ours	**52.73**	**44.91**	**41.32**	**50.29**	**41.83**	**31.81**	**40.37**	**54.65**	**60.71**	**43.76**	**34.27**	**53.33**	**41.72**

Table 2. 3D semantic occupancy prediction mIoU results on Occ3D-nuScenes validation set [1]. **Bold**/Underline: Best/second best results.

Method	mIoU	Others	Barrier	Bicycle	Bus	Car	Cons. Veh	Motorcycle	Pedestrian	Traffic cone	Trailer	Truck	Dri. Sur	other flat	Sidewalk	Terrain	Manmade	Vegetation
MonoScene [2]	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
BEVFormer [7]	23.67	5.03	38.79	9.98	34.41	41.09	13.24	16.50	18.15	17.83	18.66	27.70	48.95	27.73	29.08	25.38	15.41	14.46
TPVFormer [6]	27.83	7.22	38.90	13.67	40.78	**45.90**	17.23	19.99	18.85	14.30	26.69	**34.17**	55.65	35.47	37.55	30.70	19.40	16.78
OccFormer [23]	21.93	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97
RenderOcc [11]	26.11	4.84	31.72	10.72	27.67	26.45	13.87	18.2	17.67	**17.84**	21.19	23.25	63.2	36.42	46.21	44.26	19.58	20.72
FlashOcc [22]	31.95	6.21	39.56	11.27	36.31	43.96	16.25	14.74	16.89	15.76	28.56	30.01	**78.16**	37.52	**47.42**	51.35	**36.79**	31.42
OPUS [17]	30.86	9.68	36.17	15.86	38.65	43.41	21.81	17.21	14.63	15.43	26.92	32.04	71.42	35.96	42.65	41.92	30.61	30.26
Ours	**34.53**	**11.29**	**40.99**	**16.02**	**42.77**	45.54	**25.11**	**21.89**	**21.02**	17.11	**29.93**	32.33	76.84	**37.91**	44.77	**53.12**	36.77	**35.18**

2. Method

BePo employs a dual representation, which combines the strengths of both dense BEV grid and sparse 3D points.

BEV Branch Multi-scale features $\mathcal{F}_{im} \in \mathbb{R}^{C\times H\times W}$ are extracted from the input camera images via an image encoder, which then undergo view transform T to be projected onto BEV. Here we choose T to be LSS [12] given its efficiency. Afterwards, a BEV encoder E consisting of a stack of convolutional layers and an FPN [8] neck are used to process the BEV features to obtain $\mathcal{F}_{bev} \in \mathbb{R}^{C_b\times H_b\times W_b}$.

Sparse Points Branch We randomly initialize a set of learnable queries $\mathbf{Q}$ and 3D points $\mathbf{P}$. $\mathbf{Q}$ and $\mathbf{P}$ are used to sample image features $\mathcal{F}_{im}$ and then processed by several transformer layers. Formally, denote $\mathcal{S}_i = \{\mathbf{Q}_i, \mathbf{P}_i, \mathbf{C}_i\}_{i=0}^{\ell}$ the sets with $\mathcal{C}_i$ being the class scores for $\mathbf{P}_i$, where $\mathcal{S}_0$ is the initial set and $\mathcal{S}_{i>0}$ are the outputs from the i-th decoder stage. ℓ is the number of decoder layers. To reduce computation bottleneck, we follow [17] and make each $q_i \in \mathbf{Q}_i$ predict multiple points instead of one, denoted as M_i. A coarse-to-fine procedure such that $M_{i-1} \le M_i, i = \{1, \dots, \ell\}$ is adopted to facilitate predicting high-level semantics from low-level features.

Cross-Branch Attention and Fusion We compute cross-attention [16] between BEV features $\mathcal{F}_{bev}$ and query features $q_\ell \in \mathbb{R}^{M_i\times C_q}$ from the last decoding stage. Specifically, we treat F_{bev} as queries and q_ℓ as keys and values, injecting the more 3D-aware features into BEV. A linear layer is used to match the embedding dimensions of both sets of features. We fuse the outputs of the two branches to generate the final 3D occupancy prediction.

3. Experiments

Datasets We conduct evaluation based on ScanNet [3] which contains 1,513 room scans, and nuScenes [1] which consists of 1,000 driving scenes, covering both indoor and outdoor scenarios. Specifically, we use Occ-ScanNet [21] which curates 3D occupancy ground truth providing 11 semantic classes and Occ3D-nuScenes [15] which annotates occupancy ground-truth for nuScenes consisting of 17 semantic classes.

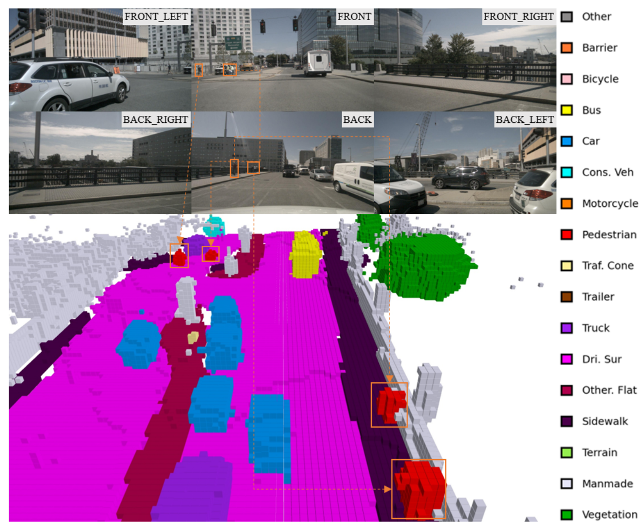

Figure 3. 3D occupancy prediction of our BePo on the Occ3D-nuScenes [15] validation set.

Results Evaluation results on OccScanNet and Occ3D-nuScenes are respectively shown in Table 1 and Table 2. It is evident that BePo improves prediction of difficult objects across the board. On ScanNet, BePo establishes a +17.11 mIoU improvement under the *Objects* category compared to the second-best method. On nuScenes, BePo consistently improves for *Others (+1.61)*, *Motorcycle (+1.90)* and *Pedestrians (+2.17)* on top of second-best, validating the effectiveness of our proposed dual representation.

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. 1, 2

[2] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[4] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023. 1

[5] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1

[6] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 1, 2

[7] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[9] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 1

[10] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. In *Proceedings of the European Confernece on Computer Vision*, 2024. 1

[11] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12404–12411. IEEE, 2024. 1, 2

[12] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2

[13] Yunxiao Shi, Hong Cai, Amin Ansari, and Fatih Porikli. Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 119–129, 2023. 1

[14] Yunxiao Shi, Hong Cai, Amin Ansari, and Fatih Porikli. H3o: Hyper-efficient 3d occupancy prediction with heterogeneous supervision. *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 1

[15] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[16] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[17] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Mingming Cheng. Opus: Occupancy prediction using a sparse set. In *Advances in Neural Information Processing Systems*, 2024. 1, 2

[18] 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. 1

[19] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1

[20] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. *arXiv preprint arXiv:2412.04380*, 2024.

[21] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *European Conference on Computer Vision*, pages 38–54. Springer, 2024. 2

[22] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 1, 2

[23] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 1, 2

[24] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 1