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

Erdao Liang[1], Jialeng Ni[2], Jiyang Wang[3] *University of Michigan, Ann Arbor, USA*
[1]erdao@umich.edu, [2]jialeng@umich.edu, [3]realwj@umich.edu [3]

Introduction

Understanding 3D scenes is crucial for autonomous driving and related applications. Among the numerous techniques available, 3D semantic occupancy prediction has gained significant attention for its ability to encode spatial information efficiently. Although Bird's-Eye View (BEV) representations have become popular due to their simplicity, they struggle to fully capture the richness of 3D space [2]. To overcome these limitations, the Tri-Perspective View (TPV) representation extends BEV by introducing three orthogonal planes, aiming to preserve detailed 3D geometric information [1].

Despite TPV's conceptual appeal, its **foundational assumptions—such as the necessity of all three planes—remain under-explored**. This gap motivates our ablation studies, where we systematically evaluate the importance of each plane configuration within the TPV framework.

Summary

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.

TPV Representation

The TPV representation extends the widely adopted BEV framework [2], which projects a 3D scene into a 2D feature map. Let H, W, and D represent the spatial dimensions of the 3D voxel space, and C the feature dimension. In standard BEV, the 3D scene is collapsed onto a single HW plane, $\mathbf{B} \in \mathbb{R}^{H \times W \times C}$, effectively discarding the vertical dimension.

In contrast, TPV employs three orthogonal 2D feature maps:

$$\mathbf{T} = [\mathbf{T}^{HW}, \mathbf{T}^{DH}, \mathbf{T}^{WD}], \quad \mathbf{T}^{HW} \in \mathbb{R}^{H \times W \times C}, \quad \mathbf{T}^{DH} \in \mathbb{R}^{D \times H \times C}, \quad \mathbf{T}^{WD} \in \mathbb{R}^{W \times D \times C}. \tag{1}$$

For a given voxel (h, w, d), its feature vector is derived by projecting onto all three planes and summing the sampled features:

$$\mathbf{t}_{h,w,d} = \mathcal{S}(\mathbf{T}^{HW}, (h, w)) + \mathcal{S}(\mathbf{T}^{DH}, (d, h)) + \mathcal{S}(\mathbf{T}^{WD}, (w, d)), \tag{2}$$

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

By distributing spatial information across these three orthogonal planes, the TPV representation recovers details lost in a single-plane projection. This ensures a richer, more comprehensive 3D feature representation, ultimately improving the understanding of complex 3D scenes.

[1] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen 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)*, pages 9223–9232, June 2023.

[2] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar 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.

Experiment Setup

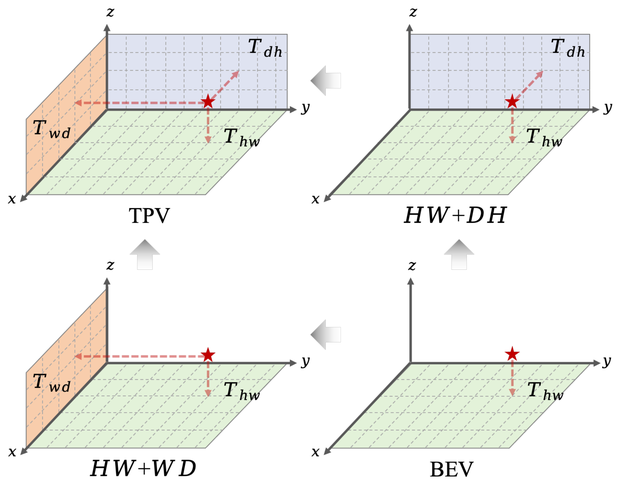

Figure 1. Illustration of different plane configurations. We examine two-plane settings ($HW + DH$ and $HW + WD$) that lie between the single-plane BEV and the full three-plane TPV, highlighting the contribution of each additional viewpoint.

