Supplementary Material to "OcRFDet: Object-Centric Radiance Fields for Multi-View 3D Object Detection in Autonomous Driving"

Mingqian Ji, Shanshan Zhang*, Jian Yang
PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology
{mingqianji,shanshan.zhang,csjyang}@njust.edu.cn

In this document, we first provide additional implementation details of our OcRFDet (Sec. 1). We then present extensive comparison experiments (Sec. 2), ablation studies (Sec. 3), and qualitative results (Sec. 4) to further validate the effectiveness of our method. Finally, we discuss the limitations of our approach and provide insights into potential future research directions (Sec. 5).

1. More Implementation Details

We implement the OcRFDet with PyTorch [12] under the framework MMDetection3D [3]. On the nuScenes `validation` dataset, we employ a 60-epoch training scheme, and use the AdamW optimizer [5] with the batch size set to 32 under 4 $\times$ RTX 3090 GPUs. The learning rate of the detector is initialized at 2×10^{-4} while the learning rate of radiance fields is set to 4×10^{-4} with a decay applied every one epoch. Following the baseline, we adopt the default data augmentation techniques. On the nuScenes `test` dataset, we employ 60-epoch training schemes with the batch size set to 8 under 8 $\times$ RTX 3090 GPUs.

2. More Comparison Experiments

More results on the Waymo dataset. We further validate the effectiveness of our method on the Waymo Open dataset [13]. As shown in Tab. 1, our method, using BEVFormer as the baseline, shows consistent improvements on Waymo-full and Waymo-mini about the LET-mAPL and LET-mAPH metrics [6]. Specifically, for the Waymo-full, our method achieves 37.1% LET-mAPL and 51.3% LET-mAPH, which outperforms BEVFormer by 2.1 pp w.r.t. LET-mAPL and 4.2 pp w.r.t. LET-mAPH, and the state-of-the-art method VectorFormer [2] by 0.3 pp w.r.t. LET-mAPL and 2.2 pp w.r.t. LET-mAPH. Moveover, for the Waymo-mini, our method also consistently achieves better results towards the baseline and the state-of-the-art method MvACon [10]. These results further demonstrate the effectiveness of our OcRFDet.

Experiments of other image encoders. To further verify the effectiveness of our OcRFDet, we validate our method with small-sized backbone ResNet-18 [4] and large-sized backbone SwinTransformer-base [11] at the same image resolution of 256 $\times$ 704. As shown in Tab. 2, when using ResNet-18 as the image backbone, our method achieves 32.9% mAP and 39.5% NDS, outperforming our baseline DualBEV by 1.2 pp w.r.t. mAP and 1.8 pp w.r.t. NDS; when using SwinTransformer-base as the image backbone, our method achieves 37.4% mAP and 41.8% NDS, outperforming DualBEV by 1.5 pp w.r.t. mAP and 1.7 pp w.r.t. NDS. These results demonstrate the effectiveness of our OcRFDet with different backbones.

Comparison of efficiency. We compare the running time, computation cost, and parameter size of the baseline method and our OcRFDet. As shown in Tab. 3, we find that our OcRFDet brings significant improvements by 1.6 pp w.r.t. mAP and 0.9 pp w.r.t. NDS with only a minimal increase in computational cost. These results demonstrate our method is friendly to applications.

Comparison at different depths. Following [1], we categorize annotation and prediction ego distances into three groups: Near (0-20m), Middle (20-30m), and Far (>30m). As shown in Tab. 4, compared to the baseline method (DualBEV), our OcRFDet consistently improves performance across all depth ranges. These results indicate our geometric feature enhancement method based on radiance fields is effective across all depth ranges.

Comparison at small-sized objects. To evaluate the effectiveness of our method for small-sized object detection, we conduct experiments on the nuScenes validation dataset, focusing on normal-sized objects at a far distance (>30m) and small-sized objects at a near distance (0-20m). In this context, cars are considered normal-sized objects, while pedestrians, motorcycles, and bicycles are categorized as small-sized objects. As shown in Tab. 5, our method consistently outperforms the baseline DualBEV for the aforementioned small objects. These results indicate that our geometric feature enhancement based on radiance fields significantly improves detection performance for small objects.

*Corresponding author.

Table 1. Comparisons on the Waymo Open dataset.

Methods	Backbone	BEV size	Waymo-full		Waymo-mini	
			LET-mAPL ↑	LET-mAPH ↑	LET-mAPL ↑	LET-mAPH ↑
BEVFormer [9]	ResNet-101	200×200	35.0	47.1	34.9	46.3
MvACon [10]			-	-	35.7	47.5
VoctorFormer [2]			36.8	49.1	-	-
BEVFormer + Ours			**37.1**	**51.3**	**37.3**	**48.5**

Table 2. Experiments of OcRFDet with other image backbones on the nuScenes `validation` set.

