N-Perspective View: Ablation Study on Tri-Perspective View on Plane Configurations

Erdao Liang†
University of Michigan, Ann Arbor
Ann Arbor, USA
erdao@umich.edu

Jialeng Ni†
University of Michigan, Ann Arbor
Ann Arbor, USA
jialeng@umich.edu

Jiyang Wang
University of Michigan, Ann Arbor
Ann Arbor, USA
realwjy@umich.edu

***Abstract*—This report investigates the impact of plane configurations in the Tri-Perspective View representation for 3D semantic occupancy prediction. Using the TPVFormer architecture, we conduct an ablation study evaluating different plane combinations on the Panoptic nuScenes dataset. Surprisingly, two-plane configurations outperform the original three-plane setup, suggesting that reducing redundancy may enhance model performance. These findings highlight a trade-off between feature completeness and simplicity, offering new insights for optimizing TPV-based frameworks in 3D scene understanding. Code: https://github.com/tomakeIT/TPVFormer.**

***Index Terms*—autonomous driving, tri-perspective view, 3D semantic occupancy prediction**

I. Introduction

3D semantic occupancy prediction is a critical task for 3D scene understanding in autonomous driving, requiring efficient representations to encode spatial information [1] [2]. The Bird's Eye View (BEV) framework has been widely adopted for simplicity but is limited in capturing 3D spatial features [3]. To address these shortcomings, the Tri-Perspective View (TPV) representation extends BEV by introducing three orthogonal planes to better preserve 3D geometric details [4]. Subsequent works have further developed upon the TPV representation, exploring its potential in enhancing 3D scene understanding [5] [6] [7].

The original TPVFormer paper demonstrated the effectiveness of this representation by achieving promising results on various 3D semantic tasks. However, it did not quantitatively explore the impact of individual planes or their combinations. Specifically, how performance changes when only one or two planes are used instead of all three remains an open question. Such an investigation is essential to fully understand the contribution of each plane to the overall performance.

In this report, we aim to bridge this gap by conducting an ablation study on TPV representation. Using the TPVFormer architecture, we evaluate various plane configurations on the Panoptic nuScenes dataset. We compare model performance across these settings using IoU and mIoU metrics. This work provides a deeper understanding of the TPV representation and offers valuable guidance for optimizing 3D scene representation frameworks.

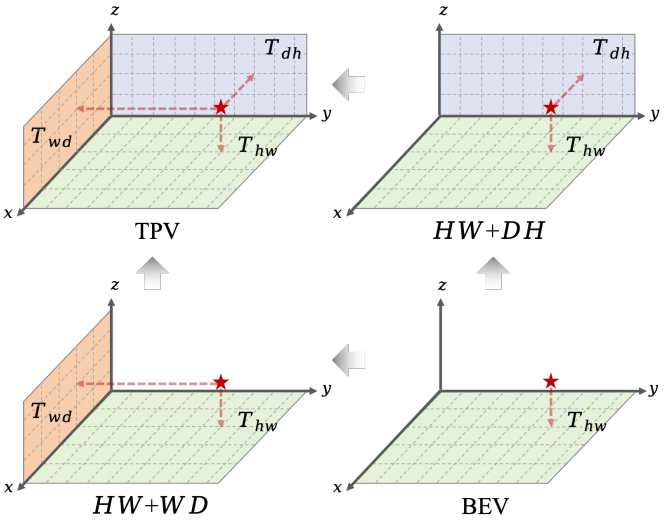

Fig. 1. Illustration of different plane configurations. In our study, two-planes settings including $HW + DH$ and $HW + WD$, which are between BEV and TPV, are tested and evaluated.

II. Methodologies

A. TPV Representation

The TPV representation builds upon the BEV framework [8] [3] [9] [10] , which is a commonly used method for encoding 3D scenes into a 2D feature map. Denote H, W, D the spatial resolution of the voxel space, and C the feature dimension, BEV compresses the 3D spatial information onto a single HW plane $\mathbf{B} \in \mathbb{R}^{H \times W \times C}$ (squeezing along the z-axis) to encode the top view of the scene. Each feature at a given 3D query voxel (h, w, d) is formulated as

$$\mathbf{b}_{h,w} = \mathcal{S}(\mathbf{B}, (h, w)), \tag{1}$$

where $\mathcal{S}$ is the sampling function.

