Ray Denoising: Depth-aware Hard Negative Sampling for Multi-view 3D Object Detection –Supplementary Material–

Feng Liu[1⋆] Tengteng Huang[2] Qianjing Zhang[2] Haotian Yao[2] Chi Zhang[2]
Fang Wan[1] Qixiang Ye[1] Yanzhao Zhou[1†]

[1] University of Chinese Academy of Sciences
liufeng20@mails.ucas.ac.cn {wanfang,qxye,zhouyanzhao}@ucas.ac.cn
[2] Mach Drive
{tengteng.huang,qianjing.zhang,haotian.yao,chi.zhang}@mach-drive.com

0.1 About the Scalability.

To verify the effectiveness of RayDN when further scaling up the backbone and image size, we conduct experiments with ViT-L on nuScenes test set. Models are trained for 24 epochs. As shown in Table 1, RayDN outperforms the Stream-PETR [10] by 1.1% mAP and 1.0 % NDS. demonstrating the scalability and effectiveness of Ray Denoising, *i.e.*, RayDN.

Table 1: Comparison on the nuScenes test set.

Methods	**Backbone**	**Image Size**	**mAP**	**NDS**
BEVDepth [4]	ConvNext-B	640×1600	52.0	60.9
AeDet [2]	ConvNext-B	640×1600	53.1	62.0
PETRv2 [6]	RevCol-L	640×1600	51.2	59.2
SOLOFusion [7]	ConvNeXt-B	640×1600	54.0	61.9
BEVFormerv2 [11]	InternImage-XL	640×1600	55.6	63.4
BEVDet4D-Gamma [3]	Swin-B	900×1600	58.6	66.4
StreamPETR [10]	ViT-L	800×1600	62.0	67.6
RayDN (Ours)	ViT-L	800×1600	**63.1**	**68.6**

0.2 About the Generalization Ability to Other Model.

We conduct experiments with more models to verify the generalization ability of RayDN. We adopt ResNet50 pre-trained on nuImages [1] as the backbone and the image size is 256×704. Models are trained for 24 epochs. As shown in Table 2, RayDN obtains 1.9% mAP and 1.2% mAP against PETR [5] and FocalPETR [9] separately, demonstrating the generalization ability of RayDN.

⋆ Work was done during internship at Mach Drive. † Corresponding Author.

Table 2: Ablation studies on the generalization ability of RayDN.

Method	Backbone	Image Size	mAP	NDS
PETR [5]	ResNet50	256×704	33.3	36.4
+ RayDN (Ours)	ResNet50	256×704	**35.2**	**37.3**
FocalPETR [9]	ResNet50	256×704	34.9	38.7
+ RayDN (Ours)	ResNet50	256×704	**36.1**	**39.9**

Table 3: Ablation studies on the training time and inference speed.

Method	backbone	Image Size	Training Time	FPS
SOTA Baseline [10]	ResNet50	704×256	7 h	10.4
+ 3DPPE [8]	ResNet50	704×256	8.5 h	9.9
+ RayDN (Ours)	ResNet50	704×256	7.5h	10.4

Cost of Ray Denoising. We analyze the computational overhead of Ray Denoising by comparing training times and inference speeds, as detailed in Table 3. Training time is benchmarked across 8 GeForce RTX 2080 Ti GPUs, while inference speed is measured on a single GeForce RTX 2080 Ti GPU. Our setup utilizes a ResNet50 backbone with an input resolution of 256×704. Ray Denoising introduces a modest increase in training time—just a 7% rise compared to StreamPETR—while 3DPPE raises it by 21%. Inference speed remains on par with StreamPETR, as Ray Denoising is only used in the training phase.

0.3 More Visualization of Detection Results.

We visualize more detection results in Figure 1. As can be seen, RayDN works well in both daytime and night.

