LowRankOcc: Tensor Decomposition and Low-Rank Recovery for Vision-based 3D Semantic Occupancy Prediction Supplementary Material

Linqing Zhao[1,2], Xiuwei Xu[1], Ziwei Wang[1], Yunpeng Zhang[3], Borui Zhang[1], Wenzhao Zheng[1], Dalong Du[3], Jie Zhou[1], Jiwen Lu[1*]
[1]Department of Automation, Tsinghua University, China
[2]School of Electrical and Information Engineering, Tianjin University, China
[3]PhiGent Robotics
linqingzhao@tju.edu.cn; {xxw21,wang-zw18,zhang-br21,zhengwz18}@mails.tsinghua.edu.cn;
yunpengzhang97@gmail.com; dalong.du@phigent.ai; {jzhou, lujiwen}@tsinghua.edu.cn

This supplementary material is organized as follows:

- Section A demonstrates the statistical basis of introducing Vertical-Horizontal (VH) decomposition for reducing spatial redundancies in 3D semantic occupancy prediction.
- Section B elaborates on the influence of the VH rank R through qualitative and quantitative results and comparisons.
- Section C describes the limitations of our LowRankOcc.

A. Statistical Basics of VH Decomposition

The fundamental concept of our VH decomposition is to factorize 3D tensors into a combination of vertical vectors and horizontal matrices. Our motivation stems from the observation of a low information density along the vertical axis. In the manuscript, we substantiate this observation experimentally by comparing the recovery capacity of the 3-way VM Decomposition with our 1-way VH Decomposition. In this section, we seek to provide an additional demonstration of this observation from a statistical perspective.

In this section, we utilize the frequency statistics of unique semantic categories as an indicator of information density, where a higher number of unique semantic categories implies more complex semantic relations. Hence, we delve into the frequency statistics of unique semantic categories along the vertical axis of the SemanticKITTI [1] validation set. The count is determined by the number of distinct semantic labels along the z-axis, excluding the 'empty voxel' class. Notably, we find that 99.99% of all pillars contain only 3 or fewer unique semantic categories. Despite this, existing methods often treat the vertical axis much like horizontal planes, neglecting to condense representations along the vertical axis. In contrast, compressing the representation parameters along the vertical dimension is shown to be a succinct yet effective alternative to voxel representations [3] and TPV representations [2].

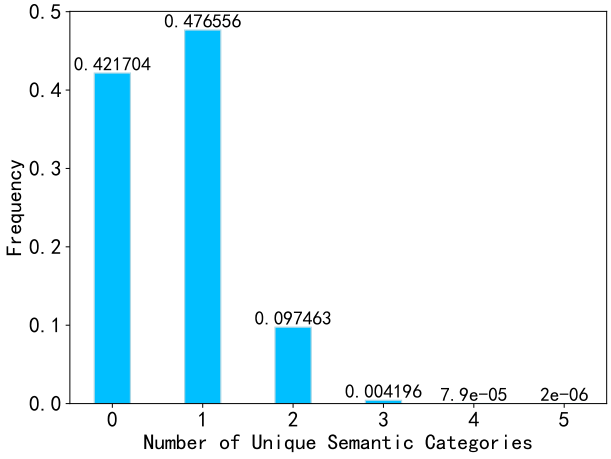

Figure 1. The frequency statistics of unique semantic categories in the vertical direction of the SemanticKITTI validation set. The count is based on the number of distinct semantic labels along the z-axis (pillar), excluding the 'empty voxel' class. Notably, **99.99%** of all pillars consist of only **3** or fewer unique semantic categories. Despite this, conventional methods [2, 3] typically treat the vertical axis similarly to horizontal planes, neglecting to compress representations along the vertical axis.

B. Comparisons of Different VH Ranks

Figure 2 illustrates that an increase in the VH rank R corresponds to an enhanced representation capability. When combined with our Recursive Residual Decomposition

*Corresponding author.

strategy, a larger R enables the capture of more high-frequency details. This includes small objects, such as the person in the 1st column, fine-grained geometries represented by the cars in the 2nd and 4th columns, and connected relations like the road in the 3rd column.

