RARE: Learn to RAnk and REtrieve for Monocular 3D Object Detection

Supplementary Material

Hyeonjeong Park[1] Peixi Xiong[2] Xiaoqian Ruan[1] Dian Jia[1] Pei Yu[3] Wei Tang[1]
[1]University of Illinois Chicago [2]Intel [3]Microsoft
{hpark233,xruan9,djia7,tangw}@uic.edu, peixi.xiong@intel.com, pei.yu@microsoft.com

A. Overview

- Appendix B: More Details of the Architecture
- Appendix C: More Details of the Loss Functions and Implementations
- Appendix D: More Experimental Results and Analysis
- Appendix E: Qualitative Results
- Appendix F: Limitation

B. More Details of the Architecture

Feature Extraction and RoI Localization. Given an RGB image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote its height and width, a 2D backbone (e.g., DLA34 [8]) extracts multi-scale feature maps $\{\boldsymbol{f}_{1/4}, \boldsymbol{f}_{1/8}, \boldsymbol{f}_{1/16}\}$, where the subscript denotes the spatial resolution relative to the input. These features are used for RoI localization and also fed into a multi-scale deformable transformer encoder. For RoI localization, we first bilinearly resize $\boldsymbol{f}_{1/8}$ and $\boldsymbol{f}_{1/16}$ to the $1/4$ scale (with a lightweight upsampling block), channel-align all maps with 1×1 convs, and then average them to obtain the aggregated feature $\bar{\boldsymbol{f}}_{1/4}$. Using this, we predict a 2D heatmap $\boldsymbol{H} \in \mathbb{R}^{H/4\times W/4\times B}$, where B is the number of categories, a 2D size map $\boldsymbol{S}^{\text{2D}} \in \mathbb{R}^{H/4\times W/4\times 2}$ for bounding box dimensions, and a 2D offset map $\boldsymbol{O}^{\text{2D}} \in \mathbb{R}^{H/4\times W/4\times 2}$ for positional adjustment. We retain top-N RoIs based on the centerness scores. Finally, RoI-Align is applied to extract RoI features $\boldsymbol{F}^{\text{roi}} \in \mathbb{R}^{N\times M\times M\times C}$. $M \times M$ denotes the RoI region size (typically M=7) and $C = 256$ is the feature dimension. The RoI features are used to generate multiple object-dependent queries.

Multi-scale Deformable Transformer Encoder. To capture long-range dependencies and multi-scale context efficiently, we employ a deformable transformer encoder. First, the multi-scale features $\{\boldsymbol{f}_{1/4}, \boldsymbol{f}_{1/8}, \boldsymbol{f}_{1/16}\}$ are projected to a shared hidden dimension d^{model} (*i.e.*, 256) using 1×1 convolutions. Since the transformer architecture is permutation-invariant, fixed sinusoidal positional encodings and learnable scale-level embeddings are added to the features. Finally, all features from each level are flattened and concatenated into a single feature sequence. These enhanced features are then processed by the encoder layers, which utilize multi-scale deformable self-attention mechanisms to aggregate information across different spatial locations and scales. The output of the encoder serves as the encoder memory $\boldsymbol{V} \in \mathbb{R}^{(H_{1/4}W_{1/4}+H_{1/8}W_{1/8}+H_{1/16}W_{1/16})\times d^{\text{model}}}$. This global memory $\boldsymbol{V}$ is subsequently used in the decoder.

