Supplementary Material for RepVF: A Unified Vector Fields Representation for Multi-task 3D Perception

Chunliang Li[1]*, Wencheng Han[2]*, Junbo Yin[1],
Sanyuan Zhao[1]†, and Jianbing Shen[2],

[1] School of Computer Science, Beijing Institute of Technology
[2] SKL-IOTSC, Computer and Information Science, University of Macau, China
`{jbji, sanyuanzhao}@bit.edu.cn`
`{wencheng256, yinjunbocn shenjianbingcg}@gmail.com`
`jianbingshen@um.edu.mo`

A Overview

This document is organized as follows:

- Appendix B contains further experimental results. We have investigated the efficiency of our single-head multi-task design, and compared different conversion functions.
- Appendix C describes more details of the experiment.
- Appendix D provides more visualization examples.

B More Experiment Results

B.1 Efficiency of Single-head Multi-task Design

Based on *RepVF*, our proposed RFTR is efficient and faster than previous approaches, as shown in Tab. 1. Notably, we observe no efficiency loss of RFTR even compared to single task experts [1, 5, 7], and a huge efficiency gain in fair comparison.

Table 1: FPS benchmark, FLOPS, and parameters of models.

Model	3D Object	3D Lane	FPS↑	GFLOPS↓	Params(M)↓
DETR3D [7]	✓		10.6	82.8	53.3
PETR [5]	✓		12.0	49.8	36.7
+Anchor [1, 5]		✓	11.2	50.1	36.8
Multi-head [1, 5]	✓	✓	5.1	51.6	46.7
RFTR-bbox	✓		14.6	49.7	37.4
RFTR-lane		✓	13.9	49.7	37.4
RFTR	✓	✓	12.0	49.7	37.4

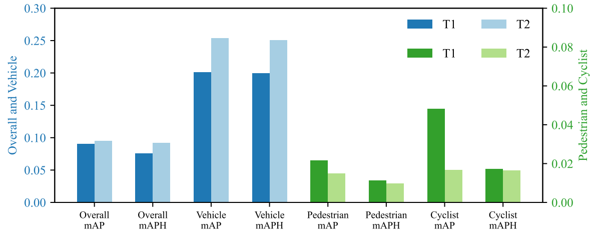

Fig. 1: Comparison of $\mathcal{T}_1$ and $\mathcal{T}_2$ for 3D object detection.

In terms of single-head multi-task design, RFTR exhibits impressive efficiency improvement over the traditional multi-head paradigm [1,5], 135% faster (+6.9 FPS) while saving 19.9% model parameters. Specifically, *RFTR-bbox* is 21.7% and 37.7% faster (+2.6 and +4.0 FPS) respectively compared to its single task counterparts [5,7], with almost equal or less computational cost. And *RFTR-lane* is 24.1% faster(+2.7 FPS) than its traditional representation baseline [1,5].

Here, plus anchor design is extended PETR [5] with traditional 3D lane anchors from [1], and multi-head design is [5] with two task specific heads under task specific prediction formats [1–3,7]. *RFTR-bbox* and *RFTR-lane* are **single-task** versions of our RFTR for 3D objects and 3D lanes, respectively. GFLOPS stands for giga (10^9) FLOPS (floating point operations per second), and the parameter scales are in millions (M).

B.2 Conversion Functions

In Fig. 1, we compared two candidate conversion functions $\mathcal{T}^b_{lwh}$ for 3D bounding box dimension computation, the **min-max** $\mathcal{T}_1$ and the **momentum-based** $\mathcal{T}_2$, and found that $\mathcal{T}_2$ works better for both overall and vehicle, but is slightly worse for smaller targets.

C More Experiment Details

More Implementation Details. We adopt the momentum-based $\mathcal{T}^b_{lwh} = \mathcal{T}_2$ for $\mathcal{T}^b$, and demonstrate its difference from min-max $\mathcal{T}_1$. FPS is calculated on a single RTX4090, and Params/FLOPS experiments are calculated based on input shape (5, 3, 512, 1408). Lightweight ablation models are implemented with half decoder layers.

Evaluation Metrics. *(1)* For 3D object detection, the official 3D object detection metrics in the Waymo Open Dataset [6] are 3D mean Average Precision (mAP) and mAP weighted by heading accuracy (mAPH). The Intersection over Union (IoU) thresholds for both metrics are set to 0.7 for vehicles and 0.5 for