C. Limitations

While our Recursive Residual Decomposition strategy enhances the representations of our VH components by capturing more high-frequency details, the recursive and serial decomposition design leads to a slowdown in the forward speed of LowRankOcc. Therefore, our method achieves inference speeds comparable to existing approaches, but exhibits a large improvement compared to methods based on 3D convolution. The specific numerical values can be found in the ablation study reported in our manuscript.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019. 1

[2] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2302.07817*, 2023. 1

[3] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023. 1

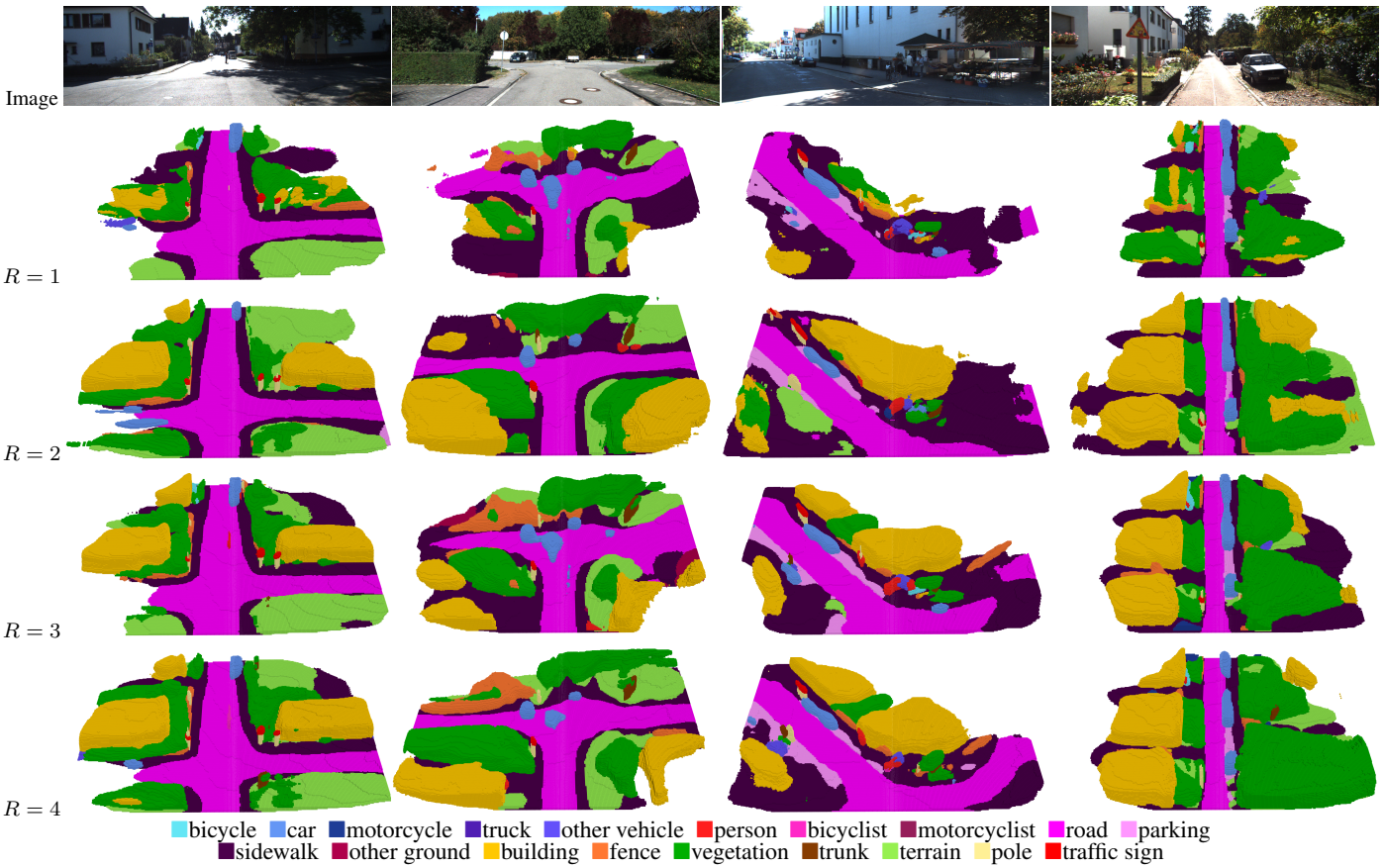

Figure 2. Qualitative comparisons on the VH rank R. We observe that an increase in the VH rank R usually corresponds to an enhanced representation capability.