Multi-scale Deformable Transformer Decoder. Unlike standard DETR approaches that use learnable query embeddings, we generate content-dependent queries directly from the RoI features. Technically, we construct a compact query set for each RoI in order to handle the ill-posed nature of monocular 3D detection. Specifically, we employ K distinct projection heads to flatten and map the enriched RoI features into K content-dependent queries $\boldsymbol{q}_{i,k} \in \mathbb{R}^{d^{\text{model}}}$, where i is the index of RoI. Simultaneously, the normalized 2D center coordinates of the RoI are used as reference points $\boldsymbol{p}_i^{\text{ref}} \in [0,1]^2$. We generate positional queries by encoding these reference points with sinusoidal functions and passing them through K separate MLPs and LayerNorms, denoted as $\boldsymbol{e}_{i,k}$. The decoder then takes the content queries $\boldsymbol{q}_{i,k}$, positional embeddings $\boldsymbol{e}_{i,k}$, and reference points $\boldsymbol{p}_i^{\text{ref}}$ as input. It performs deformable cross-attention over the encoder memory $\boldsymbol{V}$ to refine the queries, resulting in hypothesis-specific object features that fuse local RoI cues with global, multi-scale context.

The final query features are passed into task-specific prediction heads to estimate 3D parameters. For the k-th query of the i-th RoI, the heads regress: 3D dimensions $\boldsymbol{x}_{i,k} \in \mathbb{R}^3$; orientation $\theta_{i,k} \in \mathbb{R}^{m\times 2}$ using a multi-bin approach [5] with m bins; offsets for the 3D center projection $\boldsymbol{o}_{i,k} \in \mathbb{R}^2$; and depth $d_{i,k}$. Finally, a confidence head predicts a confidence score $c_{i,k} \in (0,1)$ via a sigmoid activation.

IoU /D.	Method	AP_{3D}				APH_{3D}			
		All	0-30	30-50	50+	All	0-30	30-50	50+
0.7 /L1	Baseline	2.28	6.11	0.73	**0.07**	2.26	6.06	0.72	**0.07**
	Ours	**3.02**	**8.13**	**0.77**	**0.07**	**3.00**	**8.09**	**0.76**	**0.07**
0.7 /L2	Baseline	2.14	6.09	0.70	**0.06**	2.12	6.04	0.69	**0.06**
	Ours	**2.83**	**8.10**	**0.74**	**0.06**	**2.81**	**8.05**	**0.73**	**0.06**

Table 1. **Comparison on the *Vehicle* category on the Waymo validation set.** 'D.' denotes difficulties (L1=Level_1, L2=Level_2).

C. More Details of the Loss Functions and Implementations

Loss Functions. The model is trained end-to-end using a multi-task objective function: $\mathcal{L}^{\text{all}} = \lambda^{\text{2D}}\mathcal{L}^{\text{2D}} + \lambda^{\text{3D}}\mathcal{L}^{\text{3D}} + \lambda^{\text{rank}}\mathcal{L}^{\text{rank}}$. For RoI localization loss $\mathcal{L}^{\text{2D}}$, we follow CenterNet [9] and apply loss functions for the heatmap estimation, 2D size regression, and 2D center offset regression, respectively. For 3D detection losses $\mathcal{L}^{\text{3D}}$, we have a 3D box multi-hypothesis loss with a common 3D box regression loss ℓ^{box}, including (1) an $L1$ loss between the predicted 3D size and the ground truth; (2) a MultiBin loss [1], which classifies the heading angle into bins and regresses the residual within the selected bin; (3) a Smooth-$L1$ loss on the 3D center projection offsets; and (4) an uncertainty-regularized depth loss [3, 5]. Details for the 3D box multi-hypothesis loss and the ranking-based confidence loss $\mathcal{L}^{\text{rank}}$ can be seen in the main paper.

C.1. More Implementation Details

Pre-processing. Following the conventional settings [3, 4, 6], KITTI images with a resolution of [370, 1242] and nuScenes images with a resolution of [900, 1600] are resized to [384, 1280] and [384, 672], respectively.

Training. The backbone model is initialized with ImageNet weights, while the detection heads are initialized using the Xavier algorithm [2]. We employ a cosine warmup strategy for the first 10 epochs. We set weight decay to 0.00001, and the learning rate decays by a factor of 0.1 at the 250th, 450th, and 700th epochs. For the pair-wise ranking loss, we only select pairs where the absolute difference in their 3D IoUs exceeds a threshold of 0.1 to ensure a discernible quality gap between candidates.