TPV generalizes BEV by introducing three axis-aligned orthogonal planes: the top (HW), front (DH), and side (WD) views [4]. These three planes together form three 2D feature

†These authors contributed equally to this work.

maps, aiming at capturing information aong all three spatial axes:

$$\begin{aligned} &\mathbf{T} = [\mathbf{T}^{HW}, \mathbf{T}^{DH}, \mathbf{T}^{WD}], \quad \mathbf{T}^{HW} \in \mathbb{R}^{H \times W \times C}, \\ &\mathbf{T}^{DH} \in \mathbb{R}^{D \times H \times C}, \quad \mathbf{T}^{WD} \in \mathbb{R}^{W \times D \times C}. \end{aligned} \tag{2}$$

Given a 3D query voxel (h, w, d), its feature is extracted by projecting it onto the three TPV planes, sampling the features from the respective feature maps, and the summing these features to form the final feature representation:

$$\begin{aligned} \mathbf{t}_{h,w,d} =& \mathbf{t}_{h,w} + \mathbf{t}_{d,h} + \mathbf{t}_{w,d} \\ =& \mathcal{S}(\mathbf{T}^{HW}, (h, w)) + \mathcal{S}(\mathbf{T}^{DH}, (d, h)) \\ &+ \mathcal{S}(\mathbf{T}^{WD}, (w, d)). \end{aligned} \tag{3}$$

The key insight of TPV is that the spatial information along the orthogonal direction of any single plane (e.g. HW) is complemented by features sampled from the other two planes.

B. Motivation for Ablation Studies on TPV

While the TPV representation is intuitive and conceptually extends BEV, its reliance on three orthogonal planes has not been fully examined. Specifically, the original paper does not include any ablation study exploring how the number or choice of TPV planes impacts the performance on 3D perception tasks. For example, if only HW (top) plane is kept, which is theoretically equivalent to BEV, what is the resulting performance? If only a combination of HW plane and either one of DH (front) and WD (side) planes are kept, how does the performance compare to the full TPV? These possibilities are illustrated in Fig. 1.

The original TPV paper only ablates the network design of their TPVFormer architecture but does not investigate the TPV representation itself. To address this gap, we conduct experiments that specifically ablate the TPV representation.

C. Experimental Design

We conduct this ablation study using the 3D Semantic Occupancy Prediction (SOP) task based on the TPVFormer model proposed in the original paper. This task requires predicting dense voxel semantic labels for a given 3D scene, using only sparse ground truth LiDAR points for supervision. While the original paper only reports qualitative case studies on SOP, our ablation study will present quantitative results of IoU and mIoU on this task for comparison of different choices of feature planes, regardless a lack of official benchmarks.

We test three specific plane configurations: (1) only the HW plane (equivalent to BEV), (2) $HW + DH$ planes (top and front views), and (3) $HW + WD$ planes (top and side views). For each configuration, feature maps and query interactions associated with unused planes are pruned accordingly in TPVFormer.

The task is conducted on Panoptic nuScenes datasets [11] [12]. We adopt the TPVFormer-small model with a resolution of 100x100 and a feature dimension of 256, keeping all other parameters identical to those in the original paper. For each configuration, the model is re-trained starting from the open-sourced checkpoint for the fully implemented TPVFormer-small, using a single NVIDIA RTX 3090 GPU following the same training pipeline as the original work.

III. Results

A. Quantitative Results

Table I presents the voxel mIoU and per-class IoU on Panoptic nuScenes validation set for four different configurations aforementioned. With the original TPV, the model achieves a voxel mIoU of 40.2%. When only the HW plane is used, the mIoU drops significantly to 33.5%. This demonstrates the effectiveness of TPV in capturing 3D spatial features.