Methods	Backbone	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
DualBEV [8]	ResNet-18	31.7	37.7	0.681	0.294	0.614	0.933	0.256
OcRFDet (Ours)	ResNet-18	**32.9**	**39.5**	0.661	0.272	0.616	0.907	0.240
DualBEV [8]	SwinT-Base	35.9	40.1	0.677	0.269	0.534	0.993	0.312
OcRFDet (Ours)	SwinT-Base	**37.4**	**41.8**	0.648	0.266	0.513	0.978	0.418

Table 3. Comparisons of running time, computation cost, and parameter size. For a fair comparison, the running speed of the comparison method is evaluated on one RTX 3090 with a batch size of 1.

Methods	Frames	mAP ↑	NDS ↑	Running Times (FPS)	FLOPs (G)	Parameters (M)
DualBEV [8]	1	35.2	42.5	10.7	291.51	82.86
OcRFDet (Ours)	1	36.8	43.4	8.9	302.61	83.62
DualBEV [8]	2	38.0	50.4	8.2	303.05	83.38
OcRFDet (Ours)	2	40.0	50.9	6.3	314.17	84.14

Table 4. Comparisons at different depths on nuScenes `validation` set. The numbers are **mAP/NDS**.

Methods	Near	Middle	Far
DualBEV [8]	55.5/53.2	26.5/37.4	9.8/25.4
OcRFDet (Ours)	**56.8/53.7**	**28.3/38.3**	**10.2/25.9**

Table 5. Comparisons at small-sized objects on the nuScenes `validation` set. The numbers are **AP**.

Methods	>30m	0-20m		
	Car	Pedestrian	Motorcycle	Bicycle
DualBEV [8]	23.2	55.6	53.2	48.6
OcRFDet (Ours)	**24.7**	**55.8**	**54.7**	**49.2**

3. More Ablation Studies

Effectiveness of cross-attention fusion. As shown in Fig. 6, among the evaluated strategies, cross-attention fusion achieves the highest detection performance of 35.5% mAP and 42.2% NDS, outperforming the other strategies. We think that, compared to linear fusion (weighted mean) or equal-weight fusion (concatenation and convolution), cross-attention fusion further captures more fine-grained associations between the two fields we used, effectively leveraging geometric information. These results demonstrate that cross-attention enables more effective geometry interactions, leading to detection performance improvements.

Effectiveness of multi-scale height slice attention. As shown in Tab. 8, when we progressively introduce multi-scale integration and BEV mask prediction into HSA, the detection performance improves step by step. These results demonstrate the rationality and effectiveness of our multi-scale height slice attention.

Table 6. Ablation studies of the strategy of opacity fusion.

Strategy	mAP ↑	NDS ↑
Weighted Mean	35.3	41.9
Concatenation and Convolution	35.0	42.0
Cross Attention	**35.5**	**42.2**

Table 7. HOA gains on *Car* under different scene IDs (AP).

Methods	ID: `e809···eb64`	ID: `3dd2···6a0a`	All scene
DualBEV + OcRF	44.6	66.1	59.0
+ HOA	$44.7^{\uparrow 0.1}$	$67.7^{\uparrow 1.6}$	$59.8^{\uparrow 0.8}$

Gains of HOA The gains of HOA depend on scene complexity. As shown the Tab. 7, scene `e809···eb64` has only one object category (*Car*), yielding limited improvement (+0.1). In contrast, scene `3dd2···6a0a` contains seven diverse object types, enabling better use of height-wise distinctions and resulting in a larger gain (**+1.6**) on *Car*. This shows HOA is more effective in complex scenes.

Effectiveness of random-view rendering. As shown in Tab. 9, we conduct ablations to validate the rationality of randomly selecting a single viewpoint for rendering. Specifically, when rendering all six viewpoints, we find that