Inference. We limit the maximum number of objects per image to 50. For filtering, we discard RoIs with category confidence lower than 0.2 for KITTI and 0.1 for nuScenes.

D. More Experimental Results and Analysis

Results on *Vehicle* Category of Waymo Validation Set. We further validate our method on the large-scale Waymo [7] dataset, consisting of 52,386 train and 39,848 val front images. We evaluate a 'Vehicle' category at Level 1/2 (difficulties, 'D.') over distance ranges [0,30), [30,50), [50,∞), and 'All', and report AP_{3D} and APH_{3D}. Tab. 1 shows that our method achieves clear improvements over the baseline, particularly in 'All' and the 0–30m range, while remaining comparable at longer distances.

# $\mathcal{P}$	Car (E/M/H)	Ped. (E/M/H)	Cyc. (E/M/H)
16	26.02/19.42/16.53	12.23/9.58/7.17	8.15/4.28/3.91
32	**28.58/22.05/19.21**	**13.63/10.62/8.46**	**9.54/4.96/4.79**
64	28.04/20.46/17.32	11.99/9.08/7.37	6.01/3.24/3.05

Table 2. **Impact of the number of sampled pairs $\mathcal{P}$ per class for the ranking loss** across object categories on the KITTI validation set.

# K	Car (E/M/H)	Ped. (E/M/H)	Cyc. (E/M/H)
2	26.26/19.71/16.85	13.13/10.18/8.03	7.54/3.98/3.78
3	**28.58/22.05/19.21**	**13.63/10.62/8.46**	**9.54/4.96/4.79**
4	26.03/19.64/15.73	13.17/9.63/7.89	6.89/3.59/3.47

Table 3. **Impact of the number of queries K per object** across object categories on the KITTI validation set.

Impact of Hyper-parameters. Tab. 2 reports the effect of the number of sampled pairs $\mathcal{P}$ per class for the ranking loss using the KITTI validation set. With only 16 pairs, the ranking signal is relatively weak and leads to suboptimal gains. Sampling 32 pairs strikes the best balance, providing rich yet stable supervision and achieving the highest AP_{3D} across all difficulty levels. Using 64 pairs does not bring additional benefit and even hurts performance, indicating that overly dense pairwise supervision can introduce noise and optimization overhead. Overall, these results show that RARE is most effective with a compact candidate set and a moderate number of informative pairs, rather than aggressively increasing either quantity.

Tab. 3 studies the number of queries K per object using the KITTI validation set. Using two queries already improves over the baseline but remains insufficient to capture the multimodal nature of 3D geometry. Increasing K to three yields the best performance, indicating that a small set of compact, content-dependent hypotheses is sufficient. Further increasing K degrades performance, likely due to redundant or low-quality candidates that make ranking more difficult.

Tab. 4 further analyzes the effect of K on depth estimation across distance ranges. Consistent with AP_{3D}, K=3 achieves the lowest MAE on both KITTI and nuScenes. The improvement is more pronounced at mid- and far-range distances (20–40 m and 40+ m), where monocular ambiguity is higher. Using fewer queries leads to higher errors due to limited hypothesis coverage, while larger K introduces redundant candidates that hinder reliable selection.

E. Qualitative Results

Visualization of Candidate Sets Produced by RARE. For each RoI, RARE outputs a compact set of 3D candidate boxes instead of a single estimate. As shown in Fig. 2, can-

# K	KITTI Val. (Depth MAE)				nuScenes Val. (Depth MAE)			
	0–20	20–40	40+	All	0–20	20–40	40+	All
2	0.40	1.08	1.89	0.92	0.61	1.77	4.64	1.52
3	**0.35**	**0.94**	1.67	**0.69**	**0.59**	**1.48**	**4.03**	**1.05**
4	0.40	1.05	**1.61**	0.83	**0.59**	1.55	5.09	1.17

Table 4. **Impact of the number of queries K on depth estimation accuracy (MAE)** across distance ranges on KITTI and nuScenes validation sets.