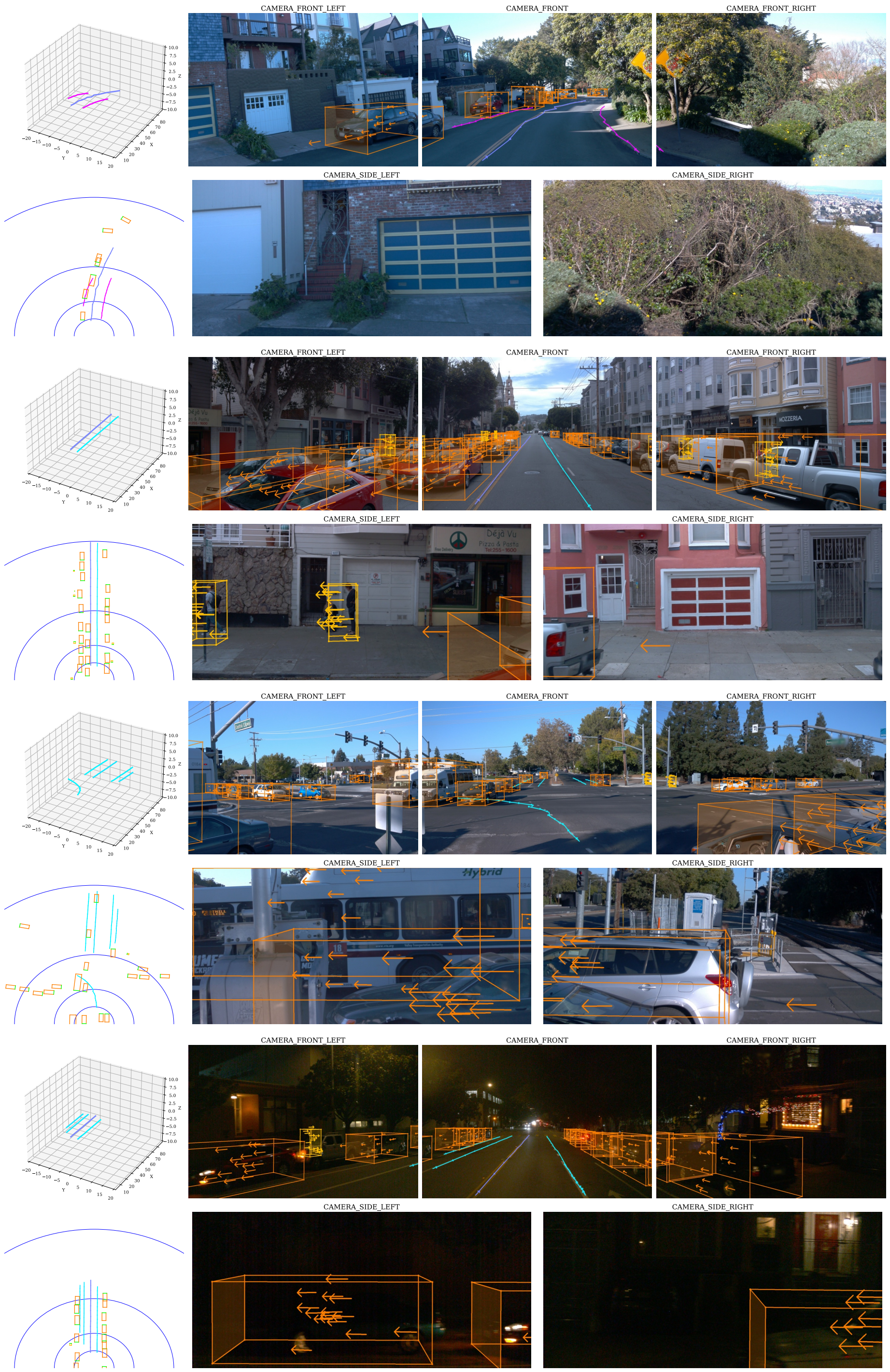

Fig. 2: More visualization results of our RFTR.

pedestrians/cyclists. The test samples are divided into two ways, depending on the distance of the objects and the number of lidar points. The first way includes 0-30m, 30-50m and >50m. The second way divides the task into two difficulty levels, LEVEL_1 for boxes with more than five LiDAR points and LEVEL_2 for boxes with at least one LiDAR point. We report the level 1 in the second way division, since our experiments do not use lidar data. *(2)* For 3D lane detection, following the official OpenLane [1,4] metrics, we use bipartite matching between the predicted and ground truth to evaluate 3D lanes. The pointwise Euclidean distance is computed at y-positions that are simultaneously covered by both the prediction and the ground truth. For a predicted lane to be considered a match, at least 75% of its covered y-positions must have a pointwise Euclidean distance of less than the maximum allowed threshold of 1.5 meters. We document the F-score, together with the errors in different ranges: near range (3-40m) and far range (>40m).

D More Visualizations

In this section, we provide more diverse visualizations (Figure 2). Our RFTR could produce robust and high quality predictions under different weather or time conditions.

References

1. Chen, L., Sima, C., Li, Y., Zheng, Z., Xu, J., Geng, X., Li, H., He, C., Shi, J., Qiao, Y., Yan, J.: PersFormer: 3D Lane Detection via Perspective Transformer and the OpenLane Benchmark. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, vol. 13698, pp. 550–567. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19839-7_32
2. Garnett, N., Cohen, R., Pe'er, T., Lahav, R., Levi, D.: 3D-LaneNet: End-to-End 3D Multiple Lane Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2921–2930. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00301
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (Jun 2012). https://doi.org/10.1109/CVPR.2012.6248074
4. Guo, Y., Chen, G., Zhao, P., Zhang, W., Miao, J., Wang, J., Choe, T.E.: Gen-LaneNet: A Generalized and Scalable Approach for 3D Lane Detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 666–681. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_40
5. Liu, Y., Wang, T., Zhang, X., Sun, J.: PETR: Position Embedding Transformation for Multi-view 3D Object Detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, vol. 13687, pp. 531–548. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19812-0_31

6. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in Perception for Autonomous Driving: Waymo Open Dataset
7. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. In: Proceedings of the 5th Conference on Robot Learning. pp. 180–191. PMLR (Jan 2022)