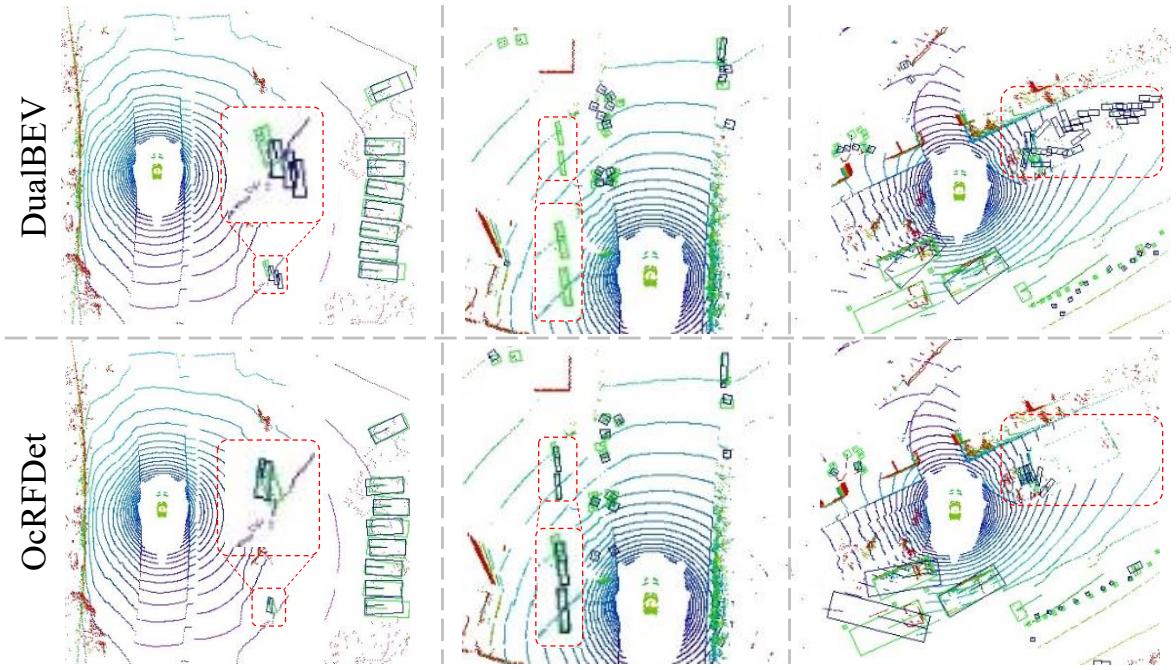

Figure 1. Comparison of detection results in the BEV space on the nuScenes `validation` set. We show the ground truth boxes in green, and the prediction boxes in blue. We use red rectangles to highlight the comparisons of ours and DualBEV.

Table 8. Ablation studies of multi-scale height slice attention.

HSA	Multi Scale	BEV Mask	mAP ↑	NDS ↑
✓			35.2	42.0
✓	✓		35.3	42.1
✓	✓	✓	**35.5**	**42.2**

Table 9. Ablation studies of rendering view selection.

View	mAP	NDS	GPU Memory (G)	Training Time of Each Iteration (s)
6	34.5	41.7	23.9	0.744
4 (Fixed)	34.4	41.4	19.2	0.625
4 (Random)	34.7	41.7	19.2	0.635
2 (Random)	34.7	**42.1**	16.9	0.527
1 (Random)	**35.0**	41.9	12.5	0.459

the detection performance achieves 34.5% mAP and 41.7% NDS, with no significant improvement. When rendering four fixed viewpoints (FRONT, FRONT_LEFT, BACK, and BACK_RIGHT), the computational cost is reduced, but the detection performance drops by 0.1 pp w.r.t. mAP and 0.3 pp w.r.t. NDS. When rendering four random viewpoints, the detection performance improves by 0.3 pp w.r.t. mAP and 0.3 pp w.r.t. NDS. Furthermore, as fewer viewpoints are randomly selected for rendering, the detection performance further improves, and the computational cost progressively decreases. We analyze these results as follows: Rendering all six viewpoints focuses the network's optimization on the radiance fields, limiting the optimization of the detection network while incurring high computational costs. Rendering randomly selected viewpoints ensures that all six viewpoints are rendered throughout the training process, making it more beneficial for detection compared to rendering the same number of fixed viewpoints. Finally, to balance detection performance and computational efficiency, we identify that randomly rendering a single viewpoint is optimal.

4. More Qualitative Results

We further compare the qualitative results of DualBEV and ours in the BEV space on the nuScenes `validation` set. As shown in Fig. 1, in the left column, our method shows a more accurate location of the predicted boxes for the distant objects. In the middle column, our method successfully obtains the detection boxes for occluded objects. In the right column, our method produces fewer false positive boxes. These results further demonstrate the superiority of our OcRFDet.

5. Limitation and Future Work

While our object-centric radiance fields effectively enhance foreground objects and suppress background noise, the learning process is heavily dependent on the quality of annotations. This leads to two interrelated constraints: (1) Critical foreground instances that lack 3D box labels (e.g.,

traffic lights in the nuScenes dataset) may be inadvertently suppressed during feature enhancement, which limits the applicability of our method in open environments. (2) Inaccurate 3D box labels in the training data could lead our method to unintentionally amplify background features that should have been suppressed.

Future work will explore leveraging prompts as priors in open-set datasets to preserve important yet unlabeled foreground object features, while addressing label noise through temporal consistency constraints in dynamic scenes. In addition, as more advanced U-Net architectures [7] continue to emerge, we envision integrating them to further refine the opacity-based attention mechanism.

References

[1] Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Objectfusion: Multi-modal 3d object detection with object-centric fusion. In *ICCV*, 2023. 1

[2] Zhili Chen, Shuangjie Xu, Maosheng Ye, Zian Qian, Xiaoyi Zou, Dit-Yan Yeung, and Qifeng Chen. Learning high-resolution vector representation from multi-camera images for 3d object detection. In *ECCV*, 2024. 1, 2

[3] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE TPAMI*, 2023. 1

[7] Jiaxin Li, Ke Zheng, Lianru Gao, Zhu Han, Zhi Li, and Jocelyn Chanussot. Enhanced deep image prior for unsupervised hyperspectral image super-resolution. 2025. 4

[8] Peidong Li, Wancheng Shen, Qihao Huang, and Dixiao Cui. Dualbev: Cnn is all you need in view transformation. In *ECCV*, 2024. 2

[9] 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 *ECCV*, 2022. 2

[10] Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. Multi-view attentive contextualization for multi-view 3d object detection. In *CVPR*, 2024. 1, 2

[11] 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 *ICCV*, 2021. 1

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1

[13] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1