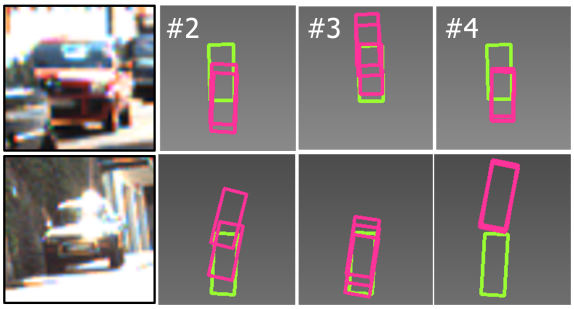

Figure 1. **Effect of the number of queries on candidate diversity.** Visualization of candidate sets with different query numbers. Using fewer queries results in insufficient coverage of plausible 3D configurations, while more queries produce highly overlapping hypotheses. The best performance is achieved at K=3, which balances coverage and redundancy.

didates for nearby objects are highly consistent and almost overlap with the ground truth box, reflecting low ambiguity at short range. In contrast, for distant objects (highlighted in the BEV views), the candidates spread out along the depth axis and form multiple plausible hypotheses that cover a wider depth range. This behavior is clearly visible in the BEV space, where the stacked boxes illustrate how RARE explicitly models depth ambiguity for far-range objects.

Qualitative Comparison across the Number of Queries. We further visualize candidate sets under different numbers of queries. As shown in Fig. 1, using too few queries results in limited coverage, where the candidate set fails to capture all plausible 3D configurations. In contrast, increasing the number of queries leads to highly overlapping hypotheses, introducing redundancy without improving coverage. These observations are consistent with our quantitative results, where K=3 achieves the best performance. It provides a good balance between sufficient mode coverage and limited redundancy, yielding a compact yet expressive set of candidates.

Qualitative Comparison on KITTI. Fig. 3 shows a qualitative comparison with the baseline (sky blue boxes) on the KITTI validation set. In the upper scene, the rearmost car is heavily occluded by another vehicle, causing the baseline to miss the object entirely, whereas RARE (pink boxes) still produces a correct 3D detection with reasonable depth and orientation. In the lower scene, both methods detect the closer car, but the box from RARE better overlaps with the ground truth (green boxes) in the BEV space, while the baseline prediction is noticeably misaligned. Moreover, the distant car (approximately 50m away) is only detected by RARE, indicating improved robustness to small, far-range objects.

F. Limitation

The current RARE framework employs a fixed query set size for each RoI, which may be suboptimal for both very easy and extremely difficult cases, and incurs additional computational costs. Designing methods that dynamically adapt the size of the query set per RoI is an important direction for future work.

References

[1] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 2

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 2

[3] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 664–683. Springer, 2022. 2

[4] Zhenjia Li, Jinrang Jia, and Yifeng Shi. Monolss: Learnable sample selection for monocular 3d detection. In *2024 International Conference on 3D Vision (3DV)*, pages 1125–1135. IEEE, 2024. 2

[5] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 1, 2

[6] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *European Conference on Computer Vision*, pages 71–88. Springer, 2022. 2

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[8] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 1

[9] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

Figure 2. **Visualization of candidate sets produced by RARE on the KITTI validation set.** Each row displays the RGB image with projected 3D boxes (left) and the corresponding Bird's-Eye View (BEV) plot (right). For each RoI, RARE generates a compact set of candidates: for nearby objects, the hypotheses tightly align with the ground truth, whereas for distant objects, they exhibit variance along the depth axis in BEV, effectively capturing the inherent depth ambiguity. Legend: (Car, Cyclist, and Pedestrian).