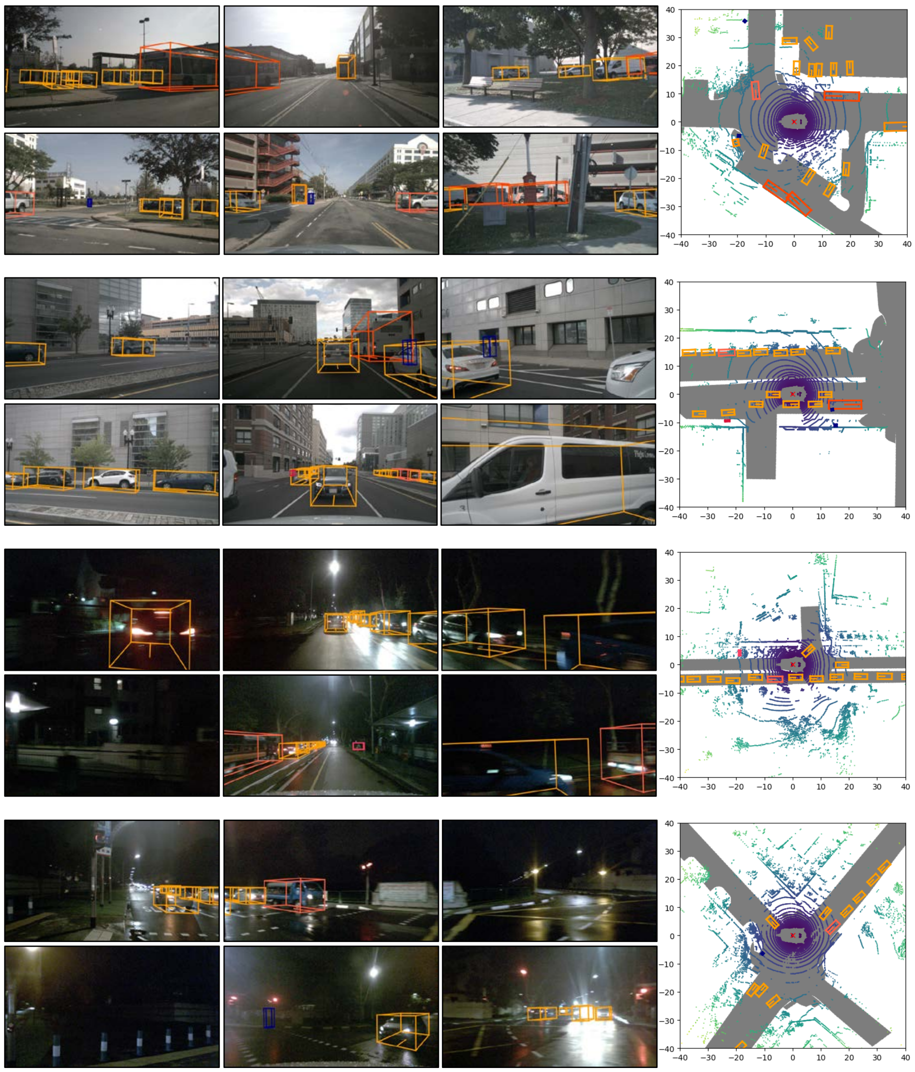

Fig. 1: Visualization of the detection results. RayDN works well under different lighting conditions (daytime, night) to suppress duplicate false positives while maintaining the ability to detect highly occluded objects on the same ray. Best viewed by zooming on the screen.

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. Feng, C., Jie, Z., Zhong, Y., Chu, X., Ma, L.: Aedet: Azimuth-invariant multi-view 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21580–21588 (2023)
3. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022)
4. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1477–1485 (2023)
5. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: European Conference on Computer Vision. pp. 531–548. Springer (2022)
6. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3262–3272 (2023)
7. Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K.M., Tomizuka, M., Zhan, W.: Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In: The Eleventh International Conference on Learning Representations (2022)
8. Shu, C., Deng, J., Yu, F., Liu, Y.: 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3580–3589 (October 2023)
9. Wang, S., Jiang, X., Li, Y.: Focal-petr: Embracing foreground for efficient multi-camera 3d object detection. arXiv preprint arXiv:2212.05505 (2022)
10. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3621–3631 (October 2023)
11. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., et al.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17830–17839 (2023)