Surprisingly, when only one of the two additional planes (DH or WD), the performance improves beyond that of original configuration. Combining HW with WD achieves an mIoU of 48.8%, while HW with DH yields an mIoU of 47.9%, which indicates that using only two planes can result in a more effective representation for SOP. The improved performance with two-plane configurations might be attributed to reduced redundancy or interference between features extracted from the three planes. Integrating all three planes could introduce conflicting information or make feature fusion more challenging. These findings suggest that the balance between spatial completeness and feature simplicity plays a crucial role in 3D semantic understanding.

B. Case Study

Fig. 2 presents a keyframe showing the ground truth and the SOP results under different plane configurations. For each configuration, both top-view and isometric-view visualizations are provided.

When using only the HW plane, the model produces reasonable predictions along the horizontal plane but fails completely to capture any meaningful structure along the z axis, which demonstrates the necessity of introducing additional vertical planes to encode height information. However, adding just one of the vertical planes (either DH or WD) significantly improves the predictions. They produce results that closely resemble those of the original TPV setting. For instance, in the given scenario, key objects such as the bus, pedestrian and truck are all correctly identified in both two-plane settings. This observation underscores the effectiveness of even partial vertical plane integration in achieving accurate 3D semantic occupancy predictions.

IV. Conclusion

This report investigates the impact of plane configurations in the TPV representation for 3D semantic occupancy prediction. Our results show that two-plane configurations ($HW + DH$, $HW + WD$) unexpectedly outperform the original TPV setup, suggesting that reducing redundancy may enhance model performance and highlighting a trade-off between feature completeness and simplicity.

TABLE I
SEMANTIC OCCUPANCY PREDICTION RESULTS ON NUSCENES VALIDATION SET WITH DIFFERENT PLANE COMBINATIONS.

Method	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation	mIoU
Original TPV	42.5	**19.5**	66.0	59.6	28.6	**32.5**	31.5	11.3	35.7	59.0	42.2	37.4	43.3	48.9	43.5	42.1	40.2
HW Only (BEV)	46.6	8.2	60.6	59.7	30.5	23.7	19.5	7.9	11.2	57.2	66.6	41.4	37.9	35.8	18.3	10.9	33.5
$HW + DH$	**60.6**	13.0	71.2	**72.9**	36.0	29.2	32.6	**15.9**	**49.4**	**71.4**	84.7	49.8	52.4	50.1	42.3	39.7	47.9
$HW + WD$	56.6	12.9	**78.2**	72.7	**38.7**	30.8	**36.4**	12.8	45.8	63.7	**84.8**	**52.4**	**54.2**	**54.5**	**45.0**	**44.4**	**48.8**