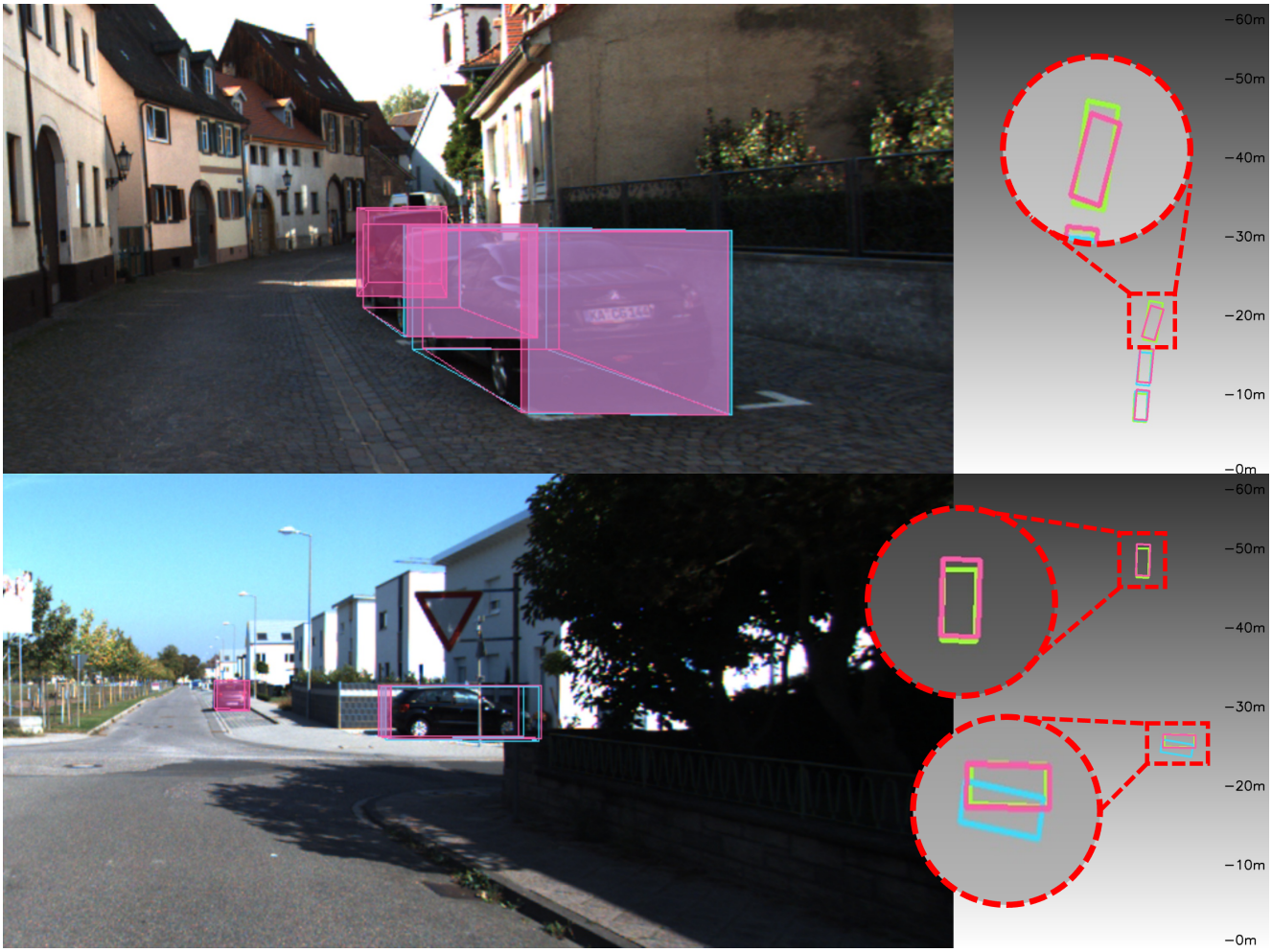

Figure 3. **Qualitative comparison of RARE to the baseline on the KITTI validation set.** Each row displays the RGB image with projected 3D boxes (left) and the corresponding Bird's-Eye View (BEV) plot (right). Compared to the baseline, RARE demonstrates superior robustness in challenging scenarios. Notably, RARE successfully recovers heavily occluded objects (top row) and detects distant vehicles (bottom row) that the baseline misses, while also exhibiting tighter alignment with the ground truth for nearby objects.