While TPV naturally extends BEV, there has been no thorough examination of how varying the number or combination of planes influences 3D perception performance. For instance, what if only the HW (top) plane is retained, effectively reverting to a BEV-like setting? Alternatively, how does the performance compare when using just $HW + DH$ or $HW + WD$ plane pairs against the full three-plane TPV?

These scenarios, illustrated in Figure 1, have remained unexplored in the original TPV work, which primarily focused on network architecture rather than the TPV representation itself. To fill this gap, we present a series of ablation studies explicitly designed to isolate and evaluate the contributions of each plane configuration.

- **Revisiting BEV**: Evaluate performance when reverting to a single top (HW) plane.
- **Partial TPV Configurations**: Compare two-plane combinations ($HW + DH$, $HW + WD$) with the full three-plane TPV.
- **Comprehensive Analysis**: Identify the relative importance of each plane and its impact on capturing 3D scene structure.

Key Insight

Our ablation studies highlight critical insights into the necessity and sufficiency of different plane combinations. By dissecting the TPV representation, we aim to clarify its fundamental mechanics and guide future research toward more efficient and effective 3D perception pipelines. This analysis provides a clearer understanding of when and why multiple perspectives are beneficial, laying the groundwork for more targeted and potentially simplified 3D representations.

Results

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**

Figure 2. Semantic occupancy prediction results on the nuScenes validation set under different plane configurations. It measures the voxel-level mean Intersection over Union (mIoU) and per-class IoU metrics on the Panoptic nuScenes validation set for four configurations

Illustrated in Figure 2, the original TPV model achieves a voxel mIoU of 40.2%. However, relying solely on the HW plane results in a substantial drop in mIoU to 33.5%, emphasizing the critical role of vertical information.

Interestingly, configurations with two planes—$HW + WD$ or $HW + DH$—outperform the full TPV. Specifically, the $HW + WD$ configuration achieves an mIoU of 48.8%, while $HW + DH$ achieves 47.9%. This indicates that reducing the number of planes can simplify the representation while enhancing performance, likely due to **reduced redundancy and more effective feature fusion**.

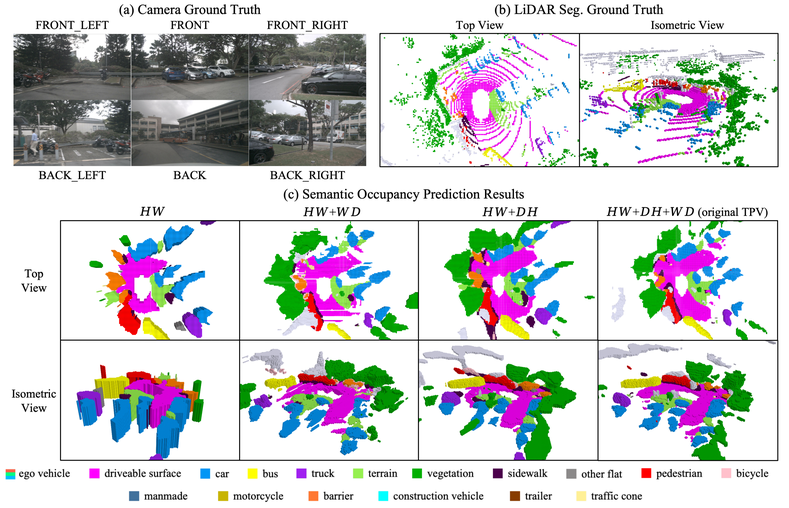

Figure 3. Visualization of 3D semantic occupancy predictions under different plane configurations.

Figure 3 visualizes a representative keyframe, displaying both the ground truth and predictions from various configurations, presented from top-view and isometric-view perspectives. When using only the HW plane, horizontal structures are preserved, but vertical information is absent. Adding a single vertical plane (DH or WD) produces predictions comparable to those of the full TPV. Notably, key objects such as buses, pedestrians, and trucks are accurately identified with just two planes. These results demonstrate that **partial vertical integration can significantly enhance 3D semantic occupancy prediction while reducing the complexity associated with a full three-plane setup**.