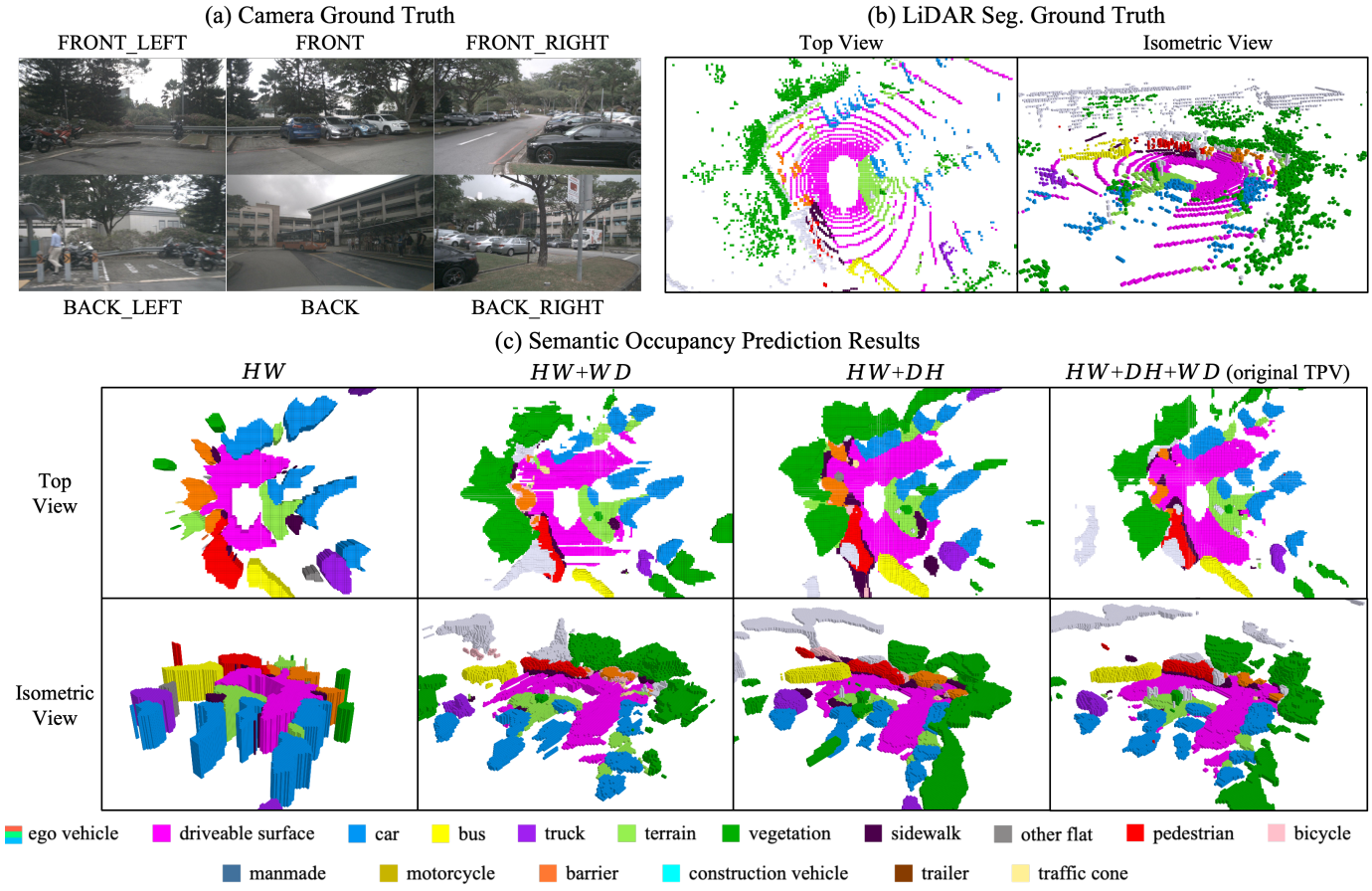

Fig. 2. Visualization results on 3D semantic occupancy prediction under different plane configurations.

However, the underlying reasons for this observation remain unclear, and it is uncertain whether this phenomenon generalizes to other tasks. Future work could analyze the feature distributions of three-plane and two-plane configurations to better understand these performance differences. Additionally, evaluating the two-plane configurations on tasks like 3D object detection or segmentation would further assess their generalizability and applicability.

REFERENCES

[1] H. Xu, J. Chen, S. Meng, Y. Wang, and L.-P. Chau, "A survey on occupancy perception for autonomous driving: The information fusion perspective," *Information Fusion*, vol. 114, p. 102671, 2025.

[2] Y. Zhang, J. Zhang, Z. Wang, J. Xu, and D. Huang, "Vision-based 3d occupancy prediction in autonomous driving: a review and outlook," *arXiv preprint arXiv:2405.02595*, 2024.

[3] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232, June 2023.

[5] S. Silva, S. Bhashitha Wannigama, R. Ragel, and G. Jayatilaka, "S2tpvformer: Spatio-temporal tri-perspective view for temporally coherent 3d semantic occupancy prediction," *arXiv e-prints*, pp. arXiv–2401, 2024.

[6] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "Pointocc: Cylindrical

tri-perspective view for point-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2308.16896*, 2023.

[7] Z. Yan, Y. Lin, K. Wang, Y. Zheng, Y. Wang, Z. Zhang, J. Li, and J. Yang, "Tri-perspective view decomposition for geometry-aware depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4874–4884, 2024.

[8] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*, pp. 1–18, Springer, 2022.

[9] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "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.

[10] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781, IEEE, 2023.

[11] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

[12] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada, "Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3795–3